# University of Southampton

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Web Science

**Connecting Peace Studies and Natural Language Processing to Rethink Hate Speech Detection as Hostile Narrative Analysis**

by

**Stephen Paul ANNING**

ORCID ID: https://orcid.org/0000-0003-4911-7907

Thesis for the degree of Doctor of Philosophy

July 2023

# University of Southampton

## Abstract

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF WEB SCIENCE

THESIS FOR THE DEGREE OF DOCTOR OF WEB SCIENCE

CONNECTING PEACE STUDIES AND NATURAL LANGUAGE PROCESSING TO RETHINK HATE SPEECH DETECTION AS HOSTILE NARRATIVE ANALYSIS

by

Stephen Paul ANNING

In response to limitations with current computational methods of hate speech detection, this research connects Peace Research and Natural Language Processing (NLP) to propose the idea of hostile narrative analysis. The corpus guiding this research contrasts Hitler's *Mein Kampf* and texts from the 'War on Terror' era with speeches from Martin Luther King, who advocated for non-violent change. Experiments using this corpus find the current computational methods of hate speech detection are unconnected to a defining theory, which questions their explanatory rigour. Hate speech itself is a polysemous term, and using the computational method of text classification skews an orator's intended meaning. The response to this finding with hostile narrative analysis draws upon Galtung's theory of cultural violence from Peace Research to detect the 'Self-other gradient'. This gradient refers to processes of violence legitimisation by elevating the Self while deflating or debasing the value of the Other. As a broad hypothesis, the steeper the gradient between the Self and the other, the more legitimate violence becomes. The computational methods for detecting the Self-Other gradient then draw upon pattern-based methods in NLP. As a general observation, problems with current computational methods arise from a technical first approach before applying theory; this paper begins with cultural violence theory to guide technological development. This paper seeks to contribute to the field of Web Science and, to the best of my knowledge, constitutes the first attempt to connect cultural violence with NLP to analyse hostile narratives.

# Table of Contents

# Table of Figures

# Research Thesis: Declaration of Authorship

Print name: Stephen Paul Anning

Title of thesis: Connecting Peace Studies and Natural Language Processing to Rethink Hate Speech Detection as Hostile Narrative Analysis

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Anning, S., Konstantinidis, G., & Webber, C. (2021, June). Social Science for Natural Language Processing: A Hostile Narrative Analysis Prototype. In *13th ACM Web Science Conference 2021* (pp. 102-111)

Signature: Date: 7th July 2023

# Acknowledgements

I would like to thank the following for their contribution to this thesis.

- Dr Nada Albunni for assistance with interpreting Arabic terms,
- Mark Neumann of AllenNLP and Dr Pavlos Vougiouklis of Southampton University for their assistance with understanding NLP methods.
- Dr Phil Taylor of Trilateral Research for assistance with Python coding.
- Dr Katrina Petterson of Trilateral Research for her advice on linguistic theory and Stuart Hall in particular.

I would also like to pay particular thanks to my supervisors Drs George Konstantinidis and Craig Webber for their insight and support.

Most of I would like to thank my wife for her unfailing support during the last five years of developing this thesis. In this time, we have also married, had two children, and moved house three times. She has been a rock.

# Definitions and Abbreviations

AI ........................................ Artificial Intelligence.

Bag of words (BOW)............ representing a text with a vector indicating the number of occurrences of each chosen word in the training corpus.

Bidirectionality .................... an NLP technique that accounts for words to the left and right of a target word when generating a word embedding.

Cultural Violence ................. Those aspects of culture, the symbolic sphere of our existence – exemplified by religion and ideology, language and art, empirical science and formal science (logic and mathematics) – that can be used to justify or legitimise direct or structural violence.

Code (technical) .................. symbols with a rigidly fixed interpretation.

Code (linguistic)................... systems of symbols that links sign-vehicles, such as words, to semantic units of meaning.

Connotative meaning........... where the implication of a word depends on the relationship between the orator and audience.

Coreference resolution ........ identifying all the pronouns and noun phrases that refer to the same named entity.

Dog whistle politics ............. coded racial appeals that carefully manipulate hostility towards non-whites.

The dominant-hegemonic position…where an orator and their audience share the same connotative meaning of a message.

Denotative meaning............ where a word has fixed interpretation between an orator and audience.

Dependency labelling........... assigning attributes to each token according to a word's grammatical function in a sentence.

Encoding and decoding ........ a theory for understanding how audiences differently interpret messages.

Entity resolution.................. assigning a unique identifier to different mentions of the same named entity across a text.

Definitions and Abbreviations

Explainable AI ...................... a movement in AI that places emphasis on human-understandable algorithms.

Explanatory dialogue ........... dialogue between an enquirer and explainer through which understanding develops and evolves over time.

Functional applications........ NLP applications whose outputs have narrow degrees of interpretation by audiences.

Grammar patterns ............... the grammatical arrangement of words in phrases, clauses, and sentences.

Grammatical ........................ the relationships between words that conform to the rules of a particular language.

Group .................................. a collection of individuals who perceive themselves to be members of the same social category, share some emotional involvement in a common definition of themselves and achieve some degree of social consensus about the evaluation of their group and their membership of it.

Historic truth........................ the lived experience of the characters in a narrative.

Homonyms........................... words that share the exact spelling but have different meanings.

Hostile narrative .................. a story used to legitimise violence against another person or group.

Hostile narrative analysis..... detecting processes of violence legitimisation in natural language.

Hybrid NLP ........................... the combination of quantitative methods for labelling a world's lexical properties and pattern-based NLP for parsing language clauses.

Hypernymy .......................... a way to discover a hyponymic lexical relationship between two or more noun phrases in a naturally occurring text.

Knowledge graph ................. a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities.

LDA...................................... latent dirichlet allocation.

Lexical ................................. the linguistic attributes of individual words and phrases of a language.

Lemmatisation .................... representing a word using its conical head, known as a lemma.

Machine learning ................ field of study that gives computers the ability to learn without being explicitly programmed.

Named entity recognition (NER)…the task of classifying tokens of interest in a sequence of tokens into specific entity types, such as a person, an organisation, or a location.

Named concept recognition…labelling a word's denotative and connotative meanings.

Narrative ............................. stories of experience used by an orator to rationalise events and create moral tales of how the world should be.

Narrative analysis................ a family of methods for interpreting texts that have a common storied form to interrogate an orator's intention and language.

Narrative truth .................... an imagined reality expressed in a narrative.

Negative peace ................... the absence of violence.

The negotiated position ....... where a recipient's interpretive framework 'contains a mixture of adaptive and oppositional elements' whereby the general meaning of the hegemonic code is accepted but with a recipient's own ground rules.

NLP ...................................... Natural language processing.

Noun chunking .................... combining grammatically related tokens to create noun phrases.

The oppositional position .... where an audience understands the denotative and connotative meanings but decodes a word within a contrary frame of reference.

Part-of-speech (POS) tagging…assigning an attribute label to each token to indicate a word's grammatical function in a sentence.

Pattern-based NLP .............. The processing of text by language clauses.

Pattern grammar................. the theory behind investigating the grammatical structures of language.

Peace Research ................... concerning the research and teaching about causes of violence, including war, and the conditions of peace.

Personal experience narrative (PEN)….verbal technique for recapitulating experience, in particular, a technique of constructing narrative units which match the temporal sequence of that experience.

Positive peace..................... the egalitarian distribution of power and resources.

Qualitative approaches........ typically used to explore new phenomena and to capture individuals' thoughts, feelings, or interpretations of meaning and process.

Quantitative approaches..... typically used to quantify the collection and analysis of data using mathematical methods.

RDF...................................... Resource Description Framework.

Self-attention...................... sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence.

Semiotics............................. the study of signs, symbols, and communication systems, including the interpretation and use of meaning in different contexts.

Semiosis.............................. process of meaning-making in which a sign, such as a word, image, or gesture, is interpreted in context to create a message.

Semantics............................ the branch of linguistics devoted to the investigation of linguistic meaning.

Social science applications... NLP applications whose outputs have wide, sometimes opposing, degrees of interpretation by audiences.

Sociotechnical research....... research that considering the interactions of humans and machines.

Text classification................ a quantitative method that draws upon statistical techniques to predict text classifications using a pre-labelled training dataset.

Text pre-processing ............ encoding a text into meaningful units for onward processing by an NLP algorithm.

Tokenisation ....................... the segmentation of a text into basic units—or tokens—such as words and punctuation.

Transformer........................ an NLP architecture using self-attention to solve sequence-to-sequence tasks while handling long-range dependencies.

Word embedding................ a quantitative method in NLP to encode word meanings in a numerical vector.

# Chapter 1 Introduction

Social volatility and the proliferation of Web technologies have increased the public awareness of online hate speech (Carter et al., 2020; Tell MAMA, 2020; Vidgen et al., 2019). Computer science has responded with a new natural language processing (NLP) literature in hate speech detection. For the computational task of detecting hate speech, NLP literature tend to draw upon quantitative methods found in machine learning algorithms like text classification (Mullah & Zainon, 2021a; Poletto et al., 2020) or vector representations (Alorainy et al., 2018; Cao et al., 2020; Kapil & Ekbal, 2020). Such algorithms are transforming the relationship between humans and machines for what this thesis categorises as *functional applications*, like virtual assistants or language translation. With reasonable reliability, the quantitative methods of these applications generate generally useful outputs from natural language inputs.

In contrast to functionally defined applications, however, hate speech detection falls within a category this thesis refers to as *social science applications*. These applications concern inferring meaning from text, whether 'opinions, speculations, beliefs, emotions, and any other evaluative views' (Hirschberg & Manning, 2015, p. 265). The difference between each application is the degrees of interpretation by audiences: functional applications have narrow interpretation whereas social science applications are more subjective, therefore, they have wide, often opposing, degrees of interpretation. With hate speech, for example, different people have vastly different opinions about what is hateful, and while some beliefs have more acceptance than others, there is no single objective truth (Röttger et al., 2021, p. 1). Indeed, the research and experimentation in this thesis question the explanatory rigour of quantitative methods to interpret hate speech and, more broadly, infer sentiment. This thesis responds by connecting Peace Research and NLP to rethink aspects of hate speech detection as *hostile narrative analysis* using qualitative methods.

## 1.1 Research Questions and Contribution

The idea behind hostile narrative analysis is to analyse processes of violence legitimisation in natural language. For the computational methods of hostile narrative analysis, this sociotechnical thesis is about integrating qualitative approaches with NLP to analyse such linguistic processes of violence legitimisation. The methodological framework presented in Chapter 3 for analysing hostile narratives draws upon Johan Galtung's (1969) theory of cultural violence from Peace Research. Accordingly, cultural violence theory provides a basis for rethinking aspects of hate speech detection as hostile narrative analysis. The computational methods of hostile narrative analysis respond to limitations with inferring meaning using quantitative methods by drawing upon qualitative research methods, like semiotics and semantic analysis. These methods then apply semantic analysis using hybrid NLP that augments quantitative methods for labelling a word's lexical properties with pattern-based NLP for processing a text by language clauses. As such, the following research hypothesis and questions guide this thesis:

> **Research Hypothesis**: Integrating qualitative methods with NLP improves the meaningful analysis of hostile narratives.
>
> **RQ1**: To what extent do quantitative methods in NLP 'understand' social science applications?
>
> **RQ2**: How can integrating Peace Research and NLP enable the meaningful analysis of hostile narratives?
>
> **RQ3**: How does augmenting quantitative NLP methods with qualitative approaches enable the meaningful analysis of hostile narratives?

The dependent variable of the research hypothesis is 'meaningful analysis of hostile narratives', while the independent variable is 'integrating qualitative methods with NLP'. The dependent variable problematises the idea of 'Explainable AI' for meaningfully analysing hostile narratives. In a shift from opaque to more human-understandable algorithms, the Explainable AI movement is gaining increased provenance in machine learning (Adadi & Berrada, 2018; Saeed & Omlin, 2021). O'Hara (2020) describes 'explanation' itself as 'the achievement of understanding of a phenomenon by the audience' which occurs as 'a process or performance that exists through time' (O'Hara, 2020, p. 3). Explanation, therefore, refers to an explanatory dialogue between an enquirer and explainer through which understanding develops and evolves over time. Dialogues themselves are generally processes for decision making. The explanatory dialogue for hate speech detection is content moderation on social media in which moderators decide upon the suitability of users' messages. For testing this hypothesis, the thesis assesses how well an algorithm's output contributes to a meaningful explanatory dialogue.

A series of experiments in Chapter 2 assess the null hypothesis of 'quantitative methods provide a meaningful analysis of hostile narratives'. These experiments test algorithms provided by

UnitaryAI[1], Gensim[2], Google[3], IBM[4] and Microsoft[5] , along with a non-machine learning API provided by TextBlob[6]. Each experiment uses the hostile narrative dataset introduced below to find limitations in each algorithms' ability to represent the accepted meaning of each text. For example, even the most sophisticated APIs failed to distinguish between the very opposite sentiments of Hitler's *Mein Kampf* and Luther King's *I Have a Dream* speech[7]. None of these algorithms, regardless of technical sophistication, provide informative inputs to an explanatory dialogue. As is explained, the limitations of these technologies arise from the treatment of text as unstructured data for common quantitative methods of NLP.

The independent variable, 'the integration of qualitative methods with NLP', responds to these limitations of using quantitative methods for processing natural language. As a research paradigm, qualitative research has no distinct methodologies and practices. As a toolbox of methods to support methodologies, qualitative researchers generally use semiotics, narrative, content, discourse, archival, and phonemic analysis-even statistics, tables, graphs, and statistics (Denzin & Lincoln, 2018, p. 12). The growing field of Digital Humanities sees increasing integration of these tools with computational methods (Dunn & Schuster, 2020; Orlandi, 2021; Van Der Zwaan et al., 2017), as a series of tools to inform explanatory dialogues about hostile narratives. The qualitative method of this thesis draws upon peace research and narrative analysis for the methodological framework, and semantic analysis for the computational methods. This thesis shows how using these methods improves upon the quantitative methods most commonly found in current NLP algorithms.

To address each research question, this thesis draws upon sociotechnical research to consider the interaction of humans and machines within an explanatory dialogue. 'Sociotechnical systems theory highlights the links between technical systems, consisting of technology and processes, and social systems, consisting of people and relationships, to focus on the joint optimization of an organization's human and technology dimensions within a given context' (Makarius et al., 2020, p. 263). As such, the research questions consider how humans and machines interact within an

---

[1] UnitaryAI (2020) Detoxify, retrieved on 17th Feb 2023

[2] Gensim (n.d.) Word2vec embeddings, retrieved on 17th Feb 2023

[3] Google (n.d.) Python Client for Natural Language API, retrieved on 17th Feb 2023

[4] IBM (n.d.) Watson Natural Language Understanding, retrieved on 17th Feb 2023

[5] Microsoft (n.d.) Text analytics, retrieved on 17th Feb 2023

[6] Loria (2020) TextBlob, retrieved on 17th Feb 2023

[7] Anning (2023) Testing Sentiment Analysis, retrieved on 17th Feb 2023

explanatory dialogue. Chapter 2 actually questions the extent to which hate speech detection literature sufficiently considers human or social elements of detecting hate, thereby questioning what an algorithm actually explains. In particular, literature from both the computer and social sciences appear to have no unifying methodology for what constitutes hate speech. Without such a unifying theory, the literature treats the detection of hate speech as a purely technical, or indeed, quantitative problem.

In a sociotechnical system of humans and machines, a machine's *output* might more appropriately be called an *input* to an explanatory dialogue. As Doran *et al*. (2017) observe, 'it is prudent for an AI to provide not only an output but also a human-understandable explanation that expresses the rationale of the machine' (Doran et al., 2017, p. 2). As informative inputs to explanatory dialogues, such rationales provide a human-understandable explanation for why a machine produced a particular output/input. In their review of multimodal AI models, Randy *et al*. (2018) shows how rationales explain an AI's interpretation of images (Randy et al., 2018). In problematising Explainable AI, therefore, the dependent variable is about how well an algorithm explains its output, preferably with some sort of accompanying rationale.

Interdisciplinarity is a central feature of sociotechnical research; Peace Research itself is an interdisciplinary topic, as is linguistics and artificial intelligence all of which contribute to this thesis. Interdisciplinarity is 'a process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline and draws on disciplinary perspectives and integrates their insights to produce a more comprehensive understanding or cognitive advancement' (Repko, 2012, p. 12). To avoid a possible recursion of interdisciplinary research about interdisciplinary subjects, the methodological framework presented in Chapter 3 provides a unifying idea for each contributing theory and discipline to this thesis. Indeed, this thesis focuses more on transdisciplinarity by offering contributions of Peace Research and Computer Science where each learns from and adapts to findings in the other.

Answering each research question then addresses the social and technical elements of this research. Chapter 2 addresses RQ1 through a series of experiments with NLP algorithms that are not integrated with qualitative methods. The remaining chapters address RQ2 and RQ3 by assessing how integrating qualitative research methods with NLP improves the explanatory value of an algorithm's output for an explanatory dialogue. In response to RQ2, Chapter 3 rethinks Matsuda's (1989) conception of hate speech as hostile narrative analysis. Chapter 4 addresses RQ 3 by showing how hybrid NLP can produce rationales as inputs to a meaningful explanatory dialogue about hostile narratives. The discussion chapter places the findings from each chapter in the policy context of tackling online abuse and growing the UK's AI industry. Further work develops upon the methodology and method with the continued integration of narrative theory and development of knowledge graphs for the computational methods.

As the conclusion explains, the finding of each chapter all point to the requirement to tackle online abuse and growing the UK' s AI industry as a sociotechnical problem. The findings from Chapter 2 and the policy discourse suggest a purely technical approach to tackling online abuse and the development of social science applications. This technical approach has become the state-of-the-art, which largely relies upon treating text as unstructured data for sophisticated quantitative methods. The interaction of Chapter 3 and Chapter 4 show how a sociotechnical approach that augments quantitative methods with qualitative approaches produces more meaningful inputs to explanatory dialogues about hostile narrative. In particular, Chapter 4 shows how treating text as structured data is consistent with linguistic theory and enables the generation of a rationale for explanatory dialogues. The corresponding central shift of the social element in this thesis, therefore, rethinks hate speech detection as hostile narrative analysis using social scientific theories, and the technical element shifts from treating text as unstructured data to treating text as structured data to generate meaningful insights from natural language.

The corresponding aim of this thesis is to develop computational methods that provide improved rigour for explanatory dialogues about hostile narratives. As will become apparent across this thesis, there is no consensus for what constitutes hate speech. The underlying premise and use of quantitative methods, however, suggest there could be. Moreover, current quantitative methods do not account for ingroup elevation, which additionally suggests the absence of an underlying explanatory methodology. Further problems arise from the term 'hate'; it is a morally loaded term that applies moral judgement against a person accused of using hate speech. Hate speech itself becomes a form of othering by the accuser. As Chapter 3 explains, this moral judgement features (at least implicitly) in disputes when using social media text as training data for hate speech detection algorithms. While the algorithm may judge a text to be hateful, the accused uses a 'free speech' defence to claim the text is morally acceptable. The methods, methodology and choice of training data do not have sufficient rigour to explain occurrences of hate speech.

Accordingly, the integration of qualitative approaches with quantitative computational methods proposed by this thesis aims to improve how to explain hostility in natural language. In the first instance, 'hostile' is intended to replace 'hate' as a much less morally loaded term. Language can be hostile whether morally 'right' or 'wrong'; irrespective of their morality, the narratives by Bush, bin Laden and Hitler used in this thesis are hostile given since all represent violence legitimisation. In contrast, the word 'hate' implies wrongfulness since there is no 'right' form of hate speech. That these narratives are more objectively hostile than social media texts is why they have been chosen for this research. Nevertheless, this thesis recognises the limits of computational methods and the necessity of human judgement in determining the extent of hostility in language. The proposed methodological framework provides a way to explain hostility, while the corresponding computational methods provide the corresponding data for this methodology. The interaction of the methodology with the computational methods, therefore, seek to provide the necessary rigour for informed explanatory dialogues about violence legitimisation in language. The rigour is found in the accompanying rationale to the algorithm's output that can be accepted, rejected or modified during an explanatory dialogue.

This introduction continues by introducing the primary contributing theories and disciplines as background to the thesis. Since the sociotechnical approach of this thesis is about integrating social and technical elements of hostile narrative analysis, these more sociological theories should be of particular interest to a technical audience. The section on NLP explains the general principles and should be of interest to the social audience. The common ground is the use of quantitative and qualitative methods to conceptualise methodologies and methods.

**Background**

## 1.2 What is a Hostile Narrative?

This thesis uses cultural violence from Peace Research to define a hostile narrative as a story used to legitimise violence against another person or group; to analyse a hostile narrative is to detect how processes of violence legitimation feature in natural language. This section introduces Peace Research as the guiding discipline of this thesis. This section begins by introducing and establishing the provenance of Peace Research, and then explains how Galtung's cultural violence provides a basis for detecting processes of violence legitimation in natural language. The section then continues by explaining the meaning of narrative and its function in intergroup relations. As such, the analytical object of this thesis are the stories used in violence legitimisation between groups. The section also explains the role of quantitative and qualitative methods in the analysis of these stories. The subsequent section then introduces the constituent elements of NLP for the computational methods of analysing hostile narratives.

Peace research is an interdisciplinary subject for exploring the conditions for Peace and War (Kelman, 1981, p. 95; Stephenson, 2020, p. 1). With 'Peace' itself referring to the absence of violence, the twin goals of peace research are preserving negative peace while promoting positive peace. The negative formulation of peace refers to the absence of violence while positive peace refers to 'social justice' as a positively defined condition whereby institutions concern themselves with the egalitarian distribution of power and resources (Galtung, 1969, p. 183). Accordingly, Kelman (1981) equates 'negative peace' with the 'absence of systematic, large-scale, collective violence, accompanied by a sense of security that such violence is improbable', and 'positive peace' with a world order 'concerned with meeting the needs and interests of the world population' (Kelman, 1981, p. 103). Critics of Peace Research find general critique on the vagueness of positive peace and an over-focus on negative peace (see: Gleditsch et al., 2014). Negative peace for online interactions refers to unlikely absence of online harm, whereby positive peace refers to the effective moderation of online abuse.

Galtung's experience and research provide a solid foundation for understanding violence legitimisation. According to a biographical chapter from his book, Galtung became motivated to study peace following World War 2 during which his father was enslaved in a Nazi concentration camp. Following his father's safe return, Galtung resolved to work for the prevention of war, but upon beginning his studies in Helsinki could only find publications on military strategy whereas books on peace could not be found. His first publication was 'Gandhi's Political Ethics' in 1955, indeed one researcher finds, 'a strong causal Gandhian underpinning to Galtung's Peace Research' (Weber, 2004, p. 42) . Galtung has since published over 160 books, 1,600 book chapters and articles in academic journals, 40 of his books have been translated into 34 languages making him the 'most cited author in the field of peace studies' (Galtung & Fischer, 2013a, p. 4).

While providing a significant academic contribution to peace research, Galtung has also founded organisations to apply his research. In 1959 Galtung and his wife founded the Peace Research Institute, Oslo (PRIO), which has since made a significant contribution to progressing the field and provides a strong reference point for this research (Gleditsch et al., 2014, p. 146). In 1993 he also founded Transcend International which now comprises 400 scholars-practitioners from more than 60 countries as members, 'to bring about a more peaceful world by using action, education/training, dissemination, and research to transform conflicts non-violently, with empathy and creativity, for acceptable and sustainable outcomes'[8]. In mobilising these organisations, he has helped mediate over 100 international conflicts and is often sought by Prime Ministers and Presidents for advice (Galtung & Fischer, 2013a, p. 4). Through this significant contribution to the pursuit of peace, Galtung is widely regarded as a pioneer of Peace Research.

As Galtung explains, cultural violence works 'by changing the moral colour of an act from red/wrong to green/right or at least to yellow/acceptable; an example being murder on behalf of

[8] Transcend International: A Peace Development Network (n.d.) About us, retrieved on 17th Feb 2023

the country as right, on behalf of oneself wrong' (Galtung, 1990, p. 292). In this example, the moral frameworks of humanitarian law and the warrior codes of warfare legitimise the intentional killing of another human during war. Outside of this legitimisation framework in peace time, however, the same act is considered murder. The moral colour of violence is a re-occurring theme of this thesis, particularly since religious and ideological frameworks provide a moral lens through which people legitimise otherwise morally questionable acts. Such lenses, therefore, are highly subjective since they are relative to an orator and their audience. Accordingly, this thesis uses moral codes to reveal cultural violence contained in hostile narratives that contribute to the legitimisation of violence.

## 1.3 What Is Narrative Analysis?

This subsection develops upon the definition of a hostile narrative from the previous section using narrative theory. There are several ways to define a narrative. In general terms, Riessman (2005) describes narratives themselves as 'stories of experience' used by an orator to rationalise events and create moral tales of how the world should be (Riessman, 2005). As moral tales, narratives provide a valuable function in group dynamics by showcasing 'the values being taught by [group] culture…in a less formal and more enjoyable manner' (Akinsanya & Bach, 2014). Concerning what narratives mean, Chase (2011) suggests narratives shape and order experience 'as a way of understanding one's own or others' actions or organising events and objects into a meaningful whole, or of connecting and seeing the consequences of actions and events over time' (Chase, 2011, p. 421). Van Dijk's (1983) exploration of the roots of racism suggests these stories provide an 'important social database on which further talk, shared opinions and attitudes are based' (van Dijk, 1983, p. 65). The rationalisation of events is common to all descriptions of narrative.

Narrative analysis 'refers to a family of methods for interpreting texts that have a common storied form to interrogate an orator's intention and language (Riessman, 2008, p. 11). Propp (1968) provides an early example of narrative analysis which reveals common syntagmatic structures within folktales (Propp, 1968). In contrast to folktales, Labov and Waletzky (1997) respond to Propp (1968) by revealing common structures in everyday stories, which they define as *personal experience narratives* (Labov, 1997; Labov & Waletzky, 1997). Through the 'narrative turn', Riessman (2008) observes how the analytical study of narrative now finds itself in virtually every field and social science discipline (Riessman, 2008, p. 12). Examples include narrative criminology (Presser, 2009), victimology (Pemberton et al., 2019), substance abuse (Larsson, 2019), and domestic abuse (Rogers, 2021; Spruin et al., 2015). This thesis offers a new application of narrative analysis for the detection of violence legitimisation.

Labov and Waletsky (1997) proposes an approach to analysing narratives using language clauses. 'A narrative is a story that contains a sequence of events that take place over a time period…it mostly follows a chronological order and usually contains a link to the present on the form of a lesson learnt by the narrator…narrative analysis seeks to find the link by analysing and evaluating various parts of the narrative' (Akinsanya & Bach, 2014, p. 1). As a subtype of narrative, Labov and Waletsky (1997) define a personal experience narrative as a 'verbal technique for recapitulating experience, in particular, a technique of constructing narrative units which match the temporal sequence of that experience' (Labov & Waletzky, 1997, p. 13). They then analyse each narrative element as a sequence of clauses that contain at least one temporal juncture. As such a narrative 'consists of at least two narrative clauses, and this sequence of clause is matched to a sequence of events which actually occurred (Labov & Waletzky, 1997, p. 12).

As will be explained across this thesis, the central point of any hostile narrative is the promotion of violence by elevating an ingroup while othering an outgroup. As Presser (2018) suggests, 'war and other mass harms…are typically promoted by stories of a virtuous protagonist facing off against a malevolent other whose forceful overcoming is necessary for salvation' (Presser, 2018). In other violent genres, Van Dijk (2014) observes how stories express and reproduce racism, by persuasively pointing out that 'we' are better than 'them', or rather 'they' fail to meet the standards set by 'our' values and norms' (Van Dijk, 2014, p. 141). These basic stories reveal a common plot to elevate the heroes while othering the villains of a story. The plot of each story type creates a status difference between the orator, their social group and some perceived 'other'. The purpose of elevation and othering in a hostile narrative, therefore, is to create distance between the ingroup and outgroup.

The central idea of rethinking hate speech detection as hostile narrative analysis is to consider different types of hate speech as different *genres* of hostile narrative. As a representative corpus of hostile narrative genres, this thesis uses two historical case studies. The hate speech case study compares *Mein Kampf* by Adolf Hitler and Martin Luther King's *I Have a Dream* speech. Text for *Mein Kampf* is from Hitler.org[9] and *I Have a Dream* is from American Rhetoric[10]. *Mein Kampf* more specifically provides data for the antisemitic and genocidal genres. In that he advocated for non-violent change[11], Luther King's *I Have a Dream* provides a case study for the non-violent

[9] Hitler.org (n.d.) Mein Kampf, retrieved on 17th Feb 2023

[10] American Rhetoric (n.d.) I Have a Dream, retrieved on 17th Feb 2033

[11] The Martin Luther King, Jr. Research and Education Institute (n.d.) Nonviolence, retrieved on 17th Feb 2023

genre and serves as control data for non-hate speech. The thesis, therefore, presumes *Mein Kampf* to be a conical text of hate speech that should generate high scores on a detection system with little fine-tuning. Conversely, it presumes that *I Have a Dream* should not generate any scores for hatefulness. These are clearly differentiated speeches.

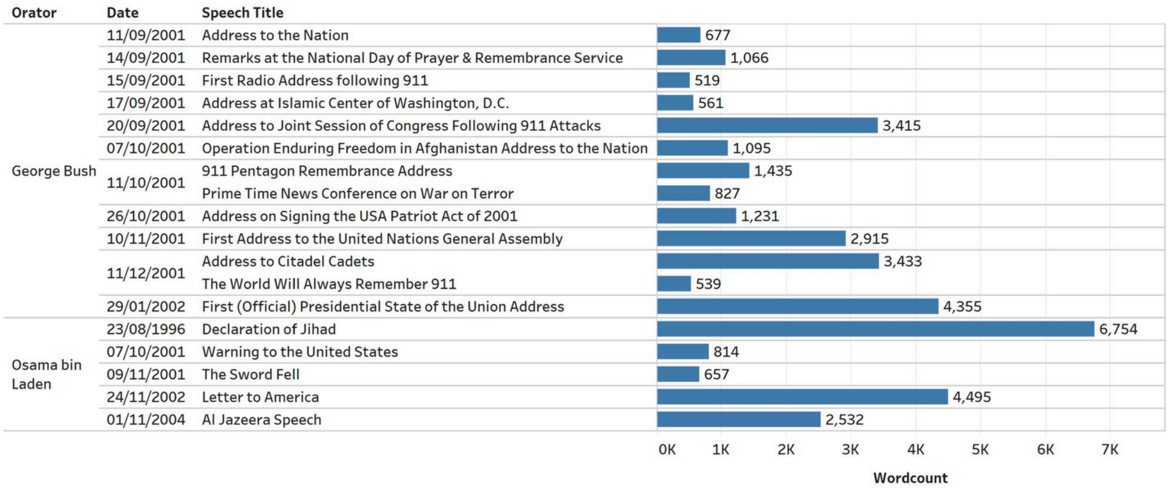



Figure 1 Texts and Word Count of Bush's and bin Laden's Declarations of War.

The 'War on Terror' provides a case study for the warfare genre of hostile narrative. The corpus comprises 18 speeches made by President George Bush and six publications by Osama bin Laden from 1996 to 2004. Bush's texts are in open source from the American Rhetoric website[12]; while bin Laden's are English translations from the 9/11 Memorial website[13] and the Guardian newspaper[14]. Figure 1 summarises these texts which are the focus for this thesis, and particular focus is given to Bush's 'Address to the Joint Sessions of Congress Following the 911 Attacks' in which he first declared the 'War on Terror'. The focus is on Bush's texts because this thesis is written from a similar Western position thereby addressing any potential for biases against bin Laden's proscribed terrorist organisation. Bin Laden's texts then provide a means to compare findings from Bush with a non-Western orator. While the analysis suggests no moral equivalence between Bush and bin Laden, Chapter 3 reveals striking similarities in how each orator legitimises violence despite being opponents of the same conflict. Indeed, this thesis reveals a functional equivalence in each orator's use of language for violence legitimisation.

[12] American Rhetoric (n.d.) George W. Bush Speeches, retrieved on 17th Feb 2023

[13] 9/11 Memorial (n.d.) Antecedents of 9/11, retrieved on 17th Feb 2023

[14] The Guardian (2002) Full text: bin Laden's 'letter to America', retrieved on 17th Feb 2023

The social identities of the ingroup and outgroup characterise each hostile narrative genre. Drawing upon the PRIO's definition of interstate and intrastate conflict[15], the governments and non-state actor groups of Bush's and bin Laden's texts characterise groups in the warfare genre. Jewish outgroups define the antisemitic genre, as represented in *Mein Kampf*. The outgroup of hate speech is generally a minority group, and Vidgen *et al.* (2020) provide a list of identities for each genre (Vidgen, Thrush, et al., 2020, p. 13). As such identities of race define genres of racism, gender defines sexism, class status defines classism, and so on.

This dataset is much smaller than the datasets used by current hate speech detection algorithms; however, the dataset size is irrelevant to the violence advocated by each orator. While there is no suggestion of causation, these texts represent the driving narratives of mass violence. Indicative of the violence linked to each narrative is data published by the PRIO. Eck and Hultman (2007) record 3000 deaths for the 2001 attacks in New York attributed to bin Laden (Eck & Hultman, 2007, p. 239), while Pettersson and Eck (2018) record 200,000 deaths by 2017 for the conflict in Afghanistan advocated by Bush (Pettersson & Eck, 2018, p. 537). The number of holocaust deaths for which *Mein Kampf* became a driving narrative is estimated to be millions[16]. Current algorithms require large datasets because of the quantitative methods they employ; conversely, this thesis draws upon qualitative methods to analyse smaller datasets. Any criticism of the small size of this dataset, therefore, does not account for the large-scale violence attributed to each and is preferring algorithmic methods over generating meaningful insights into how people promote hostility. Having applied narrative theory to hostile narratives, the chapter now turns to the methods for analysing hostile narratives.

## 1.4 What Are the Qualitative Elements of Analysing Hostile Narratives?

With no fixed approach for qualitative methods, the above subsection has established a theoretical basis of hostility using Galtung's theories of violence. This section now introduces the qualitative elements of hostile narrative analysis that are applied to the methodological framework of cultural violence presented in Chapter 3. Accordingly, the section begins by introducing the somewhat unhelpful tension between quantitative and qualitative methods. The section continues by explaining how the field of narrative analysis applies to the analysis of hostile narratives. The subsequent section then explains how this thesis aims to enable narrative analysis using the qualitative methods of semantic analysis, supported by pattern-based NLP.

---

[15] Uppsala Conflict Data Programme (n.d.) Definitions, retrieved on 17th Feb 2023

[16] Council of Europe (2023) Holocaust Remembrance, retrieved on 17th Feb 2023

While quantitative methods currently dominate NLP, there is a broader debate about the inclusion of qualitative approaches for a more hybrid approach to analysing language. A somewhat problematic tension between qualitative and quantitative methods began in the mid-twentieth century. ‘By the 1940s and 50s in sociology, psychology and some other fields, quantitative method (in the form of survey and experimental research) had become the dominant approach’ (Hammersley, 1992, p. 40). ‘The concept *qualitative research* started to spread during the 1950s and 1960s and became widespread in large parts of the social sciences during the 1970s and 1980s but in some disciplines such as psychology it did not gain momentum until the 1980s and 1990s’ (Allwood & Allwood, 2012, p. 1421). The tension is a ‘mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing fact’ (boyd & Crawford, 2012, p. 667). Yet, contrary to any claims of objectivity by quantitative researchers, all research tasks – especially the linguistic analysis – have interpretative elements and require a mixed approach of quantitative and qualitative methods.

Lindgren (2020) explains how quantitative methods generally analyse data tables with statistical tools, while qualitative methods typically involve close reading of textual data from interviews, observations and documents (Lindgren, 2020, p. 13). Quantitative methods are typically about the quantifying the collection and analysis of data using mathematical methods. The quantitative aspects of NLP are generally focus on the quantification of language. For example, latent dirichlet allocation (LDA) is a ‘generative probabilistic model’ of text often used for inferring the topics of a text (Blei et al., 2003, p. 996). Malik *et al.* (2022) provides a comprehensive review of 14 NLP models for hate speech detection, all of which use quantitative methods (Malik et al., 2022). Chapter 2 explains how text classification algorithms employ statistical models based on manually annotated training data. ‘Word embedding’ infers meaning from the mathematical relationships of ‘distributed representations of words in a vector space’ (see: Mikolov, Chen, Corrado, et al. 2013; Mikolov, Chen, Sutskever, et al. 2013; Mikolov, Yih, and Zweig 2013). While quantitative methods enable large-scale systems development, Chapter 2 explains how these methods provide questionable value for more social science applications like hate speech detection, sentiment analysis and hostile narrative analysis.

While quantitative methods dominate NLP, qualitative approaches are gaining provenance through the digital humanities. ‘Qualitative approaches are typically used to explore new phenomena and to capture individuals’ thoughts, feelings, or interpretations of meaning and process’ (L. Given, 2012, p. xxix). As such, Lindgren (2020) reflects on the growing integration of the social and computer sciences by proposing the idea of ‘Data Theory’ as a ‘broad label for a hybrid form of digital social science research practice that is data-intensive, computational (‘quantitative’), yet theoretically interpretative (‘qualitative’)’ (Lindgren, 2020). To qualitatively assess a text, linguists typically use semiotics, narrative, content, discourse, archival, and phonemic analysis, in addition to statistics, tables, graphs, and numbers (Denzin & Lincoln, 2018, p. 12). The growing field of Digital Humanities sees increasing integration of these tools with computational methods to understand social phenomena (see: Dunn and Schuster 2020; Orlandi 2021; van der Zwaan et al. 2017). As Chapter 3 explains, however, hate speech detection systems are yet to integrate qualitative approaches into their computational methods. Accordingly, the review from Malik *et al.* (2022) is only concerned with assessing the models from a quantitative perspective and does not consider the more qualitative aspects mentioned here. To fill this methodological gap, this thesis applies narrative theory, semiotics and semantic analysis to the methodological framework presented in Chapter 3.

The methodological framework connects different elements of cultural violence with a broad range of literature from across the humanities, including social psychology. The framework seeks to explain how violence legitimisation may occur through an orator's use of interacting 'ingroup elevation' and 'outgroup othering' statements to create what Galtung describes as a 'Self-Other gradient' between the subject and object of violence (Galtung, 1990, p. 302). The central feature of these statements is an orator's use of moral codes derived from religion and ideology to differentiate their ingroup from their outgroup. This methodological framework becomes the basis for explanatory dialogues about expressions of hostility in a natural language. The application of narrative theory to this framework conceptualises hostile narratives as stories and analyses them using language clauses; semantic analysis then applies to the computational methods of applying this framework.

The application of narrative theory to this framework for interpreting hostile narratives must account for how 'narrative truth' features in group stories. In general, where researchers study narrative as lived experience, narration itself is the practice of constructing meaningful selves, identities, and realities (Chase, 2011, p. 424). As such, a narrative is not necessarily a factual report of events, it is also an articulation that seeks to persuade others to see them in a particular way (Riessman, 2008, p. 187). Persuasion gives rise to a gap between 'historic truth', whose standard is accuracy, and 'narrative truth', which is judged against aesthetic criteria such as 'closure, coherence, and rhetorical appeal' (Hinchman & Hinchman, 1997, p. 1). The truth of a hostile narrative is most often prejudicial beliefs about a fictional other. As a *story* for legitimising violence, therefore, a hostile narrative should be seen as a fiction which plays out in the minds of the people who subscribe to the narrative truth.

Anderson's (2006) idea of imagined communities provides an example of how narrative truth manifests in the geopolitics of the hostile narratives in this this thesis. For Anderson (2006), an 'imagined political community' refers to the idea that nations and nationalism are social constructs that are created and sustained through shared cultural symbols, histories, and beliefs. Nations are 'imagined' because, although their members may never meet one another face to face, they share a sense of belonging and attachment to a collective identity. Anderson (2006) explains how this imagined community is maintained through the production and dissemination of cultural and ideological materials, such as maps, newspapers, and textbooks, that reinforce a shared sense of history and identity. Indeed, the extent of the 'deep horizontal comradeship' of these communities has motivated 'so many millions of people, not so much to kill, as willingly die for such limited imaginings' (Anderson, 2006, p. 7). Anderson's argument is that these imagined communities provide an origin for nationalism. For this thesis, nationalists communicate nationalism through hostile narrative.

Geopolitical imagined communities are prevalent in the hostile narrative corpus whereby each orator caricatures the greatness of their ingroup in contrast the villainy of their outgroup. Bush's characterisation of the Taliban, however, shows how these caricatures may conflict with historic truth. Bush caricatures the Taliban as a homogenous terrorist regime, whereas Simpson's (2012) analysis of the Afghan conflict explains them more as a franchise of different tribes (Simpson, 2012, pp. 77–78). To make his point, Simpson (2012) uses ethnographic studies to explain how Pashtun's (the primary ethnic group of the Taliban) have different senses of themselves. The Taliban-e jangi or Taliban-e shuri refers to the fighting or insurgent Taliban; Taliban-e darsi (madrassa students) refers to those who are students and not fighters; Taliban-e pak (clean Taliban) refers to honest individuals who are committed to Islamic principles of justice; Taliban-e duzd (the thief Taliban) refers to local bandits; and Taliban-e khana-neshin (Taliban sitting at home) refers to those who are inactive and associated with the 1990s Taliban. The homogenous representation of the Taliban in Bush's narrative truth represents a Western consensus, which is in stark contrast to the lived experience of members of each Taliban tribe. With an understanding of the role of qualitative and quantitative methods in hostile narrative analysis, the chapter continues with how they are applied with NLP.

## 1.5 What Is Natural Language Processing?

As Chapter 2 explains, NLP is a subfield of AI that seeks to give machines the ability to understand humans. The question of whether machine can understand humans has long been a philosophical question in AI. Turing (1959) first raised this question in his thought experiment, originally known as 'The Imitation Game', now more commonly referred to as the Turing test. This experiment is about whether a human subject can distinguish between another human and a machine imitating a human. If a human subject is unable to distinguish between another human and a machine in a natural language dialogue, the machine is said to exhibit artificial intelligence. Among many subsequent thought experiments that both develop and challenge Turing (1959), Searle's Chinese Room experiment argues that machines cannot understand humans because the intentionality of a symbol is 'solely in the minds of those [humans] who program them and those who use them, those who send in the input and those who interpret the output' (Searle, 1980, p. 422). The question for a sociotechnical system, therefore, is whether machines can indeed understand humans, or whether machines promote *human* understanding?

A general theme emerging from experimenting with NLP algorithms for social science applications is a tension between quantitative and rule-based methods for processing natural language. The field of NLP began in the 1950s using rule-based systems and has since evolved to employ sophisticated quantitative methods. Defining and managing rule is laborious; therefore, in the 1990s, which saw rapid adoption of the internet, 'large amounts of data became available, which enabled statistical learning methods to work on NLP tasks' (Zhou et al., 2020). Indeed, Manning and Schütze (1999) records the evolution of applying statistical methods to NLP applications (Manning & Schütze, 1999). The state-of-the-art now draws upon complex neural AI networks to process natural language (Malte & Ratadiya, 2019). Despite the popularity of these quantitative methods, however, there is something of an inconsistency between established linguistic theory and NLP literature that arises from the treatment of text as structured or unstructured data.

NLP researchers generally conceive text as 'unstructured data' while referring to grammatical structures as latent or hidden (Bengfort et al., 2018, p. 8; D.C et al., 2021; Feldman & Sanger, 2007, p. 3; Kanchinadam et al., 2021; Resende et al., 2021). NLP literature refers to 'natural language' itself as the collection of words used by humans for everyday communication captured in text, speech and audio data (Bird et al., 2009, p. ix; Patel & Arasanipalai, 2021, p. 5). Indeed, Bird, Klein and Loper (2009) describe the processing of natural language as converting 'the unstructured data of natural language sentences into the structured data' (Bird et al., 2009, p. 262). In treating text as unstructured data, NLP algorithms tend to treat text as a sequence of words over which NLP algorithms iterate. As such, they analyse words in the order in which they appear in a sentence. Yet, treating text as unstructured data is inconsistent with how linguists study language.

In contrast to treating text as unstructured data, however, linguistics generally treats a 'Language' as a structured system of words. De Saussure defined Language (langue) as 'a system of signs that express ideas' and introduced *Semiotics* as a science to 'investigate the nature of signs and the laws governing them' (Saussure, 1916, p. 16). Semiotics has since become a theory of meaning (or 'signification') that depends on the network of relationships between linguistic expressions in the minds of speakers of a speech community (Knapp et al., 2008, p. 231). The idea of Universal Grammar, mainly attributed to Chomsky, is about discovering a universal set of rules governing language generation. Indeed, in his 1957 thesis that has become a foundation for modern linguistics. Chomsky even rejected the Markov method of statistically generating language that is the current NLP state-of-the-art. As linguistic theory explains, the nodes of a linguistic system are words and the connections between them are their grammatical properties.

For this thesis, each hostile narrative genre has a unique Language whose nodes comprises both common and unique words and tropes. As Galtung (1983) observes, 'almost every linguistic act defines attributes and relations, meaning that the spoken and written language is not only a social act in the usual sense of being interactive, but in the sense of pointing out, underlying, even reinforcing social divisions and relations' (Galtung & Njshimura, 1983, p. 20). These linguistic acts signal the particular Language of a group. From the perspective of evolutionary biology, Martin (2018) suggests all languages have the basic grammatical properties reflecting underlying neural structures (Martin, 2018, p. 34). Accordingly, a Language reflects how different groups rationalise the world. For different genre, certain racial epitaphs imply racist genres, while such tropes as 'send them home' imply the anti-immigration genre. Each linguistic system is then parsed by detecting the language clauses connecting each word.

This thesis responds to this tension between quantitative and qualitative approaches by seeking to develop a hybrid approach analysing language. The computational methods outlined in Chapter 4 draws upon quantitative methods for labelling the lexical properties of words. Word labelling then enables a pattern-based approach for processing hostile language. The qualitative method firstly draws upon cultural violence theory as a methodological basis and narrative inquiry as a basis for explanation. The specific language of a potentially hostile text is then analysed through the application of semantic analysis, as applied through pattern-based NLP. This approach treats language as a system to understand the interaction of clauses in violence legitimisation. The relevant clauses then become the evidence for attributing hostility to an orator. And the use of language clauses in the computational methods gives consistency between how humans and machines process language. In effect, hostile narrative analysis becomes about detecting the narrative structures used in violence legitimation.

## 1.6 Connecting The Social and Computer Sciences

The re-occurring idea of theme of this thesis is connecting the computational methods of NLP with peace research to provide theoretical and technical contributions towards tackling online abuse. These sociotechnical contributions are as follows:

1. Propose hostile narrative analysis to rethink how to tackle online abuse.
2. Develop cultural violence as a guiding theory for analysing hostile narrative.
3. Propose a novel methodology for the process of analysing a hostile narrative.
4. Make available reproducible experiments for assessing general-purpose and state-of-the-art NLP technologies.
5. Hybrid-NLP for applying semantic analysis to the analysis of hostile narratives.
6. Explainable AI for meaningful analysis of hostile narratives.

In developing a hostile narrative analysis, a synthesis between the social and technical aspects of NLP forms the core of transdisciplinarity in this thesis. Transdisciplinarity contrasts with interdisciplinarity by how the research uses the constituent disciplines. Interdisciplinarity combines existing disciplines to conduct research, whereas transdisciplinarity seeks to modify them with new research contributions. As such, this thesis seeks to contribute to peace research by developing cultural violence theory to detect processes of violence legitimisation in natural language. In turn, this methodology contributes to Web science as a new way to conceptualise how to detect hate in online platforms. The thesis contributes to computer science with by using pattern-based NLP to show the value of treating text as structured data. A series of reproducible experiments to determine modifications to NLP technologies provide the basis for developing the computational contributions. With a focus on methodology and method, these contributions are as much about an approach to developing social science applications with NLP as much as the specifics of hostile narrative analysis.

In addition to a Web Science contribution, this paper also seeks to contribute to the field of PeaceTech with new computational methods to detecting cultural violence. The idea of Peacetech emerged in 2015 with ‘the creation of the United States Institute of Peace's (USIP) Peacetech Lab as an umbrella term for a focus on new information and communication technologies and their role in building peace’ (Rhian, 2018, p. 13). Accordingly, this theis is written from a practitioner’s perspective to inform peace researchers about applying computational methods to peace-building applications; for the technical audience, the contributions seek to explain the value of hybrid-NLP for peace research.

The practitioner’s perspective very much motivates the author’s development of hostile narrative analysis. I was previously an Infantry Officer in the British Army, during which time I served in violent conflicts in Northern Ireland, Iraq and Afghanistan. A central feature of these conflicts is how different groups legitimise violence against each other. Central to my own reflections on these conflicts is how I, and the society I served, legitimised our participation in them. Legitimising these conflicts then features in the prevailing hostile narratives of the time. Indeed, I have an innate understanding of the War on Terror texts in the hostile narrative corpus having lived through the violent actions they promoted, which is why I chose to study them in this thesis. The comparison of Bush's and bin Laden's texts reflects my general observation about a commonality between how humans legitimise violence, regardless of which side of a conflict they serve, which group they identify with, or their chosen moral framework.

During the research for this thesis, I realised a similarity between the narratives of how I legitimised violence as an Officer of the British Army and the legitimisation of violence against minorities through hate speech. This research includes a hate speech annotation project for the Alan Turing Institutes that produced two research papers (Vidgen, Botelho, et al., 2020; Vidgen et al., 2021). Recalling Galtung's point about moral colour, I make no suggestion of moral equivalence, more one of functional equivalence in legitimising violent action. Indeed, the different moral frameworks around the perpetration of violence on behalf of the state and on behalf of oneself are fascinating; pacificists may even suggest they are morally equivalent. Nonetheless, in researching hate speech I observe the same linguistic techniques in online abuse as with legitimising warfare. As explained across this thesis, these linguistic techniques centre on elevating the imagined greatness of an ingroup and the villainy of an outgroup. Current hate speech detection, however, does not account for this elevation and othering. As Chapter 3 explains, the general perspective of hate speech research is victim-focussed and does not sufficiently consider the alleged perpetrator's intent. I also seek to tackle hate speech but with a contrasting perpetrator focus as someone who has previously legitimised violent action.

To the best of my knowledge, this research constitutes the first attempt to create computational methods for analysing hostile narratives using cultural violence. Gleditsch et al. (2014) observe that cultural violence 'never caught on in mainstream Peace Research' since its introduction in Galtung (1990) but give no real explanation as to why (Gleditsch et al., 2014, p. 150). Moreover, problems with existing detection systems seemingly arise from a 'technical first' approach before applying social theory; this thesis's sociotechnical approach begins with social theory to guide technological development. By adopting this approach and by connecting peace research with NLP, this thesis seeks to support and continue the transdisciplinary work of Web Science. Having now introduced the main theories underpinning hostile narrative analysis, the thesis now continues with a review of NLP technologies for social science applications.

# Chapter 2 The Promise and Limitations of Quantitative Methods in NLP For Social Science Applications

This chapter addresses the following research question, ‘to what extent do quantitative methods in NLP ‘understand’ social science applications?’. Many descriptions of NLP imply a promise of giving machines the ability to ‘understand’ human language. Accordingly, Hirschberg and Manning (2015) suggest, NLP ‘is the subfield of computer science concerned with using computational techniques to learn, *understand*, and produce human language content’ (Hirschberg & Manning, 2015, p. 261, emphasis added). Bird et al (2009) provide a similar description in a textbook that popularised their open-source Natural Language Toolkit (NLTK) python library. They take NLP, ‘in a wide sense to cover any kind of computer manipulation of natural language’, whether simple tasks such as counting word frequencies or ‘*understanding* complete human utterances’, to the extent of generating at least ‘useful responses’ (Bird, Klein, and Loper 2009: ix, emphasis added). As such, this thesis recognises the promise of quantitative methods in NLP to provide transformative value for functional applications like chatbots and natural language translation.

In contrast, this chapter finds limitations with quantitative approaches to NLP for ‘understanding’ the language of social science applications like hate speech detection and sentiment analysis. This chapter comprises a series of experiments with quantitative NLP methods using the hostile narrative corpus. Since each text of the corpus has a generally accepted meaning, the experiments assess the extent to which the output of each algorithm represents that accepted meaning. The first section establishes the idea of the theoretically perfect text classifier that synthesises text classification with encoding and decoding theory from Hall (1974). This theory provides a way to explain how humans understand language. This perfect classifier, therefore, represents an algorithm that fulfils the promise of understanding human language. The remainder of the chapter then uses encoding and decoding to assess the effectiveness of the quantitative methods of word embeddings and text classification for hate speech detection and sentiment analysis. These experiments reveal the limitation of NLP for social science applications as the treatment of text as unstructured data for quantitative methods.

While the literature technically assesses NLP algorithms using a confusion matrix, this sociotechnical assessment focuses more on the operational domain by questioning what each algorithmic output explains for an explanatory dialogue. As such, a technical assessment provides a quantitative view of what constitutes a well-architected algorithm by giving insight into how well it performs in relation to the training data. The confusion matrix, however, does not provide insight into operational utility of an algorithm for an explanatory dialogue. This assessment contains two operational assessments which compare *Mein Kampf* with *I Have a Dream* for hate speech and Bush's and bin Laden's declarations of war for sentiment. In contrast to technical scoring, these operational assessments are more subjective; they rely upon a degree of professional judgement and background knowledge of each text. This bridge from the technical to the operational domain represents further research for assessing NLP models.

This chapter synthesises encoding and decoding theory from Hall (1974) with NLP as one way to assess how well machines can understand humans. In questioning whether machines can understand humans, there is no doubt that quantitative methods in NLP add transformative value for functionally defined applications since such tasks have a narrow scope of interpretation. Encoding and decoding theory, in contrast, applies to social science applications that have a much wider scope of interpretation by audiences. This theory has become very influential in linguistics, particularly among British theorists (see: Chandler, 2005); nevertheless, it does not seem to feature in NLP literature. As will be explained, encoding and decoding in NLP literature tends to apply a mathematical interpretation from Shannon and Weaver (1948). This chapter, therefore, is an assumed novel attempt to synthesise Hall's encoding and decoding theory with NLP.

## 2.1 How Do Encoding and Decoding Apply to Social Science Applications?

This first section introduces the idea of the 'perfect classifier' as a way to assess the promise and limitation of quantitative methods for social science applications. The section introduces text classification as a popular quantitative method for hate speech detection and sentiment analysis. The section then uses the confusion matrix to present a theoretically 'perfect classifier', which represents a machine understanding of natural language. The section then explains the subjective element of social science applications using Hall's theory of encoding and decoding. While actual definitions of 'meaning' can be deeply philosophical, encoding and decoding have practical utility for explaining the subjective interpretation of words. The section finishes with a synthesis of the perfect text classifier with encoding and decoding to suggest the limitation of text classification for social science applications.

### 2.1.1 What Is Text Classification?

Text classification is a quantitative method that draws upon statistical techniques to predict text classifications using a pre-labelled training dataset. One of the earliest examples of text classification is from Maron (1961), who experimented with statistical techniques to automatically classify documents according to content. Maron based his classifier on 'the rather straightforward notion that the individual words in a document function as clues, on the basis of which a prediction can be made about the subject category to which the document most probably belongs.' (Maron, 1961, p. 405). In more technical terms, if $d_i$ is a document of the entire set of documents D and $\{c_1, c_{2, \ldots,} c_n\}$ is a set of all the [classifications], text classification then assigns one category $c_j$ to a document $d_i$' (Ikonomakis et al., 2005, p. 966). D might be a blog, social media post or newspaper article, and $d_i$ is a lexical element of D, whether a paragraph, sentence, phrase or word (note, 'text' has since replaced 'document' and is used in the remainder of this thesis). As noted in four literature reviews, text classification is a common method for hate speech detection (Fortuna, 2018, p. 22; Kovács et al., 2021, p. 4; Mullah & Zainon, 2021a, p. 88364; Schmidt & Wiegand, 2017, p. 2) and sentiment analysis (see: Basiri et al., 2021; Beniwal & Maurya, 2021; Dowlagar & Mamidi, 2021; Mansour, 2018; Stappen et al., 2021; Zhang et al., 2021).

As Maron (1961) notes, text classification (which he refers to as automatic indexing) concerns 'the problem of deciding automatically what a given [text] is 'about' (Maron, 1961, p. 404); as such, the classification set defines the aboutness – or topic – of a text. Sentiment analysis generally applies a classification set of {positive, negative, neutral}. Hate speech classifiers most commonly use binary classifications of {hateful, non-hateful} (Burnap & Williams, 2015a, p. 231; Fortuna, 2018, p. 22; Mullah & Zainon, 2021b, p. 1; Poletto et al., 2020, p. 497), and less often multi-classifications, such as {identity-directed abuse, affiliation-directed abuse, person-directed abuse, non-hateful slurs, counter speech} (Vidgen et al., 2021, p. 2291) or {hate speech, offensive but not hate speech, neither hate speech nor offensive speech} (Abro et al., 2020, p. 486). As a simple example, the word 'parasite' might attract a label of {hateful}, therefore, a classifier would likely classify a sentence containing this word as {hateful}.

Modern text classification relies upon machine learning architectures to assign the classification of an input. The training dataset feeding a text classifier is pre-labelled with the desired classifications - $\{c_1, c_{2, \ldots,} c_n\}$ - for each constituent element of a representative corpus of texts (Bird et al., 2009, p. 222; Géron, 2017, p. 20). For sentiment analysis, developers often use reviews either from the Internet Movie Database (IMDB)[17] or Amazon Marketplace[18] as training data. These review-based datasets are useful as they contain a reviewer defined score for the product in questions. This score provides a quantifiable metric for scoring the accompanying text. For hate

---

[17] Papers with Code (n.d.) Sentiment Analysis on IMDb, retrieved on 17th Feb 2023

[18] Papers with Code (n.d.) Sentiment Analysis on Amazon Review Full, retrieved on 17th Feb 2023

speech detection, Kosisochukwu *et al* (2020) and Fortuna *et al* (2020) identify several human-annotated training datasets that are publicly available for research (Fortuna et al., 2020; Kosisochukwu et al., 2020). These datasets have a range of classification schemas.

The labelling of training data for each classification draws upon either supervised learning, unsupervised learning or semi-supervised learning (Géron, 2017, pp. 26–33). Human annotators manually label the training data according to the classification set for supervised learning. Developers use either crowdsourcing (Sabou et al., 2012; Shmueli et al., 2021) or a small group of expert annotators for annotation (Guest et al., 2021; Vidgen, Botelho, et al., 2020) . Röttger *et al*. (2021) observe how human annotation introduces high degrees of subjectivity into the dataset for what this thesis calls social science applications (Röttger et al., 2021, p. 1). In contrast, unsupervised learning algorithms analyse unlabelled datasets without human intervention. Semi-supervised learning combines unsupervised learning with human annotators. Where supervised learning is the most common approach for hate speech dection, the algorithm predicts an inputs classification based on a similarly annotated element of the training data.

### 2.1.2 What Is the Perfect Text Classifier?

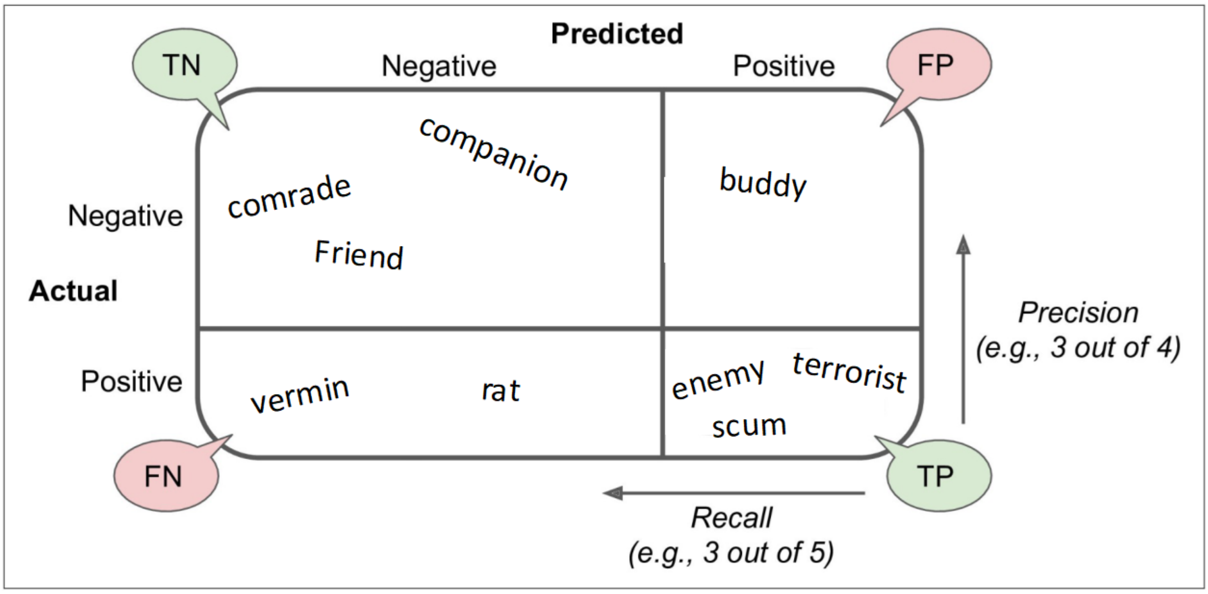



Figure 2. A confusion matrix of hateful terms (adapted from Géron 2017).

As a technical assessment, the confusion matrix evaluates how well a text classifier correctly classifies inputs compared to its training data. Figure 2 is a confusion matrix for a hate speech classifier featuring words from the hostile narrative dataset and {hateful, non-hateful} classifications. The evaluation comprises four metrics: true positive, true negative, false positive and false negative. A true positive (TP) is when a classifier correctly classifies an input in the positive class; for hate speech detection, a hateful term – enemy, scum, terrorist – is classified as hateful. Conversely, a true negative (TN) is for terms belonging to the negative class; for hate speech, a non-hateful input – comrade, companion, friend – would be classified as non-hateful. A false positive (FP) is an incorrect classification of a true positive; for hate speech, a non-hateful input – buddy – is classified as hateful. Finally, a false negative (FN) is an incorrect classification of a true negative, which means classifying a true negative as non-hateful – vermin, rat. 'A perfect classifier would have only true positives and true negatives' (Géron, 2017, p. 87) and would represent an algorithmic understanding of natural language.

Hate speech literature generally evaluates a classifier's performance using the f1 score, which summarises the confusion matrix as the harmonic mean of precision and recall (Beniwal & Maurya, 2021, p. 472; Chandra et al., 2021, p. 7; Chiril et al., 2022, p. 333; Mullah & Zainon, 2021c, p. 88370; Vidgen, Botelho, et al., 2020, p. 7; Vidgen et al., 2021, p. 2297). *Precision* is the proportion of correctly predicted positives to all positives, while *recall* is the fraction of known positives (Kowsari et al., 2019, p. 45). Precision assesses the classifier's accuracy when making a prediction; in Figure 2, the precision shows how the classifier made three out of four correct predictions. Recall evaluates the classifier's ability to sift the negative from the positive; the recall in Figure 2 is three out of five predictions. In this calculation, high recall would correctly classify hateful terms but at the cost of low precision, which would see many non-hateful terms incorrectly classified. Conversely, low recall systems would miss many hateful terms, but those detected would be correctly classified with high precision. In practice, the extent to which text classification understands hate speech is a trade-off between precision and recall. To summarise, Maron (1961) initially intended text classification to classify what a text is *about*, as is now explained, however, encoding and decoding provide a way to understand what a text *means*.

### 2.1.3 How Does Hall's Theory of Encoding and Decoding Apply to the Perfect Classifier?

Hall's (1974) theory of encoding and decoding provides a way to understand a text's connotative and denotative meaning for the perfect classifier. As Murdock (2016) explains, Hall (1974) responds to Shannon's and Weaver's 'The Mathematical Theory of Communication', a volume that 'quickly became an obligatory point of reference for scholars working in the emerging field of information studies' (Murdock, 2016, p. 1). Shannon and Weaver (1948) viewed communication as an engineering problem and sought to maximise the efficiency of broadcasting messages in a transmission system (Shannon, 1948, p. 2). By treating communication as an engineering problem, they applied a fixed interpretation of messages between two sides of communicative exchange. Correspondingly, Shannon (1948) applies what this thesis refers to as an assumption of fixed interpretation between an orator and an audience. This assumption of fixed interpretation assumes that both orators and recipients share the same interpretation of a text, which removes the subjective element of understanding language.

Hall's theory of encoding and decoding directly challenged this assumption of fixed interpretation by understanding how words are encoded by orators and differently decoded by their audiences. The term 'code' features in both the technical aspects of information theory and linguistic aspects of communication theory. The technical meaning of code is about finding methods to efficiently encode and decode messages within an information system, software code being one example. In contrast, Eco (1976) describes linguistic codes as systems that link 'sign-vehicles', such as words or images, to 'semantic units' of meaning (Eco, 1976b, p. 67). As Chandler (2005) notes, semantic units of meaning refer to 'interpretive frameworks which are used by both producers and interpreters of texts...to simplify phenomena in order to make it easier to communicate messages' (Chandler, 2005). For example, the words of the confusion matrix in Figure 2 are codes of a hostile narrative that orators use to communicate to their audiences; each is encoded with meaning that in some cases promote hostility.

According to encoding and decoding theory, connotation and denotation define the interpretative frameworks for understanding a text. The denotative (literal) meaning of words is generally fixed, but their connotative (associative) meaning very much depends on the audience (S. Hall, 2006, p. 168). As Davidson *et al* (2019) note, for example, 'nigger' in racist speech 'can be extremely racist or [routine], depending on the speaker, the context, and the spelling' (Davidson et al., 2019, p. 33). The word denotes a black or dark-skinned person, whereas it connotes either racist intention towards black communities or friendship within those communities depending on the orator-audience relationship. In its friendly connotative meaning, this word is one example of 'reclaimed words', which are once pejorative terms that the target communities have redefined (Vidgen et al., 2021, p. 2293). While a single word may have fixed denotative meaning, therefore, an orator encodes it with connotative meaning, which is then decoded by an audience with the same or different interpretation.

According to Hall, denotative and connotative interpretations of words depend on one of three orator-audience relationships that he calls the dominant-hegemonic, negotiated, and oppositional positions. The dominant-hegemonic position is where an orator and their audience share the same connotative meaning of a message (S. Hall, 2006, p. 171). The second two positions, however, challenge the hegemony of the dominant code. In the negotiated position, a recipient's interpretive framework 'contains a mixture of adaptive and oppositional elements' whereby the general meaning of the hegemonic code is accepted but with a recipient's own ground rules (S. Hall, 2006, p. 172). The oppositional position is when an audience understands the denotative and connotative meanings but decodes a word within a contrary frame of reference (S. Hall, 2006, p. 173). The subjective element of decoding is in 'the ability of audiences to produce their own readings and meanings, to decode texts in aberrant or oppositional ways, as well as the 'preferred' ways in tune with the dominant ideology' (Durham & Kellner, 2006, p. 95).

Regarding hostile narratives, commentators commonly used metaphor to communicate the dominant hegemonic position of the War on Terror. Regarding hostile narratives, commentators commonly use metaphor to communicate the dominant hegemonic position of the War on Terror. Steuter and Wills (2010) link the US military's use of linguistic codes for the War on Terror to the metaphors of infestation, cancer, corruption, and decay used by the Third Reich. The purpose is 'to dehumanise its hated Others, those disenfranchised from citizenship and ultimately from humanity itself' (Steuter & Wills, 2010, p. 153). In contrast to dehumanising the other, Gregory (2010) also observes how such pathological metaphors in warfare 'make military violence appear to be intrinsically therapeutic' whereby soldiers become the surgeons – the heroes – who kill insurgents to save the body politic (Gregory, 2010, p. 277). These metaphors denote hunting and disease; when applied to humans, however, they connote a 'language of annihilation, eradication, and extermination', which scholars identify as 'classically propagandistic language' that precedes and enables genocide (Steuter & Wills, 2010, p. 163).

'Dog whistle politics' provides a more subtle example of how connotative meaning features in expressions of racism. 'Dog whistle politics' are 'coded racial appeals that carefully manipulate hostility towards non-whites' (Haney-López, 2014, p. IX). Regarding promoting hate, dog whistles allow 'politicians to speak about taboo subjects while retaining plausible deniability that they violated any social norms' (Drakulich et al., 2020, p. 372). Dog whistles have an everyday denotative meaning but connote hostile intent towards racial groups. For example, humour and sarcasm are types of dog whistles that gain mention in some, but not all, hate speech detection papers (Agarwal & Chowdary, 2021, p. 2; Alrehili, 2019, p. 5; Fortuna, 2018, p. 85). Another prescient example is the concept of Free Speech that often features in hate speech discourse. While Free Speech denotes a human right, some commentators argue that it has taken on racist connotations. As such, Titley (2020) argues that the idea of Free Speech itself has been 'retooled as a technology for racist amplification' (Titley, 2020). Whether or not these racist connotations are valid, such concepts as Free Speech have contradictory connotative meanings to different audiences that either promote a human right or dog whistle hate speech.

Examples from the hostile narrative corpus also show how the narrative truth of both hate speech and declarations of war is a contest for the dominant-hegemonic position. Luther King's *I Have a Dream* speech challenges the dominant and hegemonic position of racism against black Americans during the 1960s. Luther King challenged the hegemony of this position by invoking a promise contained in America's Constitution and Declaration of Independence: 'all men — yes, Black men as well as white men — would be guaranteed the unalienable rights of life, liberty and the pursuit of happiness'. The words 'life, liberty and the pursuit of happiness' are part of a linguistic system contained in America's constitution that includes 'Justice', 'domestic Tranquillity' and 'the Blessing of liberty'[19]. These words constitute a system of codes that connote a sense of righteousness. In Hall's terminology, Luther King argues that the hegemonic position of inequality promoted by 'vicious racists' in 1960s America was, in fact, in opposition to what should be a dominant position of equality enshrined in America's constitution.

While there might exist a narrative truth on the successes of ending racism in America, there is a more likely historical truth that lies in the negotiated position. Concerning Luther King, Bonilla-Silva and Dietrich (2010) suggest 'a mythology that emerged in post-civil rights America has become accepted dogma among whites with the election of Barack Obama: the idea that race is no longer a central factor determining the life chances of Americans' (Eduardo Bonilla-silva & Dietrich, 2011, p. 191). This mythology – or narrative truth – signals the dominant-hegemonic code of equality in which race is no longer a determining factor of life chances. Mondon and Winter (2020) contest this 'post-racial' myth to describe a negotiated position of 'liberal racism', which accepts the dominant-hegemonic codes of equality while uncritically accepting other dog whistles that subtlety reinforce inequality. They argue that this negotiated position has created a space to mainstream far-right sentiment by pitching 'racist notions and ideas without explicitly naming them as such' (Kapoor, 2021, p. 2395). In contrast to a narrative truth, while people may claim they take the oppositional position to racism, they are more likely to take one of many negotiated positions.

### 2.1.4 What Challenges Do Encoding and Decoding Present to The Perfect Classifier?

To support any claim of understanding natural language, the perfect classifier in social science applications should incorporate connotative meaning into its classification decision. Nevertheless, encoding and decoding challenge the utility of text classification for social science applications in that one word connotes different meanings to various people. Words have a generally fixed denotative meaning for their representation of real worlds objects and ideas. Nevertheless, encoding and decoding explain how orators often encode these words with different connotative meanings. Connotative meaning in hate speech is how a word implies malicious intent towards a person or group. For sentiment analysis, connotative meaning is whether a word projects positivity, negativity or neutrality. Interpretation within a communicative exchange, therefore, is

---

[19] Archives.gov (n.d.) The Constitution of the United States, retrieved on 17th Feb 2023

how an audience decodes a text in accordance with the orator's or their own interpretative frameworks. The challenge is that in the negotiated or oppositional position, words may connote an entirely different meaning to the position an orator sought to communicate.

A second challenge for the perfect classifier concerning encoding and decoding is distinguishing between what a text is about and what it means. Following Maron (1961), text classifiers can detect what a document is about by the occurrence of certain words in a text. For example, reclaimed words can appear in either hate speech or criticisms of hate speech, and the overall topic of hate speech would be correct despite the contrasting positions of each document. Hate speech detection and sentiment analysis, however, seek to reveal what a text means. As discussed, reclaimed words connote racist intent in a dominant-hegemonic position of equality but can also connote friendship in a negotiated position within black communities. Hunting and disease metaphors also connotate obviously hostile meanings when applied to humans. The connotative meaning of dog whistles, however, is much more subtle but should still be detectable by the perfect classifier. The aboutness of text is a functional interpretation, but to understand meaning requires knowledge of the orator's and audience's interpretative frameworks. The next sections reveal these two challenges through experiments for social science applications.

## 2.2 How Effectively Do NLP Methods Encode Natural Language?

Having introduced a theoretically perfect text classifier in the first section, the following sections draw upon Hall's theory of encoding and decoding to assess text classification in practical terms for sentiment analysis and hate speech detection. The first section covers the standard practices in NLP for encoding a text. This section provides a general observation about pre-processing problems, which leads to a misrepresentation of an orator's use of words. This section then continues with an experimental review of word embeddings for encoding the meaning of a text. The subsequent sections experimentally assess the extent to which state-of-the-art research models and production systems can then decode hate speech and sentiment, respectively. All the experiments are available online and aim to be reproducible[20]. They reveal problems with decoding connotative meaning using quantitative methods for social science applications.

Note that these experiments purposefully do not assess NLP for denotative meaning. Accordingly, this thesis accepts NLP's transformative potential for the following tasks:

- Parts of speech tagging

[20] Anning (2023) Quantitative Methods, retrieved on 7th Jan 2023.

- Dependency Parsing
- Lemmatisation
- Named entity recognition
- Co-reference resolution

Each task is classified here as a functional task that seeks to encode and decode denotative meaning for word labelling. Indeed, each tasks forms part of the hostile narrative method as reviewed in Chapter 4. While these tasks are mainly descriptive, there is a degree of subjectivity, which means the denotative meaning is subject to some interpretation. Nevertheless, subjectivity in decoding a text does not rely on differences in connotation for the dominant, oppositional, or negotiated positions.

| | Author | datestamp | title | word count | sentence count | Watson Sentiment Score | Google Sentiment Score |
|---|---|---|---|---|---|---|---|
| **0** | Hitler | 2020-06-30 | Mein Kampf | 706100 | 4527 | 0.37 | -0.23 |
| **1** | King | 1963-08-28 | I Have a Dream | 9161 | 93 | 0.36 | -0.10 |

Figure 3. Document Sentiment Scores for *Mein Kampf* and *I Have a Dream*

Having reviewed methods for encoding text, the chapter continues with experiments for assessing the effectiveness of a hate speech detection and a sentiment analysis algorithm for decoding text. While these algorithms can process each text at the document level, the experiments process text at the sentence to best understand limitations with the current computational methods. The overall document scores for sentiment, as shown in Figure 3, produce nonsensical results. IBM's Watson produces similarly positive scores for *Mein Kampf* and *I Have a Dream*, while Google generates negative scores for both. These texts are expressions of starkly negative and positive sentiment and should not generate similar scores. The discussion for each experiment will explain how each algorithm processes text by the occurrence of a word, rather than by the grammatical relations between words, thereby skewing the orator's intended meaning. Processing the texts at the sentence level enables the identification of the words to explain these counter-intuitive outputs. The section now reviews pre-processing methods for encoding natural language.

### 2.2.1 What Are the Pre-Processing Methods for Encoding Natural Language?

This next section assesses the general ideas behind how NLP encodes natural language using word embeddings. The section begins by reviewing generally accepted approaches to pre-processing a text for conversion to a word embedding. This section will explain how pre-processing can skew training data by misrepresenting an orator's intended use of words. The section then continues with an experiment with the first word embedding algorithm, word2vec. The point of reviewing this algorithm is to convey how the state-of-the-art has progressed with transformer architectures that the following sections evaluate. While technical architectures are now undoubtedly more sophisticated, this comparison between old and new algorithms reveals the elements that have not changed. The section now begins by reviewing standard pre-processing practices.

Text pre-processing is about encoding a text into meaningful units for onward processing by an NLP algorithm; these meaningful units are the linguistic codes of natural language that a classifier seeks to process. Two essential steps for pre-processing are tokenisation and noun chunking. 'Tokenisation is the segmentation of a text into basic units—or tokens—such as words and punctuation' (Bird et al., 2009, p. 121). The NLTK NLP library offers a commonly used tokeniser that segments words by several techniques, including the whitespace between words and regular expressions[21]. Token combinations are often referred to as *ngrams*, whereby a *unigram* combines one token, a *bigram* combines two, and so on (Bengfort et al., 2018, p. 14). Noun chunking is a term specific to the spaCy NLP python library that involves combining grammatically related tokens to create noun phrases (Patel & Arasanipalai, 2021, p. 15). Any NLP algorithm must correctly tokenise a text to properly encode the meaning of the words and noun phrases used by an orator. As this section will explain, nevertheless, current methods to pre-process prepositional noun phrases and conjunctions misrepresent an orator's intended use of words.

```
Original Sentence:
Sentence 1: Al Qaeda is to terror what the mafia is to crime.
Sentence 2: Deliver to United States authorities all the leaders of al Qaeda who hide in your land.
Sentence 3: But today, for al Qaeda and the Taliban, there is no shelter.
Sentence 4: On my orders, the United States military has begun strikes against Al Qaeda terrorist training
camps and military installations of the Taliban regime in Afghanistan

Tokenised Sentence:
Sentence 1: ['Al', 'Qaeda', 'is', 'to', 'terror', 'what', 'the', 'mafia', 'is', 'to', 'crime', '.']
Sentence 2: ['Deliver', 'to', 'United', 'States', 'authorities', 'all', 'the', 'leaders', 'of', 'al', 'Qae
da', 'who', 'hide', 'in', 'your', 'land', '.']
Sentence 3: ['But', 'today', ',', 'for', 'al', 'Qaeda', 'and', 'the', 'Taliban', ',', 'there', 'is', 'no',
'shelter', '.']
Sentence 4: ['On', 'my', 'orders', ',', 'the', 'United', 'States', 'military', 'has', 'begun', 'strikes',
'against', 'Al', 'Qaeda', 'terrorist', 'training', 'camps', 'and', 'military', 'installations', 'of', 'th
e', 'Taliban', 'regime', 'in', 'Afghanistan']

Sentences Without Stopwords
Stopword examples ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
Sentence 1: ['Al', 'Qaeda', 'terror', 'mafia', 'crime', '.']
Sentence 2: ['Deliver', 'United', 'States', 'authorities', 'leaders', 'al', 'Qaeda', 'hide', 'land', '.']
Sentence 3: ['But', 'today', ',', 'al', 'Qaeda', 'Taliban', ',', 'shelter', '.']
Sentence 4: ['On', 'orders', ',', 'United', 'States', 'military', 'begun', 'strikes', 'Al', 'Qaeda', 'terr
orist', 'training', 'camps', 'military', 'installations', 'Taliban', 'regime', 'Afghanistan']

BOW representation:
{'afghanistan': 1, 'al': 4, 'authorities': 1, 'begun': 1, 'camps': 1, 'crime': 1, 'deliver': 1, 'hide': 1,
'installations': 1, 'land': 1, 'leaders': 1, 'mafia': 1, 'military': 2, 'orders': 1, 'qaeda': 4, 'regime':
1, 'shelter': 1, 'states': 2, 'strikes': 1, 'taliban': 2, 'terror': 1, 'terrorist': 1, 'today': 1, 'traini
ng': 1, 'united': 2}
```

Figure 4. Example implementation for a bag of words representation of text.

[21] NLTK (2022) nltk.tokenize package, retrieved on 17th Feb 2023

NLP applications typically use a 'bag of words' (BOW) representation in pre-processing, which represents a text with a vector indicating the number of occurrences of each chosen word in the training corpus (HaCohen-Kerner et al., 2020; Sebastiani, 2002; Y. Zhang et al., 2010). While more sophisticated versions are under development (see: Yan et al., 2020), Figure 4 shows sentences from Bush's text to explain the typical process for creating a BOW representation. The first step is to remove stopwords deemed of low significance to a text's meaning, such as 'I', 'me', 'my' and many others. The next step is word tokenisation, and the final is a count of words occurring in a text. These words and word counts become the features of a machine-learning model. BOW representations have become fundamental to the success of what this thesis refers to as functionally defined tasks. As Manning and Schütze (1999) explain, nevertheless, BOW representations strip a text of its 'structure and linear ordering of words' (Manning & Schütze, 1999, p. 237), thereby removing a text's grammatical properties.

| | word | POS Tag | Dependency Label |
|---|---|---|---|
| **0** | Al | PROPN (proper noun) | Al (compound) |
| **1** | Qaeda | PROPN (proper noun) | Qaeda (nominal subject) |
| **2** | is | AUX (auxiliary) | is (root) |
| **3** | to | PART (particle) | to (auxiliary) |
| **4** | terror | VERB (verb) | terror (open clausal complement) |
| **5** | what | PRON (pronoun) | what (attribute) |
| **6** | the | DET (determiner) | the (determiner) |
| **7** | mafia | NOUN (noun) | mafia (nominal subject) |
| **8** | is | AUX (auxiliary) | is (clausal complement) |
| **9** | to | ADP (adposition) | to (prepositional modifier) |
| **10** | crime | NOUN (noun) | crime (object of preposition) |
| **11** | . | PUNCT (punctuation) | . (punctuation) |

Figure 5. Parts of speech tagging and dependency labelling.

In contrast to BOW representations, parts of speech tagging and dependency labelling link words by their grammatical properties. 'A part-of-speech tagger, or POS tagger, processes a sequence of words and attaches a part of speech tag to each' (Bird et al., 2009, p. 179). While parts of speech represent the syntactic function of each word, dependency parsing focuses on the grammatical relations. Figure 5 shows the parts of speech and dependency labels for one sentence from Figure 4. NLTK and spaCy use these labels to parse a text for the noun phrases; NLTK relies upon user-defined regular expressions and tag patterns (Bird et al., 2009, pp. 266–267). In contrast, spaCy (the focus of this review) uses language models and language-specific 'syntax iterators'[22] that

[22] spaCy (2022) syntax_iterators.py, retrieved on 17th Feb 2023

identify noun phrases using the grammatical relations between each word. Figure 5 shows the POS and dependency labels for a sentence from Bush's texts.

```
Sentence 0: ['Al Qaeda', 'is', 'to', 'terror', 'what', 'the mafia', 'is', 'to', 'crime', '.', '']

Sentence 1: ['Deliver', 'to', 'United States authorities', 'all the leaders', 'of', 'al Qaeda', 'who', 'hid
e', 'in', 'your land', '.', '']

Sentence 2: ['But', 'today', ',', 'for', 'al Qaeda', 'and', 'the Taliban', ',', 'there', 'is', 'no shelter',
'.', '']

Sentence 3: ['On', 'my orders', ',', 'the United States military', 'has', 'begun', 'strikes', 'against', 'Al
Qaeda terrorist training camps', 'and', 'military installations', 'of', 'the Taliban regime', 'in', 'Afghanis
tan', '']
```

Figure 6. Tokenised text by noun chunks.

The noun chunks shown in Figure 6 are the correctly chunked sentences. As seen in the BOW representation of Figure 4, standard tokenisation separates words that belong together in a noun phrase. For example, tokenisation separates 'al' from 'Qaeda' and 'United' from 'States'. Noun chunking, therefore, ensures a text is tokenised and then chunked by its noun phrases. In Figure 6, 'al Qaeda', 'United States Authorities' and 'the United States military', among others, are the complete noun phrases for the real-world entities they represent. While pre-processing steps described here are standard practice, the section continues by explaining how they fail to correctly represent the prepositional noun phrases and conjunctions an orator may use.

### 2.2.2 How Effectively Do Standard NLP Practices Pre-Process Prepositional Noun Phrases?

The first type of noun chunk reviewed here is the prepositional noun phrase. These phrases follow the subject->predicate->object grammatical model whereby the object modifies the subject, and the predicate is a preposition, such as 'of', 'with' or 'on'. Many prepositional noun phrases from Bush's speech define the War on Terror language. 'War on Terror' itself is a prepositional noun phrase that signifies a violent campaign in response to the World Trade Centre attack on 9/11. 'Weapons of mass destruction' signifies the threat that Bush used to legitimise the war in Iraq in 2003. 'Enemies of America' signifies the threat against whom Bush sought to legitimise violence. As such, Mahmood and Asfar (2016) identify the following three commonly used prepositional phrases in the War on Terror discourse where the word 'terrorism' as an object modifies the subject of a preposition (Mahmood & Afsar, 2016, pp. 547–550)[23].

{[pp_obj_against] + terrorism} Frame: 'War against Terrorism'

{[pp_obj_to] + terrorism} Frame: 'Support to Terrorism'

[23] pp_obj refers to prepositional object in Mahmood and Asfar (2016), this instead chapter uses <noun>

{[pp_obj_on] + terrorism} Frame: 'War on Terrorism'

In Bush's speeches from the hostile narrative corpus, the '<noun> against terrorism' preposition appears 13 times; the '<noun> to terrorism' preposition appears once; the '<noun> on terrorism' appears three times. Replacing 'terrorism' with 'terror' reveals more prepositions used by Bush. '<noun> against terror' appears 4 times and the '<noun> on terror' appears three times. Any NLP algorithm claiming to understand War on Terror texts must correctly parse these key phrases.

To assess whether NLP algorithms currently process these prepositional phrases, consider the following sentences from Bush's texts:

> *Sentence 1: Our War on Terror begins with al Qaeda, but it does not end there.*
>
> *Sentence 2: These same terrorists are searching for weapons of mass destruction, the tools to turn their hatred into holocaust.*
>
> *Sentence 3: On September the 11th, enemies of freedom committed an act of war against our country.*
>
> *Sentence 4: The face of terror is not the true faith of Islam.*
>
> *Sentence 5: The United States of America is a friend to the Afghan people, and we are the friends of almost a billion worldwide who practice the Islamic faith.*

These sentences are essential to understand the legitimisation of violence against al Qaeda. The first sentence identifies al Qaeda as the target of the War on Terror. In context, the second directly links al Qaeda ('these same terrorists') to the threat of weapons of mass destruction, thereby legitimising violence against them. 'Act of war against our country' from the third sentence is a nested prepositional phrase that denotes the attacks in New York on the 9$^{th}$ of September 2001. 'Act of War' is the subject and a prepositional phrase, the predicate is 'against', and the object is 'our country'. The fourth and fifth sentences distinguish Muslims from Bush's target of violence. The fourth sentence contains the prepositional noun phrase 'faith of Islam' while the fifth contains a reordering of these words in an adjectival noun phrase, ‘Islamic faith’. Both these prepositional and adjectival noun phrases denote the same religion.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Sentence 1 Noun Chunks** | Our war on terror | al Qaeda | it | | | |
| **Sentence 2 Noun Chunks** | These same terrorists | weapons of mass destruction | their hatred | holocaust | | |
| **Sentence 3 Noun Chunks** | September | enemies of freedom | war | an act of war against our country | | |
| **Sentence 4 Noun Chunks** | The face of terror | the true faith of Islam | | | | |
| **Sentence 5 Noun Chunks** | The United States of America | a friend to the Afghan people | we | the friends of almost a billion worldwide | who | the Islamic faith |

Figure 7. A human parse of noun chunks.

Figure 7 depicts how humans parse the noun chunks of each sentence, thereby representing a grammatically correct parsing of each sentence.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| **Sentence 1 Noun Chunks** | Our war | terror | al Qaeda | it | | | | |
| **Sentence 2 Noun Chunks** | These same terrorists | weapons | mass destruction | their hatred | holocaust | | | |
| **Sentence 3 Noun Chunks** | September | enemies | freedom | an act | war | our country | | |
| **Sentence 4 Noun Chunks** | The face | terror | the true faith | Islam | | | | |
| **Sentence 5 Noun Chunks** | The United States | America | a friend | the Afghan people | we | the friends | who | the Islamic faith |

Figure 8. Spacy processed noun chunks.

Figure 8 shows how spaCy incorrectly chunks the noun phrases from each text. SpaCy’s syntax iterator incorrectly parses critical terms such as ‘War on Terror’, ‘weapons of mass destruction’ and ‘act of war against our country’ by separating the elements of each noun phrase. As such, these prepositional noun phrases in Bush’s speech lose their specific meaning. ‘War’, ‘weapons’ and ‘enemy’ are no longer linked to the terms ‘terror’, ‘mass destruction’ and ‘America’, which feature essential terms of a War on Terror narrative. Moreover, spaCy correctly chunks ‘Islamic Faith’ but incorrectly chunks ‘faith of Islam’ despite how these terms denote the same religion. The incorrect chunking of these noun phrases, therefore, does not represent an understanding of defining concepts in a War on Terror narrative. An additional problem with pre-processing prepositional noun phrases is a problem with conjunctional phrases.

### 2.2.3 How Effectively Do Standard NLP Practices Pre-Process Conjunctional Phrases?

Much like prepositional phrases, NLP tokenisation also seem to misrepresent conjunctional phrases. Conjunctions are words such as ‘and’ and ‘or’ that join words across a sentence. Notably, these are often deemed stop words and, therefore, of low significance in the standard pre-processing practices. Conjunctional phrases also follow the subject->predicate->object model, whereby conjunctions link the subject and object. As an example, consider the sentence from bin Laden's declaration of Jihad.

> *Sentence 6: The image of that dreadful massacre in Qana, Lebanon, is still vivid in one's mind, and so are the massacres in Tajikistan, Burma, Kashmir, Assam, the Philippines, Fatani, Ogaden, Somalia, Eritrea, Chechnya, and Bosnia-Herzegovina where hair-raising and revolting massacres were committed before the eyes of the entire world clearly in accordance with a conspiracy by the United States and its allies who banned arms for the oppressed there under the cover of the unfair United Nations.*

This sentence invokes the image of 11 massacres in 11 different countries, which bin Laden uses to legitimise his violent jihad against 'the United States', 'its allies' and the 'unfair United Nations'. He refers to the image of these massacres in the conjunction, 'massacres in Tajikistan, Burma, Kashmir, Assam, the Philippines, Fatani, Ogaden, Somalia, Eritrea, Chechnya, and Bosnia-Herzegovina'. The prepositional noun phrase 'massacre in Tajikistan' represents the first event of the conjunction. The subject is 'the image of that dreadful massacre', the predicate is 'in', and the object is 'Tajikistan'. The conjunction continues by referencing 10 other countries, each of which is an object to the subject 'massacre'. A subsequent conjunction, 'a conspiracy by the United States and its allies' should resolve to two conjunctions of 'a conspiracy by the United States' and 'a conspiracy by its allies'. A correct understanding of this lengthy sentence, therefore, links the word 'massacre' to each country and, as shown here, would sound rather long and clumsy:

> *[…] massacre in Tajikistan, massacre in Burma, massacre in Kashmir, massacre in Assam, massacre in the Philippines, massacre in Fatani, massacre in Ogaden, massacre in Somalia, massacre in Eritrea, massacre in Chechnya, and massacre in Bosnia-Herzegovina […] clearly in accordance with a conspiracy by the United States and a conspiracy by its allies…*

| | Human Parse | spaCy Parse |
|---|---|---|
| 0 | massacres in Tajikistan | the massacres |
| 1 | massacres in Burma | Tajikistan |
| 2 | massacres in Kashmir | Burma |
| 3 | massacres in Assam | Kashmir |
| 4 | massacres in the Philippines | Assam |
| 5 | massacres in Fatani | the Philippines |
| 6 | massacres in Ogaden | Fatani |
| 7 | massacres in Somalia | Ogaden |
| 8 | massacres in Eritrea | Somalia |
| 9 | massacres in Chechnya | Eritrea |
| 10 | massacres in Bosnia - Herzegovina | Chechnya |
| 11 | conspiracy by the United States | Bosnia-Herzegovina |
| 12 | conspiracy by its allies | a conspiracy |
| 13 | | the United States |
| 14 | | its allies |

Figure 9. Comparing a human and spacy parse of conjunction phrases.

Much like noun chunking in Figure 8, this sentence's chunking differs from human interpretation. As Figure 9 shows, a human would process the conjunction as 11 massacres in 11 countries, whereas NLP processes the conjunction as two nouns and 11 countries. SpaCy also incorrectly parses the conspiracy conjunction into two nouns and one proper noun. This conjunction is essential to understand since bin Laden legitimises violence by attributing these massacres to a conspiracy by the United States and its allies. A reference to 11 massacres makes a compelling case for retaliation, which spaCy's syntax iterator does not capture. As such, the chunking of noun phrases and conjunctions using standard pre-processing practices misrepresents the orator's intended message. Following this review of pre-processing requirements, the chapter continues with assessing word embeddings for encoding natural language.

### 2.2.4 Experiment 1: Assessing Word2vec Representation of Bush's Declaration of War

This subsection assesses word embeddings for encoding text using Bush's declaration of war from the hostile narrative corpus. The assessment comprises an experiment that is available online[24], and the figures presented here are screenshots from these experiments. Regarding assessment aim, a paper by Bolukbasi *et al. (2016)* finds 'blatant sexism' in word embeddings from the Google New corpus that hostile narrative analysis seeks to detect (Bolukbasi et al., 2016, p. 1). This

[24] Anning (2023) Assessing word2vec, retrieved on 17th Feb 2023

sexism represents hostility towards women in western news media. This experiment, therefore, aims to detect similar hostility in Bush's text towards al Qaeda and other terrorist organisations.

This experiment reviews word2vec, which is the first algorithm to use word embeddings, and the next section assesses transformers as the current state-of-the-art. The point of reviewing this entry level algorithm with the state-of-the-art is to show what has *not* changed despite the very obvious technical advances. In effect, both algorithms process text by word co-occurrence. The experiments then show the inappropriateness of inferring meaning from processing text by word cooccurrence for social science applications. In effect, the obvious technical sophistication of transformers provides no more meaningful outputs that entry-level word2vec algorithms.

Word embedding is a quantitative method to encode word meanings in a numerical vector. Algorithms then decode meaning from the numerical distance between words in a vector distribution. For Patel and Arasanipalai (2021), 'encoding text into numbers emphasises the meaning of the text' by 'looking at the context in which they appear' (Patel & Arasanipalai, 2021, pp. 110–111). Word2vec by Google was the first model to use word embeddings, and more sophisticated models, like GPT-1, GPT-2 and GPT-3, are now state-of-the-art (see: Patel & Arasanipalai, 2021, pp. 190–196). The appeal of using word embedding is in reducing the reliance on rule-based systems and linguistic resources that otherwise constrain the development of large-scale systems. Word embeddings have become a standard practice for encoding in NLP.

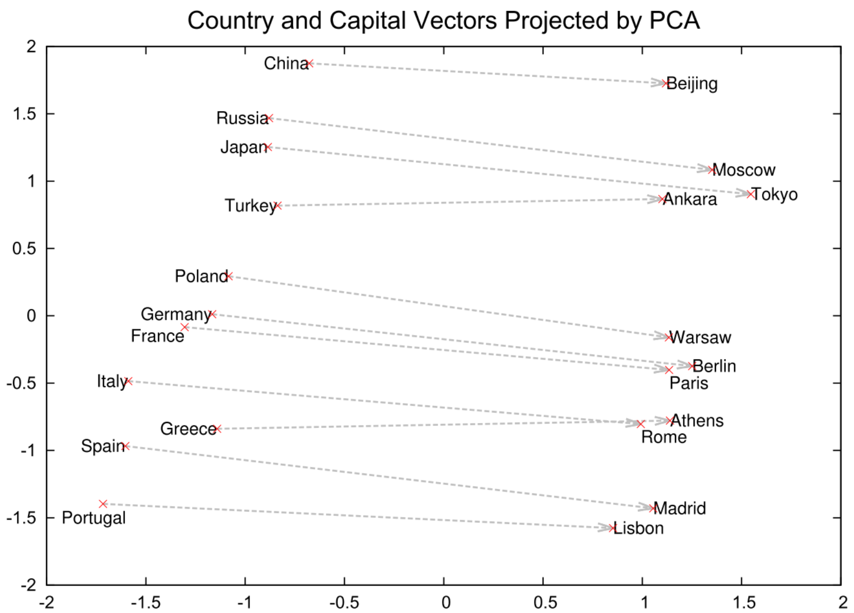



Figure 10. Extract from Mikolov's 2013 paper showing the distribution of capital cities relative to their country.

Word embeddings infer meaning and semantic relationships from word co-occurrence following the often-quoted distributional hypothesis, 'You shall know a word by the company it keeps'

(Gardner et al., 2015, p. 1084). To explain how this hypothesis applies, Figure 10 is an extract from Mikolov (2013) and shows a vector distribution of capital cities relative to their country. In such phrases as ‘Lisbon, the capital of Portugal’ or ‘the capital of Portugal is Lisbon’, a similar number of words separate each capital city and its country. Consequently, an equivalent mathematical distance separates the vector representation for each country and its corresponding capital. Word embedding then treats clusters of cities or countries as synonymous terms. Moreover, the mathematical distance between ‘Lisbon’ and ‘Portugal’ also applies to ‘Spain’ to infer its capital as ‘Madrid’. The result, therefore, is the quantification of meaning in a vector representation.

Word embeddings should contend with what this thesis refers to as the ‘abstract representation of groups’ in hostile narratives to support any claim of understanding hate speech. Abstract representations (explained in more detail in the next chapter) use pronouns and noun phrases to represent people and groups. As Galtung (1990) explains, using pronouns or noun phrases to represent groups is a process of othering. Hitler described the Jews as the ‘dangerous it’, the ‘vermin’, or ‘bacteria’; Stalin described the ‘kulaks’ in political terms as the ‘class enemy; Reagan described Qadhafi as the ‘mad dog’; Washington experts describe ‘terrorists’ as the ‘cranky criminals’ (Galtung, 1990, p. 298). In addition to othering, noun phrases also feature in elevation whereby ingroup members conversely represent themselves as the heroes or saviours of a story in response to the actions of an outgroup. These pronouns and noun phrases then interact as part of a narrative in violence legitimisation.

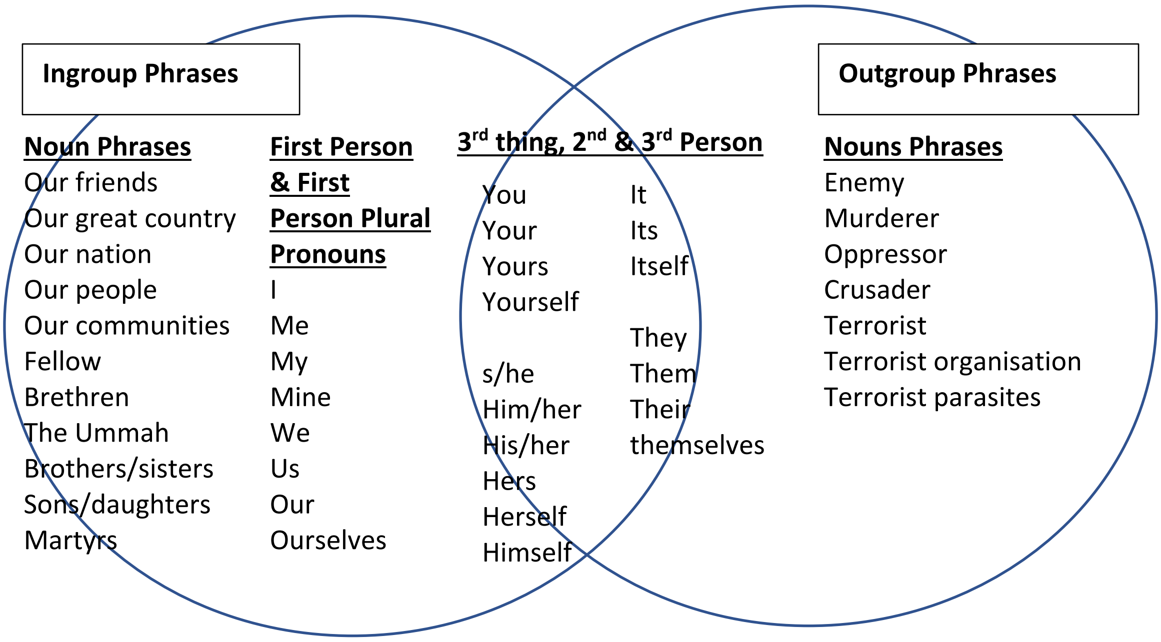



Figure 11. Abstract representations in Bush’s and bin Laden’s texts.

Figure 11 shows a sample of noun phrases and pronouns used by Bush and bin Laden to refer to their ingroup and outgroups. As the figure shows, first-person pronouns such as 'we' or 'our' refer to an ingroup, whereas second and third-person pronouns like 'you' or 'they' can refer to either an ingroup or outgroup. Notably, these are common stop words, so standard pre-processing practices would remove them despite containing valuable information about hostility. Familial noun phrases like 'brothers' and 'sisters' or 'sons' and 'daughters' signify an ingroup, while phrases like 'crusader' or 'terrorist' signify an outgroup. These noun phrases have varying degrees of intensity for elevation and othering.

| | Friend | Enemy | Terrorist | Good | Bad | Evil | Parasite |
|---|---|---|---|---|---|---|---|
| 1 | Pal, (0.748) | Enemies, (0.779) | Terror, (0.848) | Great, (0.729) | Good, (0.719) | Malevolent, (0.645) | Parasites, (0.794) |
| 2 | Friends, (0.71) | Adversary, (0.627) | Terrorists, (0.8) | Bad, (0.719) | Terrible, (0.683) | Wickedness, (0.634) | Bacterium, (0.69) |
| 3 | Buddy, (0.697) | Adversaries, (0.594) | Terrorism, (0.752) | Terrific, (0.689) | Horrible, (0.67) | Evil_Doers, (0.633) | Malaria_Parasite, (0.68) |
| 4 | Dear_Friend, (0.696) | Hostiles, (0.583) | Al_Qaeda, (0.731) | Decent, (0.684) | Bad, (0.67) | Demonic, (0.631) | Parasitic_Worms, (0.671) |
| 5 | Acquaintance, (0.684) | Old_Mariam_Sajadi, (0.56) | Terrorist, (0.718) | Nice, (0.684) | Lousy, (0.665) | Villainous, (0.628) | Protozoan_Parasite, (0.665) |
| 6 | Cousin, (0.671) | Inevitably_Schulberg, (0.548) | Al_Qaeda, (0.708) | Excellent, (0.644) | Crummy, (0.568) | Evil, (0.611) | Pathogen, (0.658) |
| 7 | Girlfriend, (0.623) | Heartless_Ruthless, (0.535) | Extremist, (0.681) | Fantastic, (0.641) | Horrid, (0.565) | Evil_Incarnate, (0.607) | Bacteria, (0.653) |
| 8 | Colleague, (0.62) | Must_Outthink_Outwork, (0.534) | Al_Qa'Ida, (0.68) | Better, (0.612) | Awful, (0.553) | Satanic, (0.599) | Microfilariae, (0.644) |
| 9 | Uncle, (0.612) | Islamofascist_Terrorists, (0.53) | Jihadist, (0.679) | Solid, (0.581) | Dreadful, (0.553) | Satan, (0.593) | Parasitic_Worm, (0.64) |
| 10 | Roommate, (0.612) | Aggressors, (0.527) | Al_Qaida, (0.675) | Lousy, (0.576) | Horrendous, (0.545) | Evilness, (0.591) | Protozoan, (0.638) |

Figure 12. Word2vec results for the Google News corpus (similarity score in brackets).

This first experiment assesses the word2vec word embedding algorithm using Bush's declaration of war against al Qaeda. The experiment is based on the following hypothesis:

*Word embeddings enable the identification of ingroups and outgroups.*

A vector distribution of the Google News corpus suggests some merit to the experimental hypothesis[25]. The corpus contains about 100 billion words with 300-dimensional vectors for 3 million words and phrases[26]. Figure 12 shows the top 10 terms deemed synonymous with various seed terms. As an initial observation, the multiple spellings of al Qaeda reveal the text pre-processing requirement to resolve different ways to denote the same entity. As such, a normalised spelling for 'al Qaeda' would produce higher-quality results. More analytically, 'friend' and 'enemy', 'good', 'bad' and 'evil' generate reasonably synonymous words. The synonymy of 'al Qaeda' and 'Islamofascist terrorists' with 'terrorist' reflects a western bias that Bush's texts sought to create. This similarity from word embeddings reflects a colloquial, as opposed to formal, synonymy that is specific to the chosen dataset. Indeed, colloquial synonymy represents a

---

[25] Anning et al (2022) Assessing Google New Corpus, retrieved on 17th Feb 2023

[26] Google Code (n.d.) word2vec, retrieved on 17th Feb 2023

bias that Bush's more hostile declaration of war should exaggerate. The hypothesis, therefore, is accepted if word embeddings for Bush's texts reveal similar synonymy for the same seed terms.

| | Friend | Enemy | Terrorist | Good | Bad | Evil | Parasite |
|---|---|---|---|---|---|---|---|
| 1 | United_States, (0.994) | United_States, (0.995) | United_States, (0.997) | United_States, (0.994) | United_States, (0.922) | Terrorism, (0.992) | Pursue, (0.708) |
| 2 | Terrorism, (0.993) | Terrorism, (0.995) | Work, (0.996) | Work, (0.993) | Enemy, (0.922) | Terror, (0.991) | September_The_11Th, (0.708) |
| 3 | Work, (0.992) | New, (0.994) | New, (0.996) | Terrorism, (0.993) | Security, (0.921) | United_States, (0.991) | Win, (0.708) |
| 4 | Military, (0.992) | Work, (0.994) | Terror, (0.995) | Terror, (0.993) | Work, (0.921) | Weapon, (0.991) | Time, (0.708) |
| 5 | World, (0.992) | Terror, (0.994) | Weapon, (0.995) | Military, (0.993) | Military, (0.921) | Work, (0.991) | Citizen, (0.706) |
| 6 | Weapon, (0.992) | Military, (0.993) | Military, (0.995) | Child, (0.993) | People, (0.921) | People, (0.99) | See, (0.706) |
| 7 | People, (0.992) | Great, (0.993) | Great, (0.995) | People, (0.993) | Home, (0.921) | Great, (0.99) | Evil, (0.706) |
| 8 | Terror, (0.992) | Weapon, (0.993) | Security, (0.995) | Weapon, (0.993) | New, (0.921) | Security, (0.99) | Building, (0.705) |
| 9 | Security, (0.991) | Help, (0.993) | Afghanistan, (0.995) | Great, (0.993) | Terror, (0.921) | New, (0.99) | Murder, (0.705) |
| 10 | Great, (0.991) | Security, (0.993) | Child, (0.995) | Help, (0.993) | Fight, (0.921) | Military, (0.99) | Important, (0.705) |

Figure 13. Word2vec results of George Bush's texts (similarity score in brackets).

Figure 13 shows that word embeddings do not support the experimental hypothesis for Bush's texts[27]. The table shows the similarity scores for the same seed terms as Figure 12 from a vector distribution of Bush's texts. The configuration of this model uses standard parameters as follows:

- word2vec algorithm = Gensim[28]
- Number of words = 111934
- vector_size = 300 (Dimensionality of the word vectors)
- window = 5 (Maximum distance between the current and predicted word within a sentence)
- skip-gram (uses the central word to predict the surrounding words)

The distribution includes normalised terms to improve the co-occurrence of named entities and, in part, address the pre-processing problems identified in the previous section. For example, Figure 13 contains 'United_States' as a normalised term to denote 'the United States of America' and 'the US', among others. To accept the hypothesis, 'United States' would be colloquially synonymous with 'Friend' and 'Good', while 'al Qaeda' and 'the Taliban' would be synonymous with the remaining othering terms. In effect, the seed terms suggest the connotative meaning for each named entity. Nevertheless, a synonymy of 'United_States' with 'Friend', 'Enemy', 'Terrorist' and 'Good' is both a contradiction and a misrepresentation of Bush's intended meaning. A smaller dataset comprising only 111,934 words has not generated the kind of output a much larger corpus generates.

[27] Anning et al (2022) Assessing Word Embeddings, retrieved on 17th Feb 2023

[28] Gensim (n.d.) word2vec, retrieved on 17th Feb 2023

| | United States | Americans | al-Qaeda | Taliban | terrorist | enemy | murderer |
|---|---|---|---|---|---|---|---|
| **Mention Count** | 187 | 37 | 16 | 25 | 118 | 41 | 7 |

Figure 14. Different representations for each named entity.

Words should frequently co-occur for the distributional hypothesis to apply to a hostile narrative; the abstract representation of named entities, however, reduces the frequency of word co-occurrence. Figure 14 shows the number of ingroup and outgroup entity mentions and outgroup terms for Bush's text. He explicitly mentions the 'United States' and 'Americans' 224 times; in contrast, 'al Qaeda' and the 'Taliban' collectively gain 41 explicit mentions. Indicative of hostility in his speech, Bush variously refers to his outgroups as 'terrorist', 'enemy' and 'murderer' 166 times. Moreover, a manual review of one speech reveals how Bush variously refers to 'al Qaeda' and 'the Taliban' as 'they' 22 times. These abstract representations contribute to the othering of Bush's outgroup and subsequent legitimisation of the War on Terror. Nevertheless, standard pre-processing practices remove these pronouns, thereby skewing the text's original meaning. While a human may subconsciously link these abstract references to 'al Qaeda' and 'the Taliban', word embeddings do not appear to make the same connection. Of additional interest, 'Egyptian_Islamic_Jihad' and 'Islamic_Movement_of_Uzbekistan' are not in the Google News vocabulary despite being named in Bush's text as an outgroup.

More advanced algorithms to word2vec still apply the distributional hypothesis to process text by word occurrence. Advances upon word2Vec began with the Embeddings from Language Models (ELMo) by Peters *et al*. (2018) from the Allen NLP institute that uses the long short-term memory (LSTM) architecture (Peters et al., 2018, pp. 2–3). ELMo was among the first technologies to use bidirectionally when generating a vector representation. Previously, encoding algorithms treated text as an iterable, therefore, they could only consider words to the left of a target word when generating a vector. Bidirectionality, conversely, accounts for words to the left and right of a target word, thereby capturing a word's full context in the vector representation rather than just relying on the context of previous words. Of particular interest here, bidirectionality disambiguates denotative meaning for homonyms - words that share the exact spelling but have different meanings. For example, ELMo generates different representations for the homonym 'bank' in the context of 'an ant went to the river *bank*' or 'that is a good way to build up a *bank* account' (Zhou et al., 2020, p. 276). Having established how NLP encodes natural language, the chapter now continues with how NLP decodes hate speech and sentiment analysis.

## 2.3 How Effective Are Transformers for Hate Speech Detection?

Having assessed word embeddings for encoding natural language in the previous section, this next experiment now assesses transformers for decoding hate speech using the Detoxify model from Hanu (2020)[29]. The experiment begins with an overview of how transformers have developed upon vector representations in word2vec. Note that this overview is about operationally assessing transformers for explanatory dialogues, therefore, it purposefully avoids technical depth. The section then continues by summarising experiments with Detoxify that compare outputs from Luther King's *I Have a Dream* text with Hitler's *Mein Kampf*. These experiments then reveal problems with quantitatively processing hate speech by word co-occurrence. Overall, while this overview acknowledges undoubted advances in the technical sophistication of NLP for detecting hate speech, it also reveals three constants. Firstly, transformers still apply the distributional hypothesis by processing text using word co-occurrence. Secondly, transformers still require human annotation to encode connotative meaning. Thirdly, language models contain a consensus bias (generally Western) that social science applications may amplify in their outputs.

Over several technical advances in word embeddings, transformer-based architectures – often referred to as large language models (LLM) - are now state-of-the-art in NLP. Vaswani *et al*. (2017) employed the idea of self-attention to introduce the transformer architectures. Self-attention, sometimes called intra-attention, is 'an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence' (Vaswani et al., 2017, p. 2). Self-attention allows transformers to process input sequences in parallel, known as parallelisation, rather than sequentially like previous architectures. Sequential processing means the algorithm can only attend to previously seen words rather than the whole context to the left and right of a target word. Parallelisation, on the other hand, attends to words in parallel, therefore, considers words following the target words when generating a vector. Parallelisation in NLP allows the transformer to process long input sentences for such applications as language translation that require a strong understanding of word context. Parallelisation in transformers, still applies the distributional hypothesis by considering the co-occurrence of words rather than their grammatical relations.

Bidirectional Encoder Representations from Transformers (BERT) by Devlin *et al*. (2018) from Google was the first transformer-based architecture in NLP that combined the ideas from

---

[29] Hanu *et al*. (2020) Detoxify, retrieved on 17th Feb 2023

transformers and ELMo. 'BERT's model architecture is a multi-layer bidirectional Transformer encoder' that comprises two steps of *pre-training* and *fine-tuning* (Devlin et al., 2018, p. 3). Pre-training refers to the general encoding of natural language, while fine-tuning is specific to the language domain. For pre-training, BERT enables bidirectionality with a 'masked language model' method that disambiguates denotative meaning, as with ELMo. The pre-training also employs 'next sentence prediction' to predict whether one sentence is likely to follow another given sentence. Fine-tuning then draws upon task-specific and labelled data using the initial pre-trained model as a checkpoint for general NLP applications such as sentiment analysis.

> *A large language model is a machine learning model that is trained to process and generate text. It is called 'large' because it has a large number of parameters, typically in the millions or billions, and is trained on a very large dataset, typically billions of words. Some examples of large language models and their sizes include:*
>
> *GPT-3 (Generative Pre-trained Transformer 3): 175 billion parameters*
>
> *GPT-2 (Generative Pre-trained Transformer 2): 1.5 billion parameters*
>
> *BERT (Bidirectional Encoder Representations from Transformers): 110 million parameters*
>
> *RoBERTa (Robustly Optimised BERT Approach): 125 million parameters*
>
> *These models are significantly larger than many other NLP models and have achieved state-of-the-art results on a variety of benchmarks. However, it is important to note that the optimal size of a language model can vary depending on the specific task and dataset, and it may be possible to achieve good performance with a smaller model in some cases.*

Suggesting the transformation potential of transformers for functional applications, the ChatGPT LLM from OpenAI generated the above description of LLMs[30]. Yet, while OpenAI have placed controls on ChatGPT for potential biases, it still contains a Western consensus bias.

> *The Taliban is a fundamentalist Islamic group that originated in Afghanistan in the 1990s. They rose to power during the country's civil war in the mid-*

---

[30] Generated and edited from ChatGPT using the question 'what is and how big are large language models', retrieved on 17th Feb 2023

> *1990s and governed Afghanistan from 1996 to 2001, when they were overthrown by the United States-led coalition after the 9/11 attacks.*

Figure 15. A ChatGPT output for the question, 'Who is the Taliban?'.

Figure 15 shows a ChatGPT output for the question 'who is the Taliban?'. This output represents a narrative truth contained in Bush's declaration of war against them. The adjective, 'fundamentalist Islamic Group' connotes Islamic terrorism that Bush resolved to fight. Recalling the introduction, however, this view does not represent the lived experience of members of different Taliban tribes, it is a consensus bias of Western opinion for legitimising the War on Terror. Moreover, this view begs the question of how do LLM learn from changes in the real-world? As of Summer 2021, the Taliban have returned to power as the government of Afghanistan[31], therefore, how much data is required to overcome the consensus bias represented in large language models that comprise trillions of words?

The potential for consensus bias is also feasible in a several papers that build language models using sentence embeddings(Kiros et al., n.d.; Lin et al., 2017; Liu & Lapata, n.d.; Wieting & Gimpel, 2017). Rather than encoding text as the level of the word, they encode text as the level of the sentence. Much like vector representations of words, sentence embeddings likely magnify any consensus expressed by the co-occurrence of sentences. Any assessment of these sentence embeddings could following the experiments in this chapter to verify the extent of this bias and the effectiveness of processing a text by sentence co-occurrence. Having introduced transformers, the section continues with a practical assessment of their utility for detecting hate speech.

### 2.3.1 Experiment 2: Assessing Detoxify for Decoding Hate Speech

While transformers are undoubtedly sophisticated and have transformative potential in many NLP applications, this experiment is about their value to explanatory dialogues about hate speech. This open-sourced experiment[32] assesses the Detoxify model for hate speech detection found in the Huggingface library[33]. Huggingface provides an open-source platform for training and deploying NLP models mainly based on the transformer architecture. They offer tools for pre-training and fine-tuning models for both functional and social science applications. Of note, while transformers provide transformative potential for functional tasks, such as natural language generation as

---

[31] BBC (2022) Who are the Taliban?, retrieved on 17th Feb 2023

[32] Anning (2022) Assessing Detoxify, retrieved on 17th Feb 2023

[33] Huggingface (n.d) Models, retrieved on 17th Feb 2023

above, human annotators are still often required to fine-tune models for their chosen application. Accordingly, the Detoxify model is based on a BERT architecture and uses human annotators for labelling[34]. The annotation draws upon three open-source challenges: Toxic Comment Classification Challenge[35], Jigsaw Unintended Bias in Toxicity Classification[36], and Jigsaw Multilingual Toxic Comment Classification[37]. The experiment uses Detoxify, therefore, because it is assumed to broadly represent a community consensus for the state-of-the-art and what constitutes toxic language[38].

| | toxicity | severe_toxicity | obscene | threat | insult | identity_attack |
|---|---|---|---|---|---|---|
| **I Have a Dream (82 sentences)** | 17 | 0 | 0 | 0 | 0 | 13 |
| **Mein Kampf (4376 sentences)** | 115 | 0 | 7 | 5 | 21 | 24 |

Figure 16. Detoxify scores greater than 0.1 for *I Have a Dream* and *Mein Kampf*.

This experiment reviews Detoxify using *Mein Kampf* and *I Have a Dream*. While *Mein Kampf* is undeniably hateful, *I Have a Dream* provides a benchmark for a non-hateful text. The experiment uses the spaCy library to segment all the sentences from each text and then uses Detoxify to classify each; Figure 16 shows the results. Detoxify classified 17 of 82 sentences in I Have a Dream as toxic and suggested Luther King made an identity attack 13 times. They also show that out of 4376 sentences in *Mein Kampf*, Detoxify identifies 115 toxic and 7 obscene sentences, 21 insults and 24 identity attacks. These are somewhat surprising results since *I Have a Dream* should have no toxic sentences, and Luther King certainly makes no identity attacks. Moreover, given its greater size and the level of antisemitism it contains, *Mein Kampf* should have many more toxic sentences and identity attacks.

The following review of specific sentences from these texts reveals the problems with the quantitative methods used by Detoxify to detect hate speech. The developers acknowledge these problems and encourage fine-tuning for specific use cases. Nevertheless, *Mein Kampf* is a conical text for hate speech, therefore, Detoxify should require minimal fine-tuning. They also note

---

[34] Hanu *et al*. (2021) How AI Is Learning to Identify Toxic Online Content, retrieved on 17th Feb 2023

[35] Kaggle (2018) Toxic Comment Classification Challenge, retrieved on 17th Feb 2023

[36] Kaggle (2019) Jigsaw Unintended Bias in Toxicity Classification, retrieved on 17th Feb 2023

[37] Kaggle (2020) Jigsaw Multilingual Toxic Comment Classification, retrieved on 17th Feb 2023

[38] Toxicity is given to refer to the 'ordinary meaning' of hate speech, as explained in the next chapter.

problems with processing texts by word co-occurrence. They comment, ‘we noticed that the inclusion of insults or profanity in a text comment will almost always result in a high toxicity score, regardless of the intent or tone of the author’[39]. This explanation develops upon the developer’s comments by reviewing selected sentences from each text to show the problems with quantitatively processing hate speech by word co-occurrence.

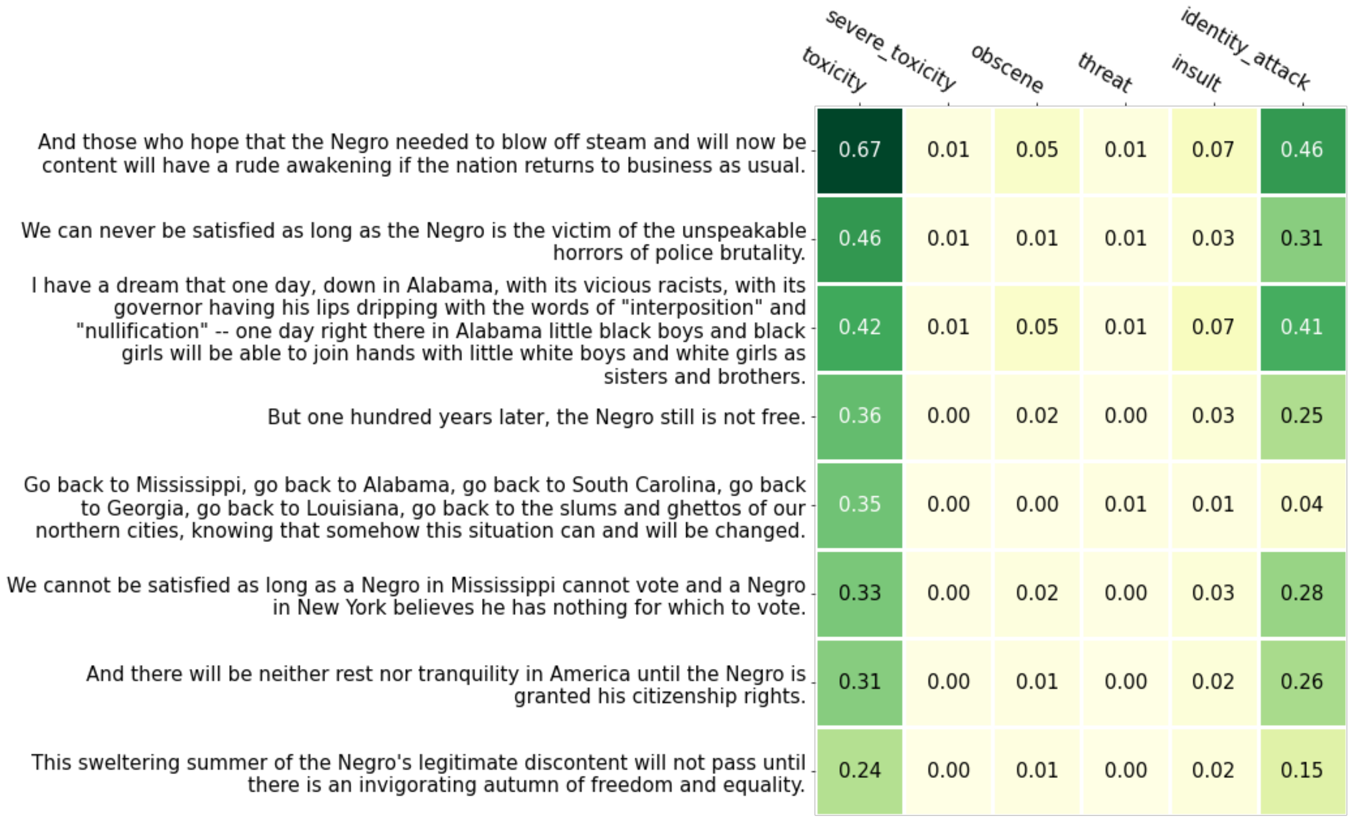



Figure 17. Benchmark data from *I Have a Dream*.

Figure 17 provides benchmark data for sentences that should not generate toxicity scores. The benchmark data is the top 8 most toxic sentences from Luther Kings’ *I Have a Dream* speech. The experiment scored each sentence and sorted them by the level of toxicity according to the algorithm. These are counter-intuitive results because Luther King expresses no toxicity in any of these sentences, and neither are they an identity attack, as the scores suggest; the expected score for these sentences is zero. Of interest, Luther King uses the word ‘negro’ in all but the third and fifth sentences to denote his ingroup and describe their experiences of inequality. The first two sentences express their disquiet towards continued inequality and police brutality. The fourth, sixth, seventh and last sentences are about the negro’s political disenfranchisement and their

[39] Hanu *et al*. (2021) How AI Is Learning to Identify Toxic Online Content, retrieved on 17th Feb 2023

desire to be free. The fifth sentence expresses the hope that inequality can be challenged but generates a toxic score. The occurrence of ‘negro’, therefore, inappropriately applies toxicity scores despite Luther King’s benign use of the term. Rather than tackling hate speech, the algorithm negatively scores Luther King’s dream of equality for black Americans.

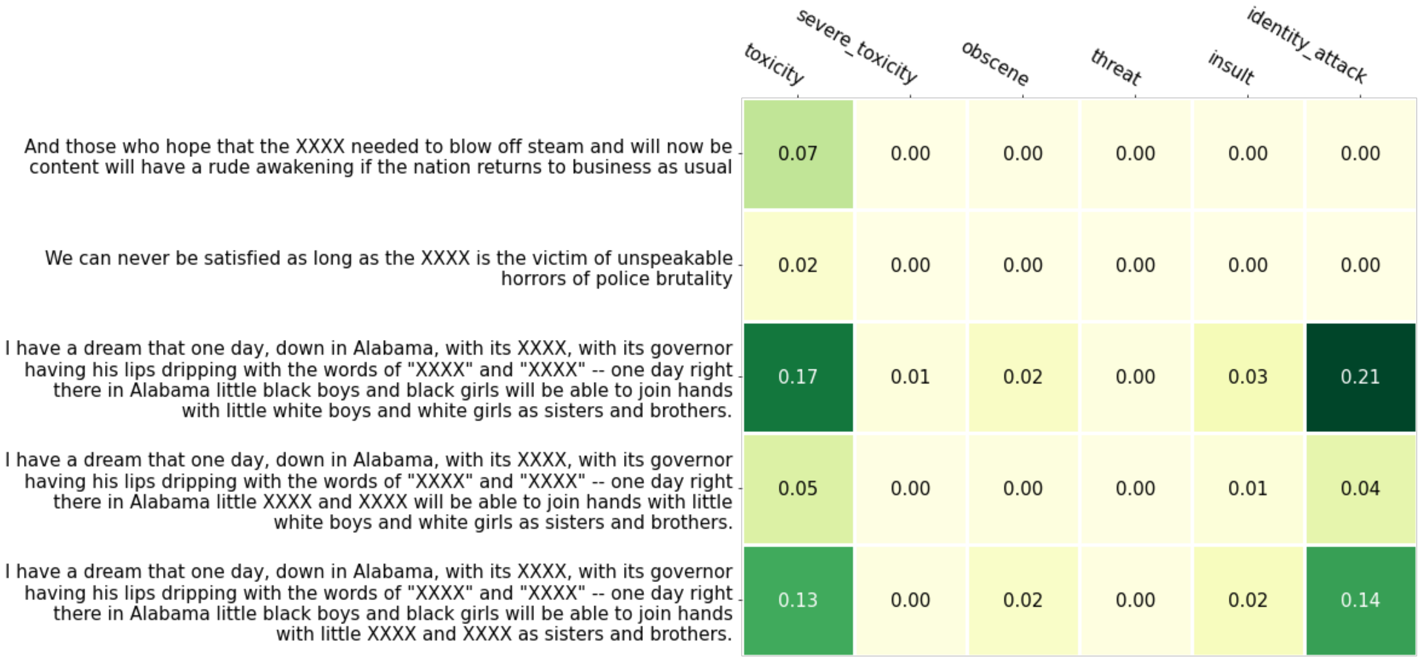



Figure 18. Altered results from *I Have a Dream*.

Figure 18 shows altered results from Figure 17 to explain the problems of processing text by word co-occurrence to detect hate speech. The problem arises from using the occurrence of words to infer meaning. Recall Maron’s (1961) original method proposed using the occurrence of words to infer what a text is about. The same method is inappropriately used here to infer meaning. Removing the word negro from the first two sentences reduces the toxicity score from 0.67 to 0.07 and 0.46 to 0.02. The simple occurrence of the word ‘negro’, therefore, increases the toxicity score by 90% and 95%, respectively. The third sentence removes the words ‘vicious racists’, ‘interposition’ and ‘nullification’, which reduces the toxicity score by 60% from 0.42 to 0.17. As such, the algorithm draws a link from the occurrence of words in the input sentence to a similarly annotated element of the training data, and not from Luther King’s intended use of each word.

A further alteration also reveals an interesting finding about how the algorithm scores the words ‘black’ and ‘white’. The third sentence envisions a world where ‘little black boys and black girls join hands with little white boys and white girls as sisters and brothers’. The removal of ‘black boys’ and ‘black girls’ from this sentence reduces the toxicity by 88% from 0.42 to 0.05 but removing ‘white boys’ and ‘white girls’ while keeping ‘black boys’ and ‘black girls’ reduces the toxicity score by 69% to 0.42 to 0.13. Somewhat problematically, the algorithm generates a higher toxicity score for a sentence containing the word black. In effect, the counter-intuitive scores arise from the occurrence of words rather than how Luther King uses them.

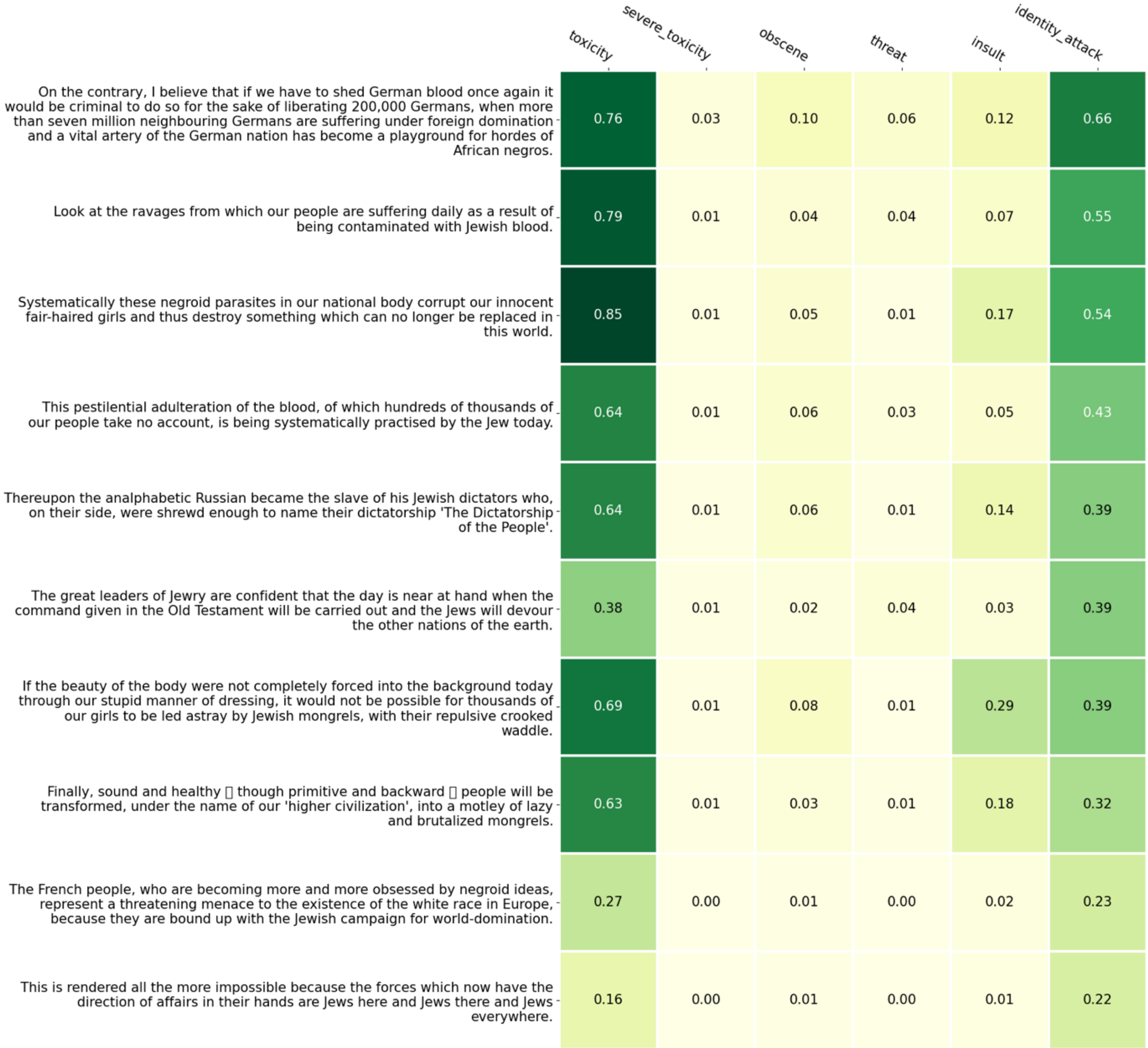



Figure 19. Detoxify outputs for identity attacks in *Mein Kampf*.

Figure 19 shows the top 10 sentences for identity attacks from Mein Kampf. These results are promising as they each contain an attack on a particular identity. The highest scoring sentence references 'African negros', and the ninth sentence refers to 'negroid ideas'. Where 'negroid' is the adjective of 'negro', nevertheless, these scores highlight problems with decoding connotative meaning. For Luther King, the term 'negro' connotes his ingroup, while the same term connotes an outgroup for Hitler. Nevertheless, while these sentences are undoubtedly identity attacks, numerical values do not identify the target. Determining the target and why the sentence constitutes an attack against them still requires human interpretation.


Figure 20. Detoxify outputs for severe toxicity in *Mein Kampf*.

As a text which became the driving narrative of genocide, scores for severe toxicity in *Mein Kampf* should be high. Accordingly, Figure 20 shows the top 10 results for severe toxicity in *Mein Kampf*. Against expectations, the most severely toxic sentence (0.09) is a low score, and only two other sentences generate a score above 0.01. Decoding the most toxic sentence (0.94) requires knowing to whom the pronoun 'his' refers. In context, 'his' refers to 'Sparcicists', Hitler's abstract representation of several named people he also refers to as 'political pigmies of the Revolution'.

For an explanatory dialogue, why the second sentence generates a score of 0.75 is not apparent. The third, fourth, fifth, seventh, eighth and ninth sentences also feature in Figure 19 as high-scoring identity attacks. Moreover, the highest scoring *I Have a Dream* sentence (0.67) is equivalent to the seventh most toxic sentence in *Mein Kampf.* While these *Mein Kampf* sentences are undoubtedly toxic to human interpretation, the algorithm generates unexpectedly low scores.


Figure 21. Detoxify outputs for sample sentences from *Mein Kampf*.

Figure 21 shows the scores from selected *Mein Kampf* sentences. These sentences are explicitly antisemitic and metaphorically use the word ‘parasite’, which features in genocidal language. The first sentence contains the phrase ‘negroid parasites’ and generates a high score of 0.85; a more toxic sentence with a score close to 1.0 is hard to imagine. The second and third sentences explicitly invoke an antisemitic trope that Jews dominate world economies; nevertheless, these sentences generate lower scores than the *I Have Dream* sentences in Figure 17. Such tropes should generate high scores for toxicity but are instead scored comparatively lower (0.27 and 0.22) than sentences from a non-violent text. As such, this transformer architecture for hate speech detection does not understand antisemitism.

The low score (0.01) for the fourth sentence is particularly interesting. By implication, ‘the best racial elements’ creates an ingroup abstract representation of all races Hitler considers to be ‘the best'. To tend to those best racial elements at the cost of everyone else is the central premise of racism, therefore, the toxicity score should be much higher. Instead, the score is among the lowest for toxicity even when compared with Luther King's non-violent text. As such, this sentence contains a subtle reference to ingroup elevation and outgroup othering that Detoxify does not account for in its toxicity score.

The fifth and sixth sentences interact to elevate the Aryan race while othering Jews as parasites. The sentence containing 'parasite' generates a score of 0.22 which is also lower than the *I have a* Dream score in Figure 17. More problematically, combining these two sentences in the final sentence reduces the toxicity score by 81% from 0.22 to 0.04. The combination of these two sentences is still horrifically antisemitic, but an increase in word count drowns out the toxicity the algorithm claims to detect. These results show how the decoding of both *Mein Kampf* and *I Have a Dream* by word co-occurrence for classifying a text completely misrepresents each orator's original meaning. While Detoxify is still in the research domain, the following section develops upon these findings to show how the problems persist in commercially available applications.

## 2.4 How Effective Are Quantitative Methods for Sentiment Analysis?

This next section uses sentiment analysis to assess the effectiveness of quantitative methods for decoding sentiment in a text. Sentiment analysis is an NLP application to detect whether a natural language input connotes positivity, negativity, or neutrality. In one survey, Mäntylä *et al.* (2018) observe how 'the outbreak of modern sentiment analysis happened only in mid- 2000s and focused on the product reviews available on the Web' (Mäntylä et al., 2018, p. 2). Developers have since applied sentiment analysis to domains beyond product reviews, such as terrorism (see: Mansour, 2018) and bullying (see: Beniwal & Maurya, 2021). The requirement for human annotation largely remains, Atteveldt *et al.* (2021) explain how 'a machine learning algorithm is used to create a statistical model based on [manually coded] training data which is then used to predict the sentiment of unlabelled texts' (Atteveldt et al., 2021, p. 4).

| Statement | Source | Text | IBM Watson | Google | Microsoft |
|---|---|---|---|---|---|
| 1 | *Mein Kampf* | The Aryan himself was probably at first a nomad and became a settler in the course of ages. | Neutral | Neutral | Neutral |
| 2 | *Mein Kampf* | The Jew has never been a nomad, but always a parasite, battening on the substance of others. | Negative (-0.85) | Negative (-0.5) | Neutral |
| 3 | *Mein Kampf* | The Aryan himself was probably at first a nomad and became a settler in the course of ages. The Jew has never been a nomad, but always a parasite, battening on the substance of others. | Neutral | Negative (-0.2) | Neutral |
| 4 | *I Have a Dream* | I have a dream that one day, down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of "interposition" and "nullification" – one day right there in Alabama little black boys and black girls will be able to join hands with little white boys and white girls as sisters and brothers. | Negative (-0.81) | Negative (-0.7) | Positive |

Figure 22. Comparison of sentiment scores for statements from *Mein Kampf* and *I Have a Dream*

To connect the commercial applications in this section to Detoxify in the previous, Figure 22 shows a summary of sentiment scores for sentences from Figure 21. These scores, generated from an open-source test[40], use algorithms from IBM[41], Google[42] and Microsoft[43]. Apart from Microsoft, the scores for statement two from *Mein Kampf* are negative, as might be expected. Nonetheless, combining the first two *Mein Kampf* statements in statement three generates a less negative score for what remains horrifically antisemitic sentiment. Somewhat counter-intuitively, the *I Have a Dream* statement has a similarly negative score to the antisemitic *Mein Kampf* statement. These commercial applications share the same processing problems as Detoxify.

Most problematically, there is no obvious way to assess the merit of each algorithm's output. Numerical scores alone do not explain how an algorithm generated them, and each vendor only gives access to technical documentation. While Microsoft's documentation does provide transparency notes to limit its algorithm's scope to product and service reviews, none of these vendors explain the annotation schema, training data or methodological documentation that would otherwise inform a rigorous explanatory dialogue. The Explainable AI movement would categorise these technologies in problematic terms as black-box algorithms.

Of particular interest to this sociotechnical assessment, and much like hate speech detection, there appears to be no defining methodology for sentiment analysis. Of note, sentiment analysis literature does not present an agreed unit of measurement for the numerical outputs of a sentiment analysis system. Contemporary literature tends to focus on improving technical architectures through advanced machine-learning techniques, and the first section explained how the literature generally uses f1-scores to evaluate architectural performance. From a more sociotechnical perspective, the literature does not appear to contain any attempts to connect these architectures to a defining sociological theory of sentiment. Sentiment analysis is undoubtedly a sophisticated technology; the following experiment using the hostile narrative corpus, however, raises questions about the explanatory value of the underpinning computational methods, especially when there is no corresponding methodology to explain sentiment.

---

[40] Anning (2020) Testing Sentiment Analysis, retrieved on 17th Feb 2023

[41] IBM Watson Natural Language Understanding Text Analysis, retrieved on 17th Feb 2023

[42] Google Cloud Natural Language, retrieved on 17th Feb 2023

[43] Microsoft Text Analytics, retrieved on 17th Feb 2023

### 2.4.1 Experiment 3: Detect the Ingroup and Outgroup

This open-sourced experiment[44] assesses the explanatory rigour of production systems to infer sentiment. It uses declarations of war from Bush and bin Laden to assess sentiment analysis by testing the following experimental hypothesis:

> *An orator's ingroup will generate positive scores, while their outgroup will generate negative scores.*

This experimental hypothesis assesses whether an NLP algorithm can decode the sentiment and orator asserts towards their ingroup and outgroup. The hypothesis is accepted if positive or neutral sentiment scores correlate with ingroup annotations and if negative scores correlate with outgroup annotations. Conversely, the hypothesis is rejected if positive or neutral scores correlate with outgroup annotations and if negative scores correlate with ingroup annotations. Bush's text is his ‘Address to Joint Session of Congress Following 9/11 Attacks’, made on the 20 September 2001; bin Laden's text is his ‘Declaration of Jihad Against the Americans Occupying the Land of the Two Holy Places’, published on 23 August 1996. The experiment uses IBM's sentiment analyser, Watson, since the marketing material claims an ability to ‘analyse target phrases in the context of the surrounding text for focused sentiment and emotion results’[45]. Since IBM provides an option to create custom models[46], the experiment assumes a text classification architecture. According to its marketing claim, Watson should provide useful outputs to test the hypothesis.

The experiment uses three annotation methods to create test data of target phrases: seed term, entity disambiguation, and inference. The seed term method annotates grouping according to words the orator associates with a named entity. For example, named entities associated with the terms ‘my fellow’ or ‘brethren’ were annotated as an ingroup. Conversely, entities related to such terms as ‘enemy’ were annotated as an outgroup. Entity disambiguation annotates multiple mentions of the same entity with the same annotation. Where the seed term ‘enemy’ identifies an outgroup, all other mentions of the same entity are given the same annotation. According with Watson's claim, inference annotations are a manual evaluation using the surrounding text as context when no direct seed term is available. The annotations are available online[47].

---

[44] Anning (2021) Obj 2 - detect ingroup and outgroup, retrieved on 17th Feb 2023

[45] IBM (2021) Watson Natural Language Understanding, retrieved on 17th Feb 2023

[46] IBM Watson NLU - Creating custom entities and relations models, retrieved on 17th Feb 2023

[47] Anning (2021) Dataset Annotations, retrieved on 17th Feb 2023

Ingroup scores for George Bush has a True Positive Score of 90% from a total of 20 Entities

| | Entity Text | Sentiment Score | Label | Grouping | Test Result |
|---|---|---|---|---|---|
| 0 | Pentagon | 0.979300 | positive | ingroup | pass |
| 1 | The Armed Forces | 0.979300 | positive | ingroup | pass |
| 2 | Great Britain | 0.965023 | positive | ingroup | pass |
| 3 | The Office Of Homeland Security | 0.897459 | positive | ingroup | pass |
| 4 | Muslims | 0.759801 | positive | ingroup | pass |
| 5 | New Yorkers | 0.755239 | positive | ingroup | pass |
| 6 | Mayor Rudolph Giuliani | 0.755239 | positive | ingroup | pass |
| 7 | Governor George Pataki | 0.755239 | positive | ingroup | pass |
| 8 | American | 0.680431 | positive | ingroup | pass |
| 9 | Lisa Beamer | 0.643663 | positive | ingroup | pass |
| 10 | Fbi Agents | 0.536320 | positive | ingroup | pass |
| 11 | Republicans | 0.430625 | positive | ingroup | pass |
| 12 | Democrats | 0.430625 | positive | ingroup | pass |

Outgroup scores for George Bush has a True Positive Score of 31% from a total of 13 Entities

| | Entity Text | Sentiment Score | Label | Grouping | Test Result |
|---|---|---|---|---|---|
| 0 | Arlene | -0.866308 | negative | ingroup | fail |
| 1 | Muslim | -0.832574 | negative | ingroup | fail |
| 2 | Christians | -0.818949 | negative | ingroup | fail |
| 3 | Jews | -0.818949 | negative | ingroup | fail |
| 4 | Islamic Movement Of Uzbekistan | -0.818724 | negative | outgroup | pass |
| 5 | Al Qaeda | -0.738021 | negative | outgroup | pass |
| 6 | United States Authorities | -0.648871 | negative | ingroup | fail |
| 7 | Taliban Regime | -0.573796 | negative | outgroup | pass |
| 8 | Taliban | -0.555951 | negative | outgroup | pass |
| 9 | The United States Of America | -0.535138 | negative | ingroup | fail |
| 10 | Americans | -0.460159 | negative | ingroup | fail |
| 11 | America | -0.384844 | negative | ingroup | fail |
| 12 | The United States | -0.307953 | negative | ingroup | fail |

Ingroup scores for Osama bin Laden has a True Positive Score of 80% from a total of 10 Entities

| | Entity Text | Sentiment Score | Label | Grouping | Test Result |
|---|---|---|---|---|---|
| 0 | Clinton | 0.835665 | positive | outgroup | pass |
| 1 | Brother Muslims | 0.642764 | positive | ingroup | pass |
| 2 | Israel | 0.443557 | positive | outgroup | pass |
| 3 | Us Enemy | 0.427792 | positive | outgroup | pass |
| 4 | Afghanistan | 0.427279 | positive | outgroup | pass |
| 5 | Gabriel | 0.387794 | positive | ingroup | pass |
| 6 | Messenger Muhammad | 0.000000 | neutral | ingroup | pass |
| 7 | Mujahidin Leaders | 0.000000 | neutral | ingroup | pass |
| 8 | Secretary William Perry | 0.000000 | neutral | outgroup | fail |
| 9 | Us Troops | 0.000000 | neutral | outgroup | fail |

Outgroup scores for Osama bin Laden has a True Positive Score of 45% from a total of 31 Entities

| | Entity Text | Sentiment Score | Label | Grouping | Test Result |
|---|---|---|---|---|---|
| 0 | United Nations | -0.935647 | negative | outgroup | pass |
| 1 | Jew | -0.883336 | negative | outgroup | pass |
| 2 | Jews | -0.803315 | negative | outgroup | pass |
| 3 | United States | -0.784973 | negative | outgroup | pass |
| 4 | Christians | -0.783726 | negative | outgroup | pass |
| 5 | Serbs | -0.778184 | negative | outgroup | pass |
| 6 | The United States | -0.765797 | negative | outgroup | pass |
| 7 | King Fahd | -0.726643 | negative | outgroup | pass |
| 8 | Russians | -0.718231 | negative | outgroup | pass |
| 9 | Jewish-Crusade Alliance | -0.672888 | negative | outgroup | pass |
| 10 | Us Defense Secretary | -0.663733 | negative | outgroup | pass |
| 11 | Army | -0.643972 | negative | ingroup | fail |
| 12 | Ulema | -0.638510 | negative | ingroup | fail |

Figure 23. Sentiment analysis results for detecting the ingroup and outgroup Bush's and Bin Laden's declarations of war.

Each declaration of war and list of target entities was passed to Watson; the full results are available online[48] and Figure 23 shows the top 13 results; these results do not support the experimental hypothesis since sentiment scores do not sufficiently correlate with each orator's ingroup or outgroup. As a summary of these results:

- Ingroup scores for George Bush have a success score of 90% from a total of 20 Entities.
- Outgroup scores for George Bush have a success score of 31% from a total of 13 Entities.
- Ingroup scores for Osama bin Laden have a success score of 80% from a total of 10 Entities.
- Outgroup scores for Osama bin Laden have a success score of 45% from a total of 31 Entities.

---

[48] Anning (2021) Testing Sentiment Analysis, retrieved on 17th Feb 2023

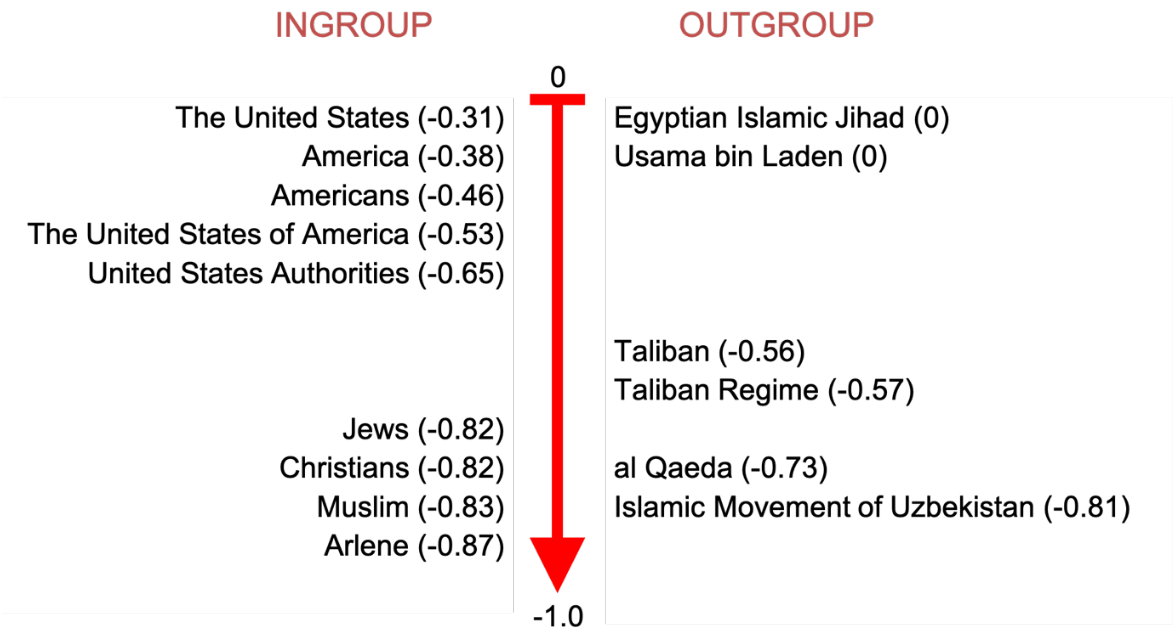



Figure 24. Outgroup sentiment scores for George Bush.

While Watson claims an ability to provide contextual sentiment scores, analysing these entities in context does not support this claim. Figure 24 depicts a summary of Bush's somewhat counter-intuitive outgroup scores. As expected, the named entities ‘al Qaeda’ (-0.73), ‘Taliban’ (-0.56), ‘the Taliban Regime’ (-0.57), and ‘Islamic Movement of Uzbekistan’ (-0.82) correlate with Bush's outgroups by generating negative scores between -0.56 and -0.82. Despite being Bush's ingroup, however, the named entities, ‘the United States’ (-0.31), ‘America’ (-0.38), ‘Americans’ (-0.46), ‘the United States of America’ (-0.53) and ‘United States Authorities’ (-0.64) generate overlapping negative scores between -0.31 and -0.65. Relative to the expected outputs, these scores do not distinguish between Bush’s ingroup and outgroup, which his speech clearly defines.

The following quote provides the context for mentions of Bush’s outgroups.

> *This group and its leader -- a person named Usama bin Laden - are linked to many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan. There are thousands of these terrorists in more than 60 countries.*

While co-occurring in the first sentence, ‘Usama bin Laden’ (0) and ‘Egyptian Islamic Jihad’ (0) generate neutral scores, but ‘Islamic Movement of Uzbekistan’ (-0.81) generates the most negative outgroup score. The Islamic Movement of Uzbekistan only gets one mention in the whole text but generates a more negative score than al Qaeda (-0.73) against whom Bush declares war. In context, Bush refers to each group as ‘these terrorists’ in the second sentence, but the algorithm fails to make the same link.

Bush praises his most negatively scored entity, ‘Arlene’ (-0.87), who is mentioned once in the following context,

> *And I will carry this: It is the police shield of a man named George Howard, who died at the World Trade Center trying to save others. It was given to me by his mom, Arlene, as a proud memorial to her son. It is my reminder of lives that ended, and a task that does not end.*

No words from this context explain why ‘Arlene’ is the most negative score, especially when less negatively scored outgroup entities like ‘terrorist’ or ‘enemy’ correlate with strong negativity.

The following two sentences contain the only mentions of ‘Christians’ (-0.82) and ‘Jews’ (-0.82):

> *The terrorists' directive commands them to kill Christians and Jews, to kill all Americans, and make no distinctions among military and civilians, including women and children.*

> *They want to drive Christians and Jews out of vast regions of Asia and Africa.*

While each named entity co-occurs with the verb ‘kill’, the first sentence expresses negative sentiment towards the concept of the ‘terrorists' directive’ and not ‘Christians’ or ‘Jews’. Similarly, the second sentence expresses negative sentiment towards the pronoun ‘they’, which in context refers to ‘the terrorists’. In contrast to Watson’s claim, none of the sentiment scores reflects the sentiment a President would express towards his country in response to a national tragedy and a declaration of war. A supplementary experiment provides some insight into why these counter-intuitive scores may occur.

### 2.4.2 Experiment 4: The Effect of Co-Occurring Words on Sentiment Scores

Entity: 'Americans' - sentiment -0.46

| | Positive, (11 Terms) | Negative, (14 Terms) | Neutral, (38 Terms) |
|---|---|---|---|
| 0 | course (0.68) | casualties (-0.7) | members (0) |
| 1 | state (0.31) | war (-0.93) | events (0) |
| 2 | thousands (0.53) | surprise attacks (-0.88) | Presidents (0) |
| 3 | directive (0.68) | civilians (-0.84) | come (0) |
| 4 | commands (0.4) | attacked (-0.78) | chamber (0) |
| 5 | win (0.64) | terrorists (-0.98) | report (0) |
| 6 | expect (0.51) | kill (-0.91) | known (0) |
| 7 | measures (0.66) | kill (-0.91) | wars (0) |
| 8 | protect (0.49) | civilians (-0.84) | the past 136 years (0) |
| 9 | thank (0.99) | hate (-0.98) | wars (0) |
| 10 | done (0.6) | fight (-0.71) | soil (0) |
| 11 | | war (-0.93) | known (0) |
| 12 | | battle (-0.92) | center (0) |
| 13 | | terrorism (-0.99) | city (0) |
| 14 | Average (0.59) | Average (-0.88) | Average (0) |

Entity: 'America' - sentiment -0.38

| | Positive, (16 Terms) | Negative, (11 Terms) | Neutral, (56 Terms) |
|---|---|---|---|
| 0 | sounds (0.56) | tragedy (-0.86) | touched (0) |
| 1 | honored (0.99) | forget (-0.66) | evening (0) |
| 2 | unity (0.57) | streets (-0.43) | see (0) |
| 3 | practiced (0.8) | enemy (-0.92) | joined (0) |
| 4 | counts (0.37) | atrocity (-0.98) | steps (0) |
| 5 | hope (0.41) | retreating (-0.9) | singing (0) |
| 6 | freedom (0.66) | forsaking (-0.84) | will (0) |
| 7 | uphold (0.96) | fight (-0.71) | playing (0) |
| 8 | values (0.87) | resolve (-0.25) | friend (0) |
| 9 | creativity (0.64) | died (-0.97) | crossed (0) |
| 10 | strengthen (0.92) | fear (-0.99) | ocean (0) |
| 11 | leaders (0.56) | | show (0) |
| 12 | lives (0.47) | | many millions (0) |
| 13 | possibilities (0.58) | | millions (0) |
| 14 | hopes (0.44) | | countries (0) |
| 15 | future (0.55) | | friends (0) |
| 16 | Average (0.65) | Average (-0.77) | Average (0) |

Entity: 'The United States' - sentiment -0.31

| | Positive, (2 Terms) | Negative, (1 Terms) | Neutral, (4 Terms) |
|---|---|---|---|
| 0 | respects (0.99) | sympathy (-0.62) | people (0) |
| 1 | support (0.63) | | nations (0) |
| 2 | Average (0.81) | Average (-0.62) | Average (0) |

Entity: 'the United States of America' - sentiment -0.54

| | Positive, (7 Terms) | Negative, (2 Terms) | Neutral, (14 Terms) |
|---|---|---|---|
| 0 | makes (0.43) | terror (-0.96) | tonight (0) |
| 1 | demands (0.8) | lies (-0.81) | following (0) |
| 2 | Deliver (0.37) | | United States authorities (0) |
| 3 | leaders (0.56) | | hide (0) |
| 4 | determined (0.73) | | land (0) |
| 5 | grant (0.84) | | will (0) |
| 6 | wisdom (0.51) | | age (0) |
| 7 | Average (0.61) | Average (-0.89) | Average (0) |

Figure 25. Sentiment scores for co-occurring nouns with specified named entities.

This supplementary experiment reveals problems with processing text by word co-occurrence. Each of the four tables in Figure 25 shows results for selected entities from Figure 24: 'Americans' (-0.46), 'America' (-0.38), 'the United States' (-0.31) and 'the United States of America' (-0.54). Each column shows the nouns and verbs co-occurring with each entity and their individual sentiment score in brackets. Individual scores focus on how the algorithm's architecture might score a word. Note neutral scores are excluded to focus on the negative and positive terms. The final row of each column is the average sentiment score for each co-occurring noun and verb.

The first observation is from the 'Americans' table where 'civilians' (-0.84), which should be a neutral term, is counter-intuitively scored more negatively than both 'casualties' (-0.7) and 'fight' (-0.71). Secondly, this table shows how many negative terms co-occur with each entity without expressing negative sentiment towards that entity. They co-occur in phrases specific to hostile narratives, such as 'the Enemy of America' – a prepositional phrase to describe an outgroup – and

'The terrorists' directive commands them to kill Christians and Jews, to kill all Americans' – an othering clause against the terrorist outgroup. Finally, the negative terms are valuable features of a hostile narrative; nevertheless, the average scores show how their weighting probably skews the outputs by their co-occurrence with a named entity. These contradictory sentiment scores are against expectations and provide limited explanatory value about Bush's and bin Laden's texts.

## 2.5 Discussion

The experiments presented in this chapter question the effectiveness of quantitative methods in NLP for social science applications like sentiment analysis and hate speech detection. The first section shows how Hall's theory of encoding and decoding presents a theoretical problem for using text classification. While a word's denotative meaning is reasonably fixed, its connotative meaning very much depends on the orator-to-audience relationship. In effect, denotative meaning is often formally defined in dictionaries while connotative meaning is colloquially understood in the minds of those who broadcast and receive messages. In turn, this dependency on an orator-audience relationship for connotative meaning explains the subjective elements of social science applications. The classification set of a text classifier represents connotative meaning and is encoded into training data through manual annotation. Nevertheless, NLP algorithms do not appear to account for this orator-audience relationship. The promise of NLP for social science applications, at least from a theoretical perspective, therefore, is with understanding denotative meaning; the quantitative methods reviewed in this chapter suggest text classification has limited understanding of connotative meaning.

Encoding and decoding also present a problem with annotating training data in text classifiers. Hate speech as a system of codes communicates the dominant position of hate groups which classifiers attempt to model. Hate speech detection literature attempts to take the oppositional position, which is similar to Luther King's position against racism in 1960s America. The reality for annotating training data, however, is likely to be in the negotiated position. In the annotation of training data, some annotators may interpret certain documents as hateful, while others may differently interpret the same document in a negotiated position. The potential for false negatives also arises from the use of metaphors and dog whistles that either humans incorrectly annotate or are too subtle for a classifier to detect. The idea of 'expert' annotators from the social sciences additionally assumes they can accurately decode racist language. In contrast, the experts for encoding racist language are the racist audiences with whom racist annotators communicate. The main problem, therefore, is about disagreement between an annotator and orators when annotators misinterpret their intended message.

Applying a confusion matrix with fixed categories additionally fixes connotative meaning for the relationship between the orator and annotators. Text classifiers in hate speech detection seek to model the relationship between a perpetrator of hate speech and their audiences. As the comparison of *Mein Kampf* and *I Have a Dream* shows, however, this model is fixed for all texts regardless of the orator's intent. The confusion matrix applies this fixed model by the occurrence of words in the dominant position of hate speech, however, they do not account for different decoding of the same words for different contexts. Consequently, false negatives arise from how orator use words non-hatefully that are hateful in other contexts. These words might be obvious, like negroid, or benign, like black, while metaphors are much less obvious. The problem is that language does not have fixed meaning, whereas confusion matrices treat language as a fixed variable. As such, assessing text classification using a confusion matrix re-applies the assumption of fixed interpretation of language that Hall has debunked over several decades.

From theory into a methodological perspective, the experiments in this chapter show how text classifiers conflate the aboutness and meaning of a text. As Maron (1961) initially observed, the occurrence of words provides a clue as to what a text is about, but these experiments show how they do not provide clues as to what it means. Both hate speech detection and sentiment analysis are about decoding meaning rather than aboutness. Decoding meaning is about understanding denotation and connotation, which are specific to the orator-audience relationship. For example, the word 'parasite' signifies a topic of disease and denotes an organism that survives to the cost of another organism. When the same word denotes Jewish people, as in *Mein Kampf*, it connotes antisemitism and becomes a feature of genocidal language. Moreover, the presence of such words as 'negro' might signal a topic of racism, as is the case with *Mein Kampf* and *I Have a Dream*, but it does not indicate a racist intention. In effect, text classification conflates the same computational method for two different tasks of determining aboutness and decoding meaning.

Beyond theoretical and methodological problems, the experiments of this chapter reveal practical processing problems for encoding natural language. This chapter has explained how standard pre-processing methods can misrepresent an orator's intended use of noun phrases and conjunctions. Text pre-processing that does not account for prepositional noun phrases or conjunctions, therefore, will skew an orator's original text. Moreover, the removal of stop words removes vital information when pronouns represent real-world entities. Resolving pronouns to their real-world entities is essential to detect othering in a hostile narrative. This misrepresentation of an orator's original text is a fundamental problem when using f1 scores to assess performance against training data. A high f1 score relative to the training data may mask a low f1 score relative to the orator's intended message. Accordingly, when training data misrepresents an orator's intended message, the f1 score misrepresents an algorithm's operational performance.

The assessment of presumed state-of-the-art hate speech detection and sentiment analysis algorithms also reveals how the assumed quantitative methods generate counter-intuitive outputs. They are assumed in the absence of technical documentation to explain how they work. Word embedding algorithms, regardless of technical sophistication, apply the distributional hypothesis to encode texts by word co-occurrence. Rather problematically, sentence scores change by adding or removing words in a sentence. Linguistic theory, conversely, explains how humans (at least subconsciously) interpret meaning from the grammatical relations. Moreover, some sentences are benign, but their toxicity is in their interaction with other sentences, which vector distributions do not capture. For sentiment analysis, the algorithm problematically produced different scores for the same named entities and words in their singular or plural form. Overall, these scores do not reflect the sentiment either Bush or bin Laden sought to portray.

## 2.6 Conclusion

The promise and limitations of word embedding for social science applications returns to Hall's encoding and decoding theory. Words must frequently co-occur for the distributional hypothesis to apply, and their co-occurrence will depend on the orator-audience relationship. The distributional hypothesis works well for descriptive statements, such as the relationship between countries and their capital cities; these words consistently co-occur to generate similar distances in a vector distribution. For opinions, however, the statements containing target words depend on the orator and are much less consistent. Moreover, frequent abstract representations of groups increase the variety of references to a named entity in a vector distribution. Texts written by western media for western audiences will broadly portray the dominant-hegemonic consensus of western thought. Applying a dataset representing this consensus, such as the Google News Corpus or a large language model like ChatGPT, to a non-Western individual or group in the oppositional or negotiated positions skews their intended meaning. Drawing on a cliché, words co-occurring with 'al Qaeda' in Google News are synonymous with 'terrorist', but the same term in a theoretical distribution of bin Laden's texts is more likely synonymous with 'freedom fighter'.

While NLP may promise to give machines the ability to understand natural language, the underpinning quantitative methods have limited utility for social science applications. The remainder of this thesis responds by connecting peace research and NLP to develop qualitative methods for the social science application of hostile narrative analysis. The next chapter presents the definitional problem of hate speech in more detail. As will be explained, even before developing a classifier, hate speech detection does not have an agreed methodology, therefore, humans struggle to define it for annotation. The chapter, therefore, presents cultural violence to fill the methodological gap.

# Chapter 3 Rethinking Hate Speech Detection as Hostile Narrative Analysis


Figure 26. The methodological framework of cultural violence.

This chapter responds to the research question 'how can integrating peace research and NLP enable the meaningful analysis of hostile narratives?' by rethinking hate speech detection as hostile narrative analysis. Hate speech detection responds to the propagation of hate over the Web and is attracting significant attention. The Web's role in hate speech propagation and how to respond is a common feature of Web Science conferences[49] and research (see: Squire, 2021; Zannettou et al., 2020). The UK's Alan Turing Institute's website records a range of academic interest in hate speech detection across research groups, journals, workshops and conference sessions[50]. In the commercial domain, the emerging UK Online Harms Bill will obligate web companies to tackle the propagation of hate on their platforms[51]. While attracting such attention,

[49] 13th ACM Web Science Conference - Session: Problematic Online Content, retrieved on 17th Feb 2023

[50] Online Hate Research Hub, retrieved on 17th Feb 2023

[51] Online Harms White Paper, retrieved on 17th Feb 2023

nonetheless, a 2019 Alan Turing Institute review finds ‘the field is beset with terminological, methodological, legal and theoretical challenges’ (Vidgen et al., 2019, p. 3). In support of this review, this thesis finds that hate speech has become a polysemous term giving rise to definitional ambiguity for the computational methods of hate speech detection.

This chapter rethinks what is known as the ordinary meaning of hate speech as a hostile narrative using a methodological framework derived from Galtung’s (1990) theory of cultural violence shown in Figure 26. The chapter begins by introducing an underpinning hypothesis of this framework to guide the explanatory dialogues about hostile narratives. The first section then explains how hate speech has become a polysemous term through an interdisciplinary transition from critical legal studies, to the social, political and computer sciences. The consequence is a disconnection between a defining theory of hate speech and the computational methods that questions the explanatory relevance of detection systems. The second section responds by using theories of violence from peace research to develop the methodological framework. The concluding section applies this framework to Bush’s and bin Laden’s declarations of war to verify its empirical relevance. This chapter is a novel attempt to theoretically develop cultural violence beyond Galtung’s (1990) paper.

The qualitative aspects of analysing hostile narratives are about detecting aspects of cultural violence in natural language using the proposed methodology. Using the framework as a methodological basis, the corresponding methods are about detecting what Galtung (1990) describes as the ‘Self-Other gradient’ in violence legitimisation. This gradient is about how an orator may elevate their ingroup while othering an outgroup to legitimise harm. Proposing the following underpinning hypothesis, therefore, provides a basis for explanatory dialogues about hostile narratives:

> *The steeper the Self-other gradient created by ingroup elevation and outgroup othering, the more legitimate violence against an outgroup becomes.*

This hypothesis has broad applicability to the different types of violence hate speech literature generally refers to, whether racism, sexism or homophobia. Violence against the targets of each type of violence could be verbal abuse, physical violence or more systemic harm arising from pernicious societal structures. Cultural violence is then about understanding how each violence type is legitimised. This framework and hypothesis, along with the method presented in the next chapter then provide a way to rethink the ordinary meaning of hate speech as a hostile narrative using cultural violence.

## 3.1 What Is Hate Speech Detection?

This section introduces the field of hate speech detection to tackle online abuse. The section begins by introducing the field of hate speech detection as a response to the propagation of hate over the Web. The section then introduces Matsuda's (1989) original legal characterisation of hate speech and explains how its ordinary meaning has become polysemous through an interdisciplinary transition from critical legal studies to the social, political and computer sciences. The subsequent section additionally explains how a disconnection between hate computational methods and a definitive sociological theory questions the explanatory relevance of hate speech detection systems. The proposal to rethink the ordinary meaning of hate speech as a hostile narrative in the remainder of this chapter is then motivated by improving explanatory dialogues for detecting such hostile language in natural language.

Note, the criticism and commentary to hate speech has generated a vast literature. Delagado (1993) provides a response to some good faith criticism (Delgado, 1993), while others engage in reactionary politics that draws upon complex emotions of 'resentment and ressentiment, blending anger, fear, nostalgic hope, betrayal, and a sense of perceived injustice' (Capelos & Katsanidou, 2018, p. 1272). Engaging with this literature would become a distraction to the focus of this thesis. Instead, this thesis takes a narrow review of Matsuda (1989) paper and her specific conceptualisation of hate speech. A response to both the good and bad faith criticism, therefore, is that hate speech has unhelpfully become a polysemous term.

### 3.1.1 Why Is Hate Speech Detection Required?

Hate speech detection is required to address the propagation of hate over the internet. The propagation of hate and development of information technology have an unfortunate co-existent relationship. During the 1980s, and prior to the Web's invention, Coates (1995) records that extremist groups used bulletin board systems to propagate hate among their members (Coates, 1995, p. 194). Citing StormFront[52] as among the first online hate groups, Schafer (2002) provides a content analysis of hate propagation websites compiled by HateWatch. This early non-profit organisation, now hosted by the Southern Poverty Law Centre, monitors the growing and evolving threat of online hate groups[53]. Schafer concludes with a finding applicable to any era of information technology development (Schafer, 2002, p. 80).

---

[52] StormFront (2021) StormFront.org, retrieved on 17th Feb 2023

[53] SPLC (2021) HateWatch, retrieved on 17th Feb 2023

> *...the internet offers users the opportunity to involve themselves in debating and advancing hate-based extremist ideologies with a high degree of anonymity and with considerable convenience...groups that are actively advancing their agenda may share ideas and resources at any hour of the day, from anywhere in the world, at very little expense.*

Contemporary, research generally focuses on examining the propagation of hate over such social media platforms as Twitter, Reddit and Facebook (Alorainy et al., 2018; Burnap & Williams, 2015a; Gomez et al., 2020; Nithyanand et al., 2017; Vidgen, Botelho, et al., 2020). Muller and Schwarz (2017) use Facebook data to investigate the link between online hate and violent crime. They find short-lived and localised bursts in sentiment 'have substantial effects on people's behaviour, and that social media plays a role in their propagation' (Muller & Schwarz, 2017, p. 33). To exploit the potential link between crime and online content, CrimeTelescope seeks to predict and visualise crime hotspots 'based on the fusion of heterogeneous urban and social media data' (Yang et al., 2018, p. 1325). In a particularly insightful analysis, Squire (2021) reveals an emerging industry of hate for profit whereby 'a regularly produced live stream show on a niche platform like DLive, far-right actors can earn over $100,000 in donations in less than a year' (Squire, 2021, p. 166).

Hate speech detection is attracting significant commercial investment. With £1.8m of Economic and Social Research Council (ESRC) funding, Cardiff University established its HateLab[54] in 2017 as a global hub for data and insight into hate speech and crime and has since generated £3m in research funding[55]. Kaggle, a Data Science research company, hosts regular competitions for detecting insults in social commentary with prizes of up to $10k. Like most commercial online platforms Facebook is already developing hate speech detection systems and hosts a challenge for developing algorithms that identify multimodal hate speech with a $100k prize fund[56]. Nevertheless, Whistle-blower Frances Haugen accuses Facebook of 'unquestionably making hate worse', and of being 'unwilling to accept even little slivers of profit being sacrificed for safety'[57]. The terminological, methodological, legal and theoretical problems Vidgen *et al*. (2019) identify are a problem for an industry which attracts so much interest and investment. As is not explained, these problems arise from the polysemy of hate speech as a term.

---

[54] Cardiff University (2021) HateLab, retrieved on 17th Feb 2023

[55] Cardiff University (2021) Professor Matthew L. Williams Bio, retrieved on 17th Feb 2023

[56] Facebook (2020) Hateful Memes Challenge, retrieved on 17th Feb 2023

[57] BBC (2021) Frances Haugen says Facebook is 'making hate worse', retrieved on 17th Feb 2023

### 3.1.2 Why Is Hate Speech A Polysemous Term?

'Hate speech' has become a polysemous term through an interdisciplinary transition from critical legal studies to the social and political sciences, and then to computer science. The term hate speech originated in critical legal studies in Matsuda (1989), but as Brown (2017) observes, it has since taken on a 'legal' and 'ordinary' meaning. Critical legal studies emerged in the 1970s and 1980s within the field of law and legal scholarship. Critical legal theorists critique the prevailing view of legal jurisprudence as an apolitical and objective, and instead argue that it is shaped by power relations, ideology, and social and historical context (Delgado, 1993, p. 744). Singer, a prominent critical legal theorist, claims 'Lawyers, judges, and scholars make highly controversial political choices, but use the ideology of legal reasoning to make our institutions appear natural, and our rules appear neutral' (Singer, 1984, p. 5). Critical legal theory became highly influential and led to the development of other such critical theories as radical feminism, and critical race theory (Delgado & Stefancic, 2017, p. 5).

Matsuda (1989) provides a legal meaning of hate speech to criminalise 'a narrow, explicitly defined class of racist speech, to provide public redress for the most serious harm, while leaving many forms of racist speech to private remedies' (Matsuda, 1989a, p. 2380). The focus of criminalising hate speech was US jurisprudence, and Matsuda was careful to respect the First Amendment rights of Free Speech. Matsuda (1989) provides three identifying characteristics of hate speech to distinguish harmful types of hate speech from other forms of racist and non-racist speech (Matsuda, 1989a, p. 2357):

- The message is of racial inferiority.
- The message is directed against a historically oppressed group; and
- The message is persecutorial, hateful and degrading.

The first characteristic is the primary identifier of hate speech, whereby all target group members are at once generalised as homogeneous and inferior. The second recognises structural elements of racism for which racist speech is a mechanism of subordination that reinforces historical injustice. While recognising historical injustice, Matsuda (1989) concedes, 'should history change course, placing former victim groups in a dominant or equalised position, the newly equalised group will lose the special protection suggested here' (Matsuda, 1989a, p. 2362). The third characteristic recognises the potential of written or spoken words to incite hatred or violence against a target. And where these characteristics focus on racism, Matsuda intended to take a more general view of hate speech for other minority groups.

To motivate the need for criminalising hate speech, Matsuda (1989) uses 'Outsider Jurisprudence' as 'a methodology grounded in the particulars of [people of colour's] social reality and experience' (Matsuda, 1989a, p. 2324). This methodology finds its origin in the field of critical legal studies and draws on the lived experiences of minority groups to challenge perceived power structures in legal norms and practices. The idea is to present stories as evidence for the harm that racist messages cause minority groups. As such, Matsuda (1989) presents the methodology as an attempt 'to know history from the bottom…from the fear and namelessness of the slave, from the broken treaties of the indigenous Americans' using hitherto ignored sources of 'journals, poems, oral histories, and stories from their own experiences of life in a hierarchically arranged world' (Matsuda, 1989a, p. 2324). In effect, these stories do not necessarily explain why a message is harmful, instead they explain the need for criminalising hateful language.

Criticism of hate speech in its legal characterisation generally takes a libertarian position of protecting Free Speech, especially on college campuses (Sandmann, 1995; Weinstein, 1991). Matsuda was aware of this tension; her 1989 paper sought to begin 'a conversation about the First Amendment that acknowledges both the civil libertarians' fear of tyranny and the victims' experience of loss of liberty in a society that tolerates racist speech' (Matsuda, 1989b, p. 2380). In addressing this tension, her paper contrasts United States' jurisprudence with international standards and the UK's Race Relations Act. She finds the UK act of that time makes restricting hatred a legitimate object of the law by properly criminalising certain forms of racist expression. As Matsuda acknowledges, the genuine debate centres around the limits of restricting speech, not around the basic decision to control racism since, 'racist hate propaganda is illegitimate and properly subject to control under the international law of human rights' (Matsuda, 1989b, p. 2345). This tension between Free Speech is an enduring feature of hate speech detection.

The ordinary meaning of hate speech arises from its transition from critical legal studies to the social and political sciences. For its ordinary meaning, Brown (2017) explains how 'different kinds of people who are not legislators, legal professionals or scholars of law use the term "hate speech" in countless different types of contexts about a tremendous diversity of phenomena' (Brown, 2017, p. 424). Accordingly, Brown (2017) observes two opposing positions in the ordinary meaning of hate speech (Brown, 2017, p. 425). The first is a progressive position where:

> *'hate speech' has been perhaps most often associated with liberal progressives, or people on the left of politics – who use it to highlight and problematise speech that they view as racist, xenophobic, homophobic, Islamophobic, misogynistic, disablist, or in some other way targeted at minority groups in ways that supposedly violate ideals of respect, solidarity, tolerance, and so forth.*

In an alternate view, Brown (2017) observes an opposing reactionary position whereby,

> *…political and religious conservatives repudiate such uses of the term and view them simply as crude attempts to close down meaningful debate on what they believe are the evils of open-border policies, the failures of multiculturalism as a social experiment, the lamentable decline of traditional moral values, political correctness gone mad, and so on.*

For this second position, Mondon and Winter (2020) note that Free Speech has become a 'reactionary tool' without any concrete legal basis to skew meaningful debate and push racist agendas (Mondon & Winter, 2020, pp. 75–79). With such reactionary uses of hate speech in all sides of political debates, definitions greatly depend on who uses the term.

### 3.1.3 How Does Hate Speech Detection Literature Define Hate Speech?

The adoption of hate speech as a concept in the computer sciences has only exaggerated its polysemy. Following the liberal progressive position identified by Brown, a literature review by Schmidt and Wiegand (2017) characterises hate speech as a 'broad umbrella term for numerous kinds of insulting user-created content' and references 18 developer-defined definitions of hate speech, some including the problematic prefix 'cyber' (Schmidt & Wiegand, 2017, p. 1). Subsequent literature shows how the absence of a generally accepted definition remains (Abro et al., 2020, p. 484; Chiril et al., 2022, p. 323; Fortuna et al., 2020, p. 6786; Kosisochukwu et al., 2020, p. 155; Mullah & Zainon, 2021c, p. 88366). In the absence of a generally agreed definition, developers tend to offer an application-specific definition of hate speech and bespoke annotation schemas for training data.

While critical legal studies use outsider jurisprudence, the polysemy of hate speech means the computer science literature does not appear to have a generally accepted methodology for detecting hate speech. In contrast to the use of victim stories, hate speech detection literature generally draws upon what researchers regard as far-right and extremist content from websites such as Twitter and Gab (Burnap & Williams, 2015b; Nithyanand et al., 2017), Reddit (Vidgen et al., 2021) or a combination of website content with offensive word lists (Nithyanand et al., 2017). Given the low prevalence of online abuse compared to all online messages, a key difficulty when creating datasets for annotation is collecting enough instances of the 'positive' class to be useful for machine learning (Guest et al., 2021, p. 1337; Vidgen et al., 2021, p. 2293). Consequently, 'more focused sampling strategies have to be applied which cause biases in the resulting datasets (Wiegand et al., 2019, p. 602). Yet, where these sources are public for online discussion, the choice of training data often reflects the contrasting positions observed by Brown (2017). Once the resultant training data are applied to real-world research tasks, 'definitions of abuse fast become embroiled in contentious debates around privacy, freedom of speech, democracy, discrimination, and the power of big tech companies' (Vidgen, Thrush, et al., 2020, p. 10).

The annotation of training data is similar to outsider jurisprudence in that the determination of hatefulness relies upon the subjective assessment of a team of annotators. Sang and Stanton (2022) note that subjectiveness in annotation usually arises from what some researchers refer to as 'latent content' which refers to the connotative meaning of words (Sang & Stanton, 2022, p. 426). For labelling training data, researchers develop an annotation schema which is then applied by a diverse team of annotators (see: Guest et al., 2021; Vidgen et al., 2021), but the choice of annotators is important. Waseem (2016) compared expert and annotator annotations to find that expert labelling outperforms amateur labelling (Waseem, 2016). Subjective assessments may also mean any determination of hatefulness for labelling may also depend less on an orator's intended meaning and more on annotator subjective interpretations (see: Davidson et al., 2019; Geva et al., 2019; Gonen & Goldberg, 2019; Wiegand et al., 2019). Without an agreed definition or methodology, hate speech detection relies on subjective assessment, which may not match the hate speech perpetrator to audience relationship.

Problems with the polysemy of hate speech and developing training data means the outputs of detection systems produce outputs of questionable explanatory rigour. In the first instance, outsider jurisprudence does not provide an explanatory theory for decoding hateful messages, it only provides an explanation for why hate speech should be criminalised. In hate speech detection, the polysemy of hate speech means computational methods are unconnected to a defining methodology. Without a defining methodology, therefore, the literature treats hate speech detection as a purely technical task using text classification. Reflecting what might the consequences for this disconnect between computational methods and theory, Lindgren (2020) suggests theoretical analysis 'may fade into the light of sparkling infographics' (Lindgren, 2020, p. 13). The next section responds by using cultural violence from peace research to fill the methodological gap arising from the polysemy of hate speech.

## 3.2 How Does Cultural Violence Explain Hostile Narratives?

This section introduces how Galtung's theory of cultural violence explains violence legitimisation in hostile narratives. Galtung (1990) defines cultural violence as:

> *Those aspects of culture, the symbolic sphere of our existence – exemplified by religion and ideology, language and art, empirical science and formal science (logic and mathematics) – that can be used to justify or legitimise direct or structural violence.*

This section draws upon this definition to develop the methodological framework of cultural violence shown in Figure 26 as one way to enable the *human* understanding of a hostile narrative. Each aspect of culture and the different types of violence from of this definition provide a structure for its explanation. This subsection begins by elaborating on Galtung's three types of direct, structural and cultural violence. The subsequent subsection connects each aspect of culture to academic literature as a basis for the methodological framework. The section then continues by augmenting the framework with Tajfel's and Turner's social identity theory from social psychology (Tajfel, 1974; Tajfel & Turner, 1979). By doing so, cultural violence itself becomes a unifying idea for transdisciplinarity. The remaining sections then develop upon the hostile narrative analysis method derived from this framework using Bush's and bin Laden's declarations of war. The emphasis on *human* understanding applies to rigorous explanatory dialogues about hostile narratives into which machines provide meaningful inputs.

### 3.2.1 How Does Galtung Define Violence?

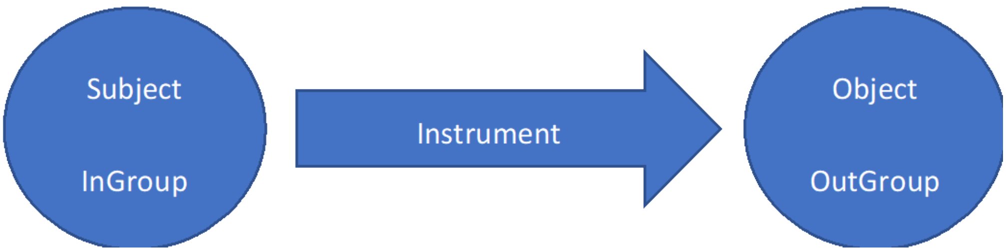



Figure 27. Galtung's model of violence.

Galtung defines violence itself as 'any avoidable insult to basic human needs, and, more generally, to sentient life of any kind, defined as that which is capable of suffering pain and enjoying well-being' (Galtung & Fischer, 2013a, p. 35). This broad definition is modelled in Figure 27 to show the essential ingredients of violence as a subject, object, and instrument for enacting harm. Galtung conceives the subject as the perpetrator of violence, while the object can be individuals or collectives of human beings who have been made nameless, faceless, de-individualised (Webel & Galtung, 2007, p. 23). This thesis extends Galtung's theory of violence with a social psychology approach pioneered by Tajfel and Turner to explain group formation and intergroup relations between an ingroup (subject of violence) and outgroup (object of violence).

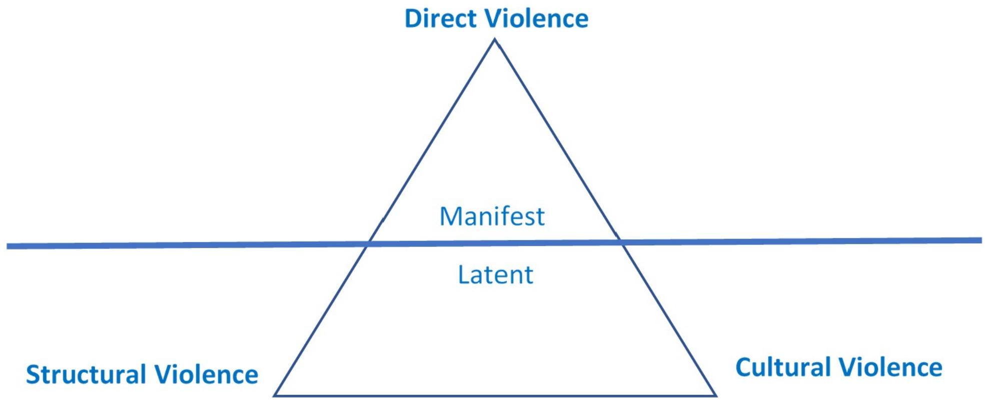



Figure 28. Galtung's violence triangle.

Galtung (1990) situates three types of direct, structural, and cultural violence within a triangular framework shown in Figure 28. These different types of violence go beyond the general understanding of violence as a physical act (Webel & Galtung, 2007, p. 23). 'Direct violence is harming others with intention; structural violence is the harm done by socio-political structures and decisions that deprive access to the basic needs necessary for fulfilling one's potential; cultural violence is the cultural justification of direct and structural violence' (Galtung & Fischer, 2013c, p. 151). Returning to the twin goals of peace research from the thesis introduction, preserving 'negative peace' refers to the minimising of direct violence, such as war, while promoting 'positive peace' refers to tackling structural violence (Galtung, 1969, p. 183).

'Instrument', as shown in Figure 27 refers to the device used by the subject against the target. Direct violence is about 'hurting or killing people' and 'includes verbal violence' (Galtung & Fischer, 2013a, p. 22). The instruments of such physical acts can be anything from a fist between individuals to a nuclear missile between countries. Accordingly, the goal of negative peace is about preventing such acts from taking place. In contrast, the instruments of structural violence might be legislation that intentionally harms an outgroup, as with 1967's Trade with Africa Act that enabled the slave trade. Accordingly, the goal of tackling structural violence removes such pernicious legislation to create institutions that enable equitable intergroup relations. In the case of verbal hostility, certain words trigger neural structures to produce social emotions for moderating status, belonging in and exclusion from social groups (Martin, 2018, p. 34). Relevant to this thesis, the instruments of verbal hostility are narratives, stories, and words that hostile narrative analysis seeks to detect and analyse.

As an 'avoidable' insult to human basic human needs, the intentionality of harm is the crucial difference between a violent and non-violent act: in Galtung's terms, violence must be intended and attributable to a person for it to exist (Galtung & Fischer, 2013c). For direct violence, the intended use of fists or nuclear missiles is easily attributable to whoever initiated the act. While the intended enactment of harmful legislation is also obviously attributable, structural violence also arises from numerous acts of omission. Galtung argues that inaction is as equally violent as intended action; for example, to not revoke legislation attributable to structural violence is just as violent as enacting that legislation in the first place (Galtung & Fischer, 2013b, p. 12). While words and stories are attributable to a person, the challenge of hostile narrative is to detect violent intention contained within them. As is now explained, intention can be inferred from what cultural violence refers to as the Self-other gradient.

### 3.2.2 How Do Each Aspect of Culture in The Cultural Violence Definition Feature in Violence Legitimisation?

This section elaborates on how each aspect of culture – religion and ideology, language and art, and empirical and formal science – features in violence legitimisation by creating a 'Self-other gradient' between an orator's ingroup and their outgroup. Galtung's (1990) Self-other gradient is about how an orator intentionally elevates a 'Chosen People' (Self) above those deemed 'lower down the scale of worthiness' (Other) (Galtung, 1990, p. 302). This gradient is a central theme of studying structural violence. Matsuda *et al.* (1993) explain racism as the 'structural subordination of a group based on an idea of racial inferiority (Matsuda et al., 1993, p. 36), while Augoustinos & Reynolds (2001) note how 'the power one group has over another transforms race prejudice into racism' (Augoustinos & Reynolds, 2001, p. 4). Accordingly, the theoretical framework in Figure 26 a range of literature to explain how aspects of each cultural domain create a Self-other gradient through elevation and othering.

The first domains of religion and ideology refer to how cognitive belief systems guide morality. Literature on religion and ideology is vast; this explanation draws upon three scholars who seemingly best summarise Galtung's ideas. In his theological analysis, Wright (2010) describes scriptures from Christianity, Judaism and Islam as 'maps of the landscape of religious tolerance and intolerance, maps that amount to a kind of code for the salvation of the world' (Wright, 2010, p. 322). Equally, for van Dijk (1998), 'ideologies allow people as group members, to organise the multitude of social beliefs' about 'good or bad, right or wrong, for *them*, and to act accordingly' (Dijk, 1998, p. 8). As Martin (2018) observes, while the moral codes underpinning ideologies emerge from within societies, religion adds belief in an external supernatural deity as the arbitrator of those codes (Martin, 2018, p. 163). Whether divine or secular, orators use these codes as cognitive frameworks for moral judgement to elevate the Self in contrast to the other and create a gradient between each.

The field of Semiotics explains how the third and fourth cultural domains of language communicate moral codes through linguistics and visual systems. De Saussure defined language (langue) as ’a system of signs that express ideas’ and introduced Semiotics as a science to ‘investigate the nature of signs and the laws governing them’ (Saussure, 1916, p. 16). Semiotics has since become a theory of meaning (or ‘signification’) that analyses the network of relationships between terms of linguistic systems, referred to as vocabularies, within the minds of a speech community (Knapp et al., 2008, p. 231). Semantic analysis (applied in Chapter 4) is the specific treatment of words as signs. An orator’s use of contrastive terms in a vocabulary signifies the Self-other gradient. Words like ‘friends’ contrast with such antonyms as ‘enemies’ to categorise the Self and the other. An orator creates a gradient between each by using framing words that signify such ‘moral oppositions’ as ‘good’ and ‘evil’ or ‘right’ and ‘wrong’ (Chandler, 2017, p. 114). Accordingly, language and art communicate moral codes to create the gradient between the Self and other.

For the fifth cultural domain, empirical science, Galtung (1990) describes how a ‘law of comparative advantage’ is embedded into economic activity to create status grievances. The idea of ‘Relative Deprivation’ (RD), first introduced by Stouffler *et a*l. (1949) and elaborated by W. G. Runciman (Runciman, 1966; Webber, 2007, 2021) explains how people derive their status in comparison to others. Gurr (2010) later used RD to explain how a decline in status compared to expectations may lead to violence (Gurr, 2010). Conversely, in the ‘Greed and Grievance’ debate, Collier and Hoeffler use macroeconomic analysis to show how pursuing power and profit to gain status can lead to violent rebellion (Collier et al., 2009; Collier & Hoeffler, 1998, 2004). Arising from economic activity, therefore, is the potential for violence in response to the perceived decline or pursuit of status.

The sixth cultural domain, formal science, is explained by prospect theory to understand how people rationalise choices under risk. From 1979, Kahneman and Tversky pioneered this theory to challenge existing economic assumptions giving rise to the field of behavioural economics. Kahneman (2011) subsequently explains how the quick-thinking brain develops heuristics (roughly rules of thumb) to rationalise the world, often at the expense of more considered thinking (Kahneman, 2011, p. 7). Prospect theory has become a rejoinder to the rational choice models of behaviour prevalent in the social sciences (Webber, 2021). Arising from heuristics is the potential for biases manifest as discriminatory beliefs about the Self and Other. A Self-other gradient itself could be considered a heuristic used by either a person or group to rationalise harm.

Aspects of these six cultural domains of religion and ideology, language and art, and formal and empirical science create a gradient between the Self and Other through elevation and othering. As belief systems, religion and ideology provide cognitive frameworks of moral codes which elevate the Self in contrast to the Other. Semiotics explains how linguistic and visual systems signify moral codes to communicate elevation and othering. Empirical science explains how economic activity may create elevation and othering through perceived status differences. Formal science, as behavioural economics, consolidates the other cultural domains to show how violence against the Other might be pre- or post-rationalised. According with the underpinning hypothesis, the steeper the gradient between the Self and Other created by elevation and othering, the more legitimate direct or structural harm becomes. With the basis for cultural violence established, the following subsection elaborates on the Self and Other using social identity theory.

### 3.2.3 How Does Social Identity Theory Augment Cultural Violence Theory?

Social identity theory augments cultural violence theory with an understanding of groups and group formation in violence legitimisation. As the originators of social identity theory, Tajfel and Turner (1979) describe groups as (Tajfel & Turner, 1979, p. 283):

> *...a collection of individuals who perceive themselves to be members of the same social category, share some emotional involvement in a common definition of themselves and achieve some degree of social consensus about the evaluation of their group and their membership of it.*

Recalling the introduction, emotional involvement and strict definitions of group membership strongly feature on the nationalism of Anderson's Imagined Communities. For group membership, Goffman (1959) theorises how individuals gain their sense of Self through interactions with other members. As Johnstone (2003) describes, in many respects, personal experience narratives referred to by Labov and Waletzky (1997) are how people make sense of themselves as individuals and group members. The process of violence legitimisation then begins with the representation of the Self and others as members of differentiated groups.

The process of 'social categorisation' leads to group formation and the abstract representation of groups in natural language. As Hogg (2008) observes, 'categorisation renders the world more predictable' thus facilitating planning for effective action (Hogg, 2008, p. 74). When social categories become salient, people see themselves and others less as individuals and more as homogenous group members, whether an ingroup or outgroup. Some members gain the status of group prototypes. 'The prototype is not an objective reality, but rather a subjective sense of the defining attributes of a social category that fluctuates according to context' (Hornsey, 2008, p. 208). This prototype then forms the abstract representation of groups in language. As Galtung (1990) explains, 'when the other is not only dehumanised but has been successfully converted into an 'it', deprived of [humanity], the stage is set for any type of direct violence' (Galtung 1990: 298). But these prototypes can also be the most favoured group member as an example for others to follow. In many respects the abstract representation of groups through such prototypes as 'hero' or 'enemy' provides heuristics for evaluating the social world and represent the narrative truth of hostile narratives.

In terms of group evaluation, Hogg (2016) explains how 'intergroup differentiation' suggests people are concerned with ensuring their ingroup is positively distinctive, clearly differentiated from and more favourably evaluated than identified outgroups (Hogg, 2016). Differentiation often manifests when people define themselves by who they are *not* rather than who they are; 'a sense of identity is founded upon a distinction between us and the rest of the world' (Chandler, 2017, pp. 108–114). Experiments by Turner and Reynolds (2011) subsequently find that having self-identified with a group, individuals assign more resource to their ingroup than their outgroup (Turner & Reynolds, 2011). They note how attributes of group formation may be as trivial as shared scores in a maths assessment. As Tajfel and Turner (1979) note, in contrast to ethnocentric notions of group formation, experiments yield 'the basic and highly reliable finding' that 'trivial, ad hoc categorisation leads to ingroup favouritism and discrimination against the outgroup' (Tajfel & Turner, 1979, p. 39).

While differentiation is a reoccurring theme in theories of intergroup relations, hate speech detection research seems to only consider othering and exclude ingroup elevation (see: Alorainy et al., 2018; Cao et al., 2020; Fortuna, 2018; Poletto et al., 2020; Schmidt & Wiegand, 2017). This omission of ingroup elevation further questions the explanatory rigour of hate speech detection systems. In contrast, the cultural violence framework as an underpinning methodology is connected to established literature to provide an explanatory basis for explanatory dialogues about hostile narratives. As such, this framework represents how a human 'understand' processes of violence legitimisation in natural language. The following chapter explores the necessary computational methods to enable this human understanding. The next section verifies this framework using Bush's and bin Laden's declarations of war from the War on Terror.

## 3.3 How Does Cultural Violence Feature in Bush's And bin Laden's Declarations of War?

This analysis of Bush's and bin Laden's declarations of war seeks to verify the framework of cultural violence depicted in Figure 26 and explained in Section 2. The section begins by explaining why these texts fit the definition of narrative from the introduction to this thesis. As texts that seek to rationalise and respond to violent events, the defining narrative clauses of each text refers to the action and response. This explanation will show how Galtung's conception of violence features in these clauses. The subsequent subsection then explains how methodological framework from Figure 26 features in each orator's violence legitimisation. The subsection will show how religion and ideology legitimises each orator's declaration of war, whether 'jihad' or 'War on Terror'.

This section seeks to represent the narrative truth of each text, the remainder of the thesis then seeks to develop the computational methods to recreate elements of this representation. This analysis, nevertheless, does not seek to engage with a critique of each narrative. There is a rich literature in the critique of both orator's texts and actions during the War on Terror years. The Changing Character of War project from Strachan and Scheipers (2011) is a comprehensive example (Strachan & Scheipers, 2011). This analysis, on the other hand, represents an explanatory dialogue about how each orator legitimises violence in each text and seeks to represent a disinterested analysis. The quotes from each text represent inputs to a dialogue that seek to understand how each orator legitimised violence, not to assess the moral colour of each text. Having created a representation of each text, the enquirer and explainer of a dialogue may draw upon other literature to engage with the historic truth as a critique of each narrative.

### 3.3.1 Why Are Bush's and Bin Laden's Texts Hostile Narratives?

As a reminder from the introduction, a narrative is defined by having at least two narrative clauses whose order cannot be switched without changing the text's meaning. This analysis explains how Bush's and bin Laden's texts fit this definition by selecting two sequential clauses and showing how switching their order changes each text's meaning. The content and context of these clauses also reveals Galtung's violence model from Figure 27, and the violence triangle in Figure 28. As will be shown, the plot for each narrative is remarkably similar despite each orator being opponents of the same conflict.

Two sequential narrative clauses from bin Laden's text are as follows:

> *Clause 1: {Your blood}* $_{subject}$ *{has been spilled}* $_{predicate}$ *in {Palestine}* $_{object1}$ *and {Iraq}*$_{object2}$.
>
> *Clause 2: {Your brothers}* $_{subject1}$ *{and sons}* $_{subject2}$, *{the sons of the two holy mosques}* $_{appositional\ modifier}$, *{have launched}* $_{predicate}$ *{the jihad}* $_{object}$ *for {the sake of God's cause to expel the occupying enemy from the country of the two holy mosques}* $_{preposition}$.

These two clauses form the basic plot of bin Laden's narrative whereby his declaration of jihad is clause 2 responds to aggression by his enemy in clause 1. In context, 'your blood' as the subject of the clause figuratively refers to Muslim's blood, which in turn refers to bin Laden's ingroup. That the blood of his ingroup has been ‘spilled’ in Palestine and Iraq is in addition to 11 other acts of direct violence in the next sentence. He comments, ‘the image of that dreadful massacre in Qana, Lebanon, is still vivid in one's mind, and so are the massacres in Tajikistan, Burma, Kashmir, Assam, the Philippines, Fatani, Ogaden, Somalia, Eritrea, Chechnya, and Bosnia-Herzegovina’. His conclusion from these events is that Muslims are ‘the main target of the Jewish-crusade alliance aggression’ to which his declaration of jihad responds.

The use of familial terms in as the objects of clause 2 invokes the idea of social categorisation from social identity theory. Clause 2 refers to the jihad launched by ‘brothers and sons’ of Muslim people in response to the aggression cited in clause 1. Using the social categories of ‘brothers and sons’ in the subject of this clause to denote the soldiers of the mujahedeen connotes an affection towards them. Additionally referring to them as ‘the sons of the two holy mosques’ in the appositional modifier connotes a religious legitimacy. Much like bin Laden, Bush refers to his military in familial terms and makes religious references. He comments, ‘A Commander-in-Chief sends America's sons and daughters into battle in a foreign land only after the greatest care and a lot of prayer’. Using familial social categories is a way to affectionally elevate soldiers thereby lionising them as heroes to create positive prototypes of the ingroup.

The use of occupation in ‘occupying enemy’ invokes structural violence that bin Laden amplifies with religion. ‘The country of the two holy mosques’ is bin Laden’s way to refer Saudi Arabia. The two mosques are the holy sites of Al Masjid Al Haram in Mecca and the Prophet Mohammed’s Mosque in Medina. Denoting Saudi Arabia in such pious terms not only connotes a religious sacristy to the country bin Laden seeks to protect, but also strengthens the sense of wrongdoing by the ‘occupying enemy’. The reference to ‘God’s cause’ also makes God the arbiter of the righteousness of bin Laden’s jihad. Bush similarly sanctifies America with the often-used phrases, ‘May God bless you all’ and ‘God bless America’. And while America’s secular constitution is Bush’s primary arbiter of morality, he does ask for religious guidance for his War on Terror in such phrases as, ‘may God grant us wisdom’. Each orator uses religion to sanctify the countries they seek to protect and use God as an arbiter for promoting their sense of righteousness.

In relation to the ‘occupation of Jerusalem’, bin Laden’s and Bush’s speeches are a contest for the hegemonic-dominant code of Israel’s legitimacy. In a United Nations speech, Bush (2001) seeks to uphold the dominant-hegemonic code enshrined in the various United Nations resolutions that rightfully legitimise Israel’s existence. He declares to be ‘working towards a day when two states, Israel and Palestine, live peacefully together within secure and recognise borders as called for by the Security Council resolutions (UNSCR)’. Bin Laden (2002), however, contests the dominant-hegemonic position in his ‘Letter to America’ by encoding Israel’s formation as an ‘occupation’ with 'years overflowing with oppression, tyranny, crimes, killing, expulsion, destruction and devastation'. These references to both structural and direct violence signal the oppositional position to Israel's legitimacy, and implicitly encode UNSCRs as instruments of structural violence against Muslims in favour of Israel. Indeed, bin Laden declares that 'the creation and continuation of Israel is one of the greatest crimes…which must be erased'.

According to bin Laden, the instruments of structural violence in the occupation of Saudi Arabia are '[the Saudi Regime's] failure to have recourse to the Shari'ah, its confiscation of people's legitimate rights, the opening of the land of the two holy mosques to the American occupiers, and the arbitrary jailing of the true ulema, heirs of the Prophets'. These instruments are legal in the sense of replacing Shari'ah law with 'temporal law', the removal of legally bestowed rights and the jailing of people bin Laden supports. They are also economic through America's relationship with the Saudi Government. Where the imposition or ignoring of such perceived injustices constitute Galtung's (1969) definition of structural violence, bin Laden's argues that they were raised to the Saudi government by prominent Saudi's through 'petition after petition and memorandum after memorandum', but were met with 'rejection, disregard, and ridicule'. The consequent harm to Saudis was 'taxes, duties, and excises imposed on the public' through oil policies that favoured the American's over Saudi Arabia's interests.

Two sequential clauses from Bush's declaration of war reveal a somewhat similar plotline.

> *Clause 3: {On September the 11th} $_{preposition}$ {enemies of freedom} $_{subject}$ {committed} $_{predicate}$ {an act of war} $_{object}$ against {our country} $_{preposition}$.*
>
> *Clause 4: On my orders, {the United States military} $_{subject}$ {has begun} {strikes} $_{object}$ against {Al Qaeda terrorist training camps} $_{preposition1}$ and {military installations of the Taliban regime} $_{preposition2}$ {in Afghanistan} $_{preposition3}$.*

While religion does feature, political ideology is a central theme of Bush's violence legitimisation. Clause 3 refers to the attacks on America on 11$^{th}$ September 2001, which themselves are acts of direct violence. In this clause, 'enemies of freedom' are the aggressor and 'our country' (America) is the victim. Encoding these attacks as an 'act of war' legitimises the 'War on Terror' under international law. Indeed, Bush later invokes the article 7 of the NATO charter whereby, 'an attack on one is an attack on all'. This NATO article then becomes an instrument of violence by legitimising a military response from all contributing nations to the NATO alliance. Where freedom is a core idea of America's constitution, an 'enemy of freedom' is antithetical to the political values of Bush's ingroup. Indeed, Bush's declares 'Tonight we are a country called to defend Freedom' in a previous clause and calls the subsequent military response in Afghanistan 'Operation Enduring Freedom'.

Clause 4 is from Bush's address to the nation on 7th October 2001, which announces the start of Operation Enduring Freedom. Much like bin Laden's clauses, therefore, this second clause is a violent response to aggression in the first clause. The targets of these strikes in clause 4 are assets of al Qaeda and the Taliban. Perhaps identifying the targets as assets rather than people makes the violent action more palatable. Bush also refers to the NATO members of the alliance using the social categories of friendship. He comments, 'We are joined in this operation by our staunch friend, Great Britain. Other close friends, including Canada, Australia, Germany and France, have pledged forces as the operation unfolds'. While encoding these countries as 'friends' in the imagined communities of international relations is compelling in narrative truth, the reality may rightly be subject to question, do America and the UK have a special friendship?

The term '9/11' commonly denotes the attacks in New York and Washington, which themselves are acts of direct violence. He connotes of wrongdoing by encoding 9/11 as, 'despicable acts of terror', 'acts of murder' and 'these evil acts'. In relation to the model of violence in Figure 27, the perpetrator is al Qaeda, and the target is America. The instrument of violence is multi-layered. Most obviously, the instruments were the passenger jets that al Qaeda used to destroy the Twin Towers in New York and the Pentagon in Virginia. More subtly, however, the instrument could also be the terror of innocent deaths felt by American citizens. As such, in relation to Galtung's definition of violence, the 'avoidable insult to basic human needs' is not just killing of other human beings, but also the sorrow felt by those who lost loved ones and terror felt by people who became less able to live the life they choose because of terrorising acts.

The obvious instruments of Bush's War on Terror are military operations, but Bush also uses the Patriot Act of 2001 as an instrument of structural violence. The Act gave 'intelligence and law enforcement officials important new tools to fight a present danger' by changing 'the laws governing information-sharing', and by allowing 'surveillance of all communications used by terrorists, including e-mails, the Internet, and cell phones'. Objectively, the change created by the Patriot Act exemplifies structural violence as it enabled the US state to intrude on civil liberties. Subjectively, the legitimacy of this change is open to debate. While some argue the Patriot Act was necessary, others later argued, 'George Bush tipped the balance too far from liberty towards security, and it has stayed there under Barrack Obama'[58]. The perpetrator, therefore, was the American Government while the victims were people unfairly targeted by surveillance, including

[58] Economist (2013) Liberty's lost decade, retrieved on 17th Feb 2023

American citizens. The determination of harm of the Patriot Act is a trade-off between civil liberties and national security, and the assessing the moral colour of this harm changes over time.

This analysis explains how bin Laden's and Bush's text confirm to the definition of a narrative. These clauses (and many others) cannot be reversed without changing the plot of either narrative. In his narrative truth, bin Laden and his fellow Muslims are victims of an occupation of Saudi Arabia. In the reversal of these clauses, Bin Laden's jihad to liberate Saudi Arabia would precede the perceived occupation. As such bin Laden and his fellow Muslims would become the aggressor rather than victim, thereby changing the plot. Much like bin Laden's clauses, a reversal of Bush's clauses would change the plot of his narrative. In Bush's narrative, America is the victim of aggression by al Qaeda and the Taliban. To launch Operation Enduring Freedom before 9/11 would make the Taliban and al Qaeda victims of American aggression thereby reversing the meaning of his narrative. Having now established that these texts conform to the definition of narrative, the subsequent subsection shows how the methodological framework of cultural violence commonly applies to how each orator sought to legitimise violence.

### 3.3.2 How Does Elevation and Othering Feature in Bush's and bin Laden's Declarations of War?

As this subsection now explains, elevation and othering features in each text through the way in which each orator legitimises their violent campaign, whether 'jihad' or 'War on Terror'. As will be explained, both narratives feature ideology and religion as a way of elevating their ingroup and legitimising violence by othering their outgroup. The noun phrases each orator uses to represent their ingroup and outgroups also develops upon the use of social categorisation mentioned above. Rather than being representations of real people, these noun phrases represent characters in the story of each narrative. They are abstract representations of reality. Both orators then use religion and ideology to compartmentalise each character into categories of good or bad and judge their actions as either right or wrong. Bin Laden also invokes a status grievance for legitimising his jihad. As with the analysis of the narrative clauses above, the way the framework features in each orator's narrative is remarkably similar.

Bin Laden creates a victim narrative for his ingroup by using a perceived decline in Saudi Arabia's economic standing to express a status grievance against his outgroup. He amplifies this grievance using a mixture of economic ideology and religion. Bin Laden's text contains few phrases that positively identify his ingroup. He does use the phrases 'Muslim brothers' and 'Brother Muslims worldwide', and he uses familial social categories to create a sense of belonging in his ingroup. His ingroup, therefore, is implied from these noun phrases and his own identity. He expresses the status grievance as ‘economic decline, high prices, massive debts, and overcrowded prisons’ in Saudi Arabia. This decline strongly represents the lowering of status when compared other countries. Correspondingly, bin Laden discusses an apparent drop in purchasing power of the Saudi riyal ‘in comparison with other major currencies’ and ‘hundreds of millions and even billions of riyals’ in debts owed to major traders by the Saudi government.

Bin Laden represents his primary outgroup in abstract terms as a ‘Jewish-crusade alliance’ to which he adds various named entities during the narrative. He creates a gradient between his ingroup and this outgroup by attributing the ‘injustice, repression, and aggression that have befallen Muslims’ to this ‘alliance of Jews, Christians and their agents’. He invokes further injustice by attributing the above-mentioned 11 massacres to ‘a conspiracy by the United States’ under the cover of ‘the unfair United Nations’. More directly, he claims ‘Christian armies of the Americans and their allies’ occupied Saudi Arabia, which was ‘one of the worst catastrophes to befall the Muslims since the death of the Prophet’. Bin Laden even makes the Saudi Government and agent of this alliance in the phrase, ‘the main enemy, namely the Israeli-American alliance, through the Saudi regime, its agent in the country’. His incitement of violence is a call for ‘efforts…to kill him, fight him, destroy him’ where ‘him’ refers to the ‘US enemy’. He goes as far as calling America ‘the main disease and cause of the affliction’ against Saudi Arabia.

Economic ideology features in the use of economic entities to create a threat to status. Bin Laden uses oil as an ingroup status symbol that he claims is under threat from his outgroup. According to bin Laden, Saudi Arabia is ‘an important economic power in the Islamic world because it has the largest oil reserves in the world’. He claims ‘the presence of the crusader and American military forces in the Islamic Gulf states…represents the greatest threat’ to these oil reserves. He goes as far to say that his outgroup’s presence is an affront to people’s ‘religion, feelings and dignity’ and has ‘driven them to armed struggle’. He asks for the Mujahedeen to preserve Saudi’s wealth by not using this oil in battle since it is ‘a great Islamic wealth’ and an ‘important economic power for the coming Islamic state, God willing’. He additionally warns the ‘aggressive United States’ against burning any oil at the end of any war.

Religion features in the amplification of the status grievance and subsequent violence legitimisation. To amplify the grievance, he sanctifies Saudi Arabia with the phrases, ‘place of revelation’, ‘source of the Prophetic mission’, and ‘home of the Noble Ka’ba where Muslims direct their prayers’. Any action to protect these deeply sacred sites is given a pious justification. Indeed, bin Laden warns that ‘warding off an enemy who corrupts religion and the world is the top duty after faith’. Accordingly, the violence he legitimises is given religious legitimacy by describing it as a ‘jihad against the enemies of God’, against ‘your enemies the Israelis and Americans...to expel them in defeat and humiliation from the holy places of Islam'. The mention of God in this phrase (as above) is to use him as an arbiter of whether the jihad is morally just. As such, Laden declares many times that 'God is the source of all power' and that 'God says it is permissible to shed [the occupiers] blood and seize their property, and anyone who kills a person, the booty is his'.

In contrast to bin Laden's victim narrative, Bush creates a hero narrative for his ingroup to uphold the political ideals of freedom and justice against the threat of terrorism. Bush identifies his ingroup with the often-used phrase, 'my fellow Americans', the adjective 'fellow' links Bush to the geopolitical group of Americans. In relation to political ideals, his first sentence in response to the attacks declares 'our way of life, our very freedom came under attack in a series of deliberate and deadly terrorist acts'. He lionises his ingroup by claiming they were targeted because 'we're the brightest beacon for freedom and opportunity in the world'. He additionally invokes America's founding fathers as arbiters of these political ideals in the phrase, 'we are freedom's home and defender, and the commitment of our Fathers is now the calling of our time'. He establishes his ingroup's leadership role by claiming 'America will lead by defending liberty and justice'. And in a rallying cry to other nations, he declares, 'This is not, however, just America's fight. And what is at stake is not just America's freedom. This is the world's fight. This is civilization's fight. This is the fight of all who believe in progress and pluralism, tolerance, and freedom'.

Bush elevates his ingroup by characterising them with morally virtuous terms. In contrast to bin Laden's depiction of America as a violent country, Bush remarks how 'the world has seen that our fellow Americans are generous and kind, resourceful and brave'. Bush also upholds his ingroup's economic virtues in the phrase, 'America is successful because of the hard work, and creativity, and enterprise of our people'. There are also many examples of Bush extoling the virtues of first responders, the Armed Forces, political leaders and American citizens as prototypical exemplars of the virtues Bush bestows onto his ingroup. He uses these prototypes to elevate the greatness of his ingroup. Much like treating an outgroup as a homogenous entity is a form of othering, Bush also lionises America as a homogenous hero within his narrative.

Bush uses the political ideals he seeks to uphold to create a gradient between his ingroup and outgroup. Much like bin Laden's use of 'Jewish-crusade alliance', the primary abstract representation of Bush's outgroup is 'terrorist' to which he ascribes various groups across the text. To create a gradient he claims, '[terrorists] hate what they see right here in this chamber - a democratically elected government. '[terrorists] leaders are self-appointed. [terrorists] hate our freedoms - our freedom of religion, our freedom of speech, our freedom to vote and assemble and disagree with each other'. Bush is also careful to separate terrorists from the Muslim faith. For example, he contrasts terrorists against 'Islam is peace', whereby 'these terrorists don't represent peace. They represent evil and war'. The use of evil for othering is a common theme in other such clauses as 'And in the terrorists, evil has found a willing servant'. And in the statement 'Either you are with us, or you are with the terrorists', Bush divides international relations between his ingroup (us) and his outgroup (the terrorists).

The first mention of al Qaeda as a terrorist occurs across two sentences, 'Americans are asking: Who attacked our country? The evidence we have gathered all points to a collection of loosely affiliated terrorist organizations known as al Qaeda'. He then identifies bin Laden and creates an alliance of terrorist organisations in the subsequent sentence, 'This group and its leader - a person named Usama bin Laden - are linked to many other organizations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan'. As a pretext for the subsequent invasion of Iraq, Bush additionally invokes state sponsored terrorism to create a second collective outgroup term, 'axis of evil'. Over six sentences he identifies North Korea, Iran, and Iraq as 'regimes' who sponsor terrorism and threaten 'America or our friends and allies with weapons of mass destruction'. In the sentences, 'States like these, and their terrorist allies, constitute an axis of evil, arming to threaten the peace of the world.

In contrast to bin Laden's religious framing of Jihad, Bush uses political ideology to legitimise his 'War on Terror'. He invokes political ideals to frame the War on Terror as struggle between 'Freedom and fear, justice and cruelty'. Moreover, he regards his signing of the Patriot Act as an 'essential step in defeating terrorism, while protecting the constitutional rights of all Americans'. Of interest, most mentions of 'War on Terror' are also preceded by the pronoun, 'our', which confers a sense of ownership of the campaign to his ingroup. He also legitimises the War on Terror as a way to achieve Peace in the phrase, 'America and our friends and allies join with all those who want peace and security in the world, and we stand together to win the war against terrorism'. Across several clauses he additionally promotes the War on Terror as a 'great cause' and frames it positively by declaring, 'Our cause is necessary. Our cause is just'.

Drawing upon the underpinning hypothesis for the methodological framework, the social categories associated with ingroups and outgroups in elevation and othering clauses provides a basis for a measurement schema to score the Self-Other gradient. As Kahneman suggests, 'the brain responds quickly even to purely symbolic threats' and using 'emotionally loaded words quickly attract attention, and bad words (war, crime) attract attention faster than do happy words (peace, love)' (Kahneman, 2011, p. 301). Different terms for describing group attributes, therefore, suggest an elevation and othering hierarchy of semantic units. For elevation, familial concepts, such as 'brother' and 'sons', are likely to be the most significant, followed by concepts synonymous with friendship, community, and affiliation. Conversely, outgroup concepts, such as 'terrorist' or 'enemy' invoke different intensities of othering. Applying scores to these terms then enables a potential quantification of the Self-other gradient.

This application of the methodological framework to Bush's and bin Laden's declarations of war has revealed interesting similarities in how each orator legitimises violence. In the first instance, each orator uses abstract representations to define social categories in each narrative. They also use abstract representations of people to create a sense of the prototypical member of their ingroup and outgroup. Each orator uses these representations as characters in the narrative truth of each text. In relation to religion, they both sanctify assets of value to their ingroup to motivate any violent action to protect them. Each orator additionally uses God to confer a sense of righteousness on their legitimisation of violence. In relation to ideology, bin Laden rejects political ideology in favour of religious jurisprudence through Shariah law. Conversely, Bush uses political ideology as a set of ideals his ingroup must uphold to maintain world order. Moreover, both use economics to create a sense of status differences. Bin Laden's use of economic ideology is to create a status grievance that he amplifies with religion, while Bush uses 9/11 as an attack on America's economic status symbols. Despite being representing oppositional positions of the same conflict, each narrative follows a similar plot and similar uses of religion and ideology for elevation and othering to legitimise their violent campaigns.

## 3.4 Discussion

In response to the research question, this chapter has used cultural violence from Peace Research to rethink hate speech detection as hostile narrative analysis. The requirement for hostile narrative analysis develops upon the previous chapter, which found limitations in the computational methods that quantify meaning in social science applications. This chapter takes a more theoretical approach to explain how hate speech has unhelpfully become a polysemous term through a transition from critical legal studies into the social, political and computer sciences. Recalling Hall's theory of encoding and decoding, definitions of hate speech largely depend on the orator-audience relationship. In a dominant position between liberal progressives and audiences, hate speech connotes unacceptable language against minorities. In the oppositional position from liberal progressives to reactionary audience, hate speech connotes attempts to close down free speech. In the negotiated position, hate speech connotes any number of levels of acceptability. This polysemy, therefore, questions the utility of hate speech as a term for explanatory dialogues about detecting hateful speech.

Despite the polysemous meaning of hate speech, its progressive ordinary meaning does have applicability beyond race relations to detect persecutorial, hateful and degrading language.

Dynamics of racial inferiority and historical oppression can be generally applied to group identities such as gender, sexuality or specific dynamics of race relations. Nonetheless, any faithful application of Matsuda's conception of historical oppression should recognise any changes in the relative power dynamics of the target group. These power dynamics relate to Galtung's notion of structural violence that itself elicits disagreement about intention to create societal harms. Evidencing the changes in power dynamics is perhaps where the criticism and commentary about each type of hate speech emerges. And this thesis aims to overcome such politicised tensions to better understand how structural violence may be legitimised.

The polysemy of hate speech also means there is no generally accepted methodology for hate speech detection. Outsider jurisprudence is a valuable methodology for motivating change but does problematise hate speech as subjective experience from a victim's perspective. To problematise hate speech subjectively creates a vulnerability that can be exploited by reactionary commentary. Consequently, the methodology for hate speech detection is essentially unique to a particular algorithm; developers write the methodology in their coding schema and apply it during the annotation of training data. As shown in the previous chapter, the computational methods generally draw upon quantitative approaches from NLP technologies. Yet the methodologies cannot be assessed when the coding schemas are unpublished, as is the case for the sentiment analysis applications. Therefore, hate speech detection systems are largely quantitative methods without a generally accepted methodology.

In developing upon the technical findings in Chapter 2, the application of quantitative methods to detect such a polysemous concept gives only an illusory understanding of hate speech. Using such quantitative methods implies that a qualitative assessment of hatefulness is somehow quantifiable. If quantitative methods are about objectivity and identifying facts, this chapter has explained how the quantification of hatefulness is only relevant to a developer's specific definition of hate speech, annotation schema and training data. Consequently, the outputs of detection systems provide only an illusory understanding of hate speech that does not stand up to the rigour. As such, the proposal to rethink hate speech as hostile narrative analysis seeks to remedy problems with the polysemy of hate speech by using the more generally accepted theory of cultural violence to develop computational methods of improved explanatory rigour.

## 3.5 Conclusion

The methodology presented in this chapter returns hate speech to its original definition of criminal language against minority groups and replaces the progressive ordinary meaning of hate speech with hostile narrative. Firstly, Galtung's theories of violence reconceptualise the ordinary meaning of *hate speech* as *violence legitimisation* against a target groups. Violence itself might be acts of direct violence or processes of structural violence. Violence against a vulnerable group then becomes racism, sexism or homophobia depending on how the target identifies. Accordingly, as a complement to Matsuda (1989), the utterance of speech deemed unlawful is an act of hate speech, which becomes a subclass of direct violence. The instruments of that violence are the particular words, phrases or stories an orator uses for othering the target group. The logic of cultural violence is to identify how an orator may use aspects of each domain to legitimise direct or structural violence. The logic of hostile narrative analysis is to detect how each cultural domain features in elevation and othering narrative clauses. This methodology also provides the basis for qualitative methods that are further developed in the next chapter.

# Chapter 4 Applying Semantic Analysis to the Analysis of Hostile Narratives with hybrid NLP

| Objective | Description | Output |
|---|---|---|
| **Obj 1.** | Text Pre-processing | Labelled data |
| **Obj 2.** | Detect and classify named entities as either ingroup or outgroup | A list of named entities classified as either ingroup or outgroup |
| **Obj 3.** | Detect and classify ingroup elevation and outgroup othering phrases. | A list of elevation and othering statements linked to each entity |
| **Obj 4.** | Score the Self-other gradient. | Scores for each ingroup and outgroup relationship of a text |

Figure 29. The computational methods for analysing hostile narratives.

This chapter responds to the research question, 'how does augmenting quantitative NLP methods with qualitative approaches enable the meaningful analysis of hostile narratives?'. Thus far, this thesis has revealed problems with quantitative methods in NLP for social science applications like hate speech detection and sentiment analysis. Chapter 2 finds problems with processing natural language as unstructured data using quantitative methods. Chapter 3 responds with a methodology to begin the introduction of more qualitative approaches to hate speech detection. The chapter explains how the polysemy of hate speech in its ordinary meaning means there is no generally accepted methodology for determining whether a text is hateful. The second part responds by rethinking hate speech detection as hostile narrative analysis using a methodological framework derived from Galtung's theories of violence from peace research. The qualitative element of this framework applies narrative analysis to analyse the linguistic elements of a hostile narrative as a story. This chapter continues with developing qualitative computational methods by drawing upon semantic analysis for a hybrid-NLP approach to analysing hostile narratives.

This chapter shows how augmenting text classification with the qualitative method of semantic analysis enables the automatic generation of rationales to explain why a narrative is hostile. The chapter begins by showing how semantics enables hostile narrative analysis by detecting elevation and othering language clauses. This detection of language clauses treats text as structured data, which contrasts with the established methods reviewed in Chapter 2 that treat text as unstructured data. The chapter then continues by introducing a custom dependency parser to apply semantic analysis to detect these clauses. This parser uses grammar patterns to model language clauses as the basis for pattern-based NLP. The findings of this second section then inform the final section that covers the computational methods for each objective of the method in Figure 29. In contrast to the algorithms reviewed in Chapter 2, the final section then shows how applying semantic analysis through pattern-based NLP more accurately represents an orator's intended meaning. The result is a hybrid approach to NLP for analysing hostile narratives.

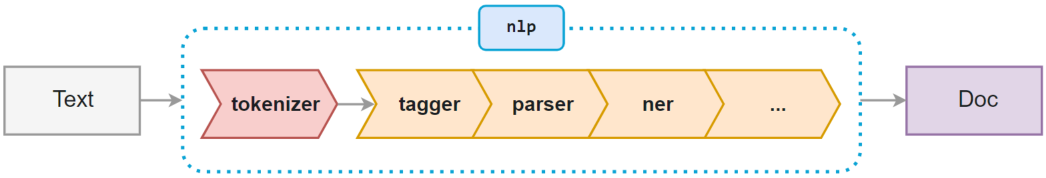



Figure 30. The spacy pipeline.

The computational methods of this chapter use the spaCy python library to develop the pattern-based approach to NLP. As an open-source python library for NLP, spaCy is based on a pipeline architecture shown in Figure 30. The purpose of each pipeline component is to create a machine-readable Doc object from text inputs. The tokeniser segments a text into tokens; the tagger assigns parts of speech (POS) tags to each token; the parser assigns dependency labels; named entity recognition (ner) detects and labels named entities. As will be explained, each of these standard components uses text classification for labelling a word's lexical and grammatical properties. This application of text classification is consistent with Chapter 2 since this thesis regards the labelling of grammatical and lexical attributes as a functional task; within a reasonable tolerance, these attributes have denotative meaning. In addition to these standard components, the second section presents a custom pattern-based parser to process text by language clauses. The resultant hybrid approach to NLP combines quantitative methods for word labelling and patterns for parsing language clauses.

The point of this chapter is not to generate a production-ready pipeline but to motivate the development of such a pipeline in further research. As such, the chapter begins with linguistic theory and finishes with applying that theory using the spaCy library. The chapter applies the theory by detecting language clauses for each objective of the method in Figure 29. This chapter focuses on developing the components of a production system for each objective, but these components alone are insufficient for a production system. The development of these components revealed the requirement for incorporating knowledge graphs into a production pipeline. The next chapter proposes this requirement as further work. As the further work explains, applying knowledge graphs means treating text as structured data, which is contrary to the algorithms reviewed in Chapter 2. For now, this chapter shows the how treating text as structured data using pattern-based NLP enables explanatory dialogues about hostile narratives by generating a rationale to explain why a particular text is hostile.

## 4.1 How Does Semantic Analysis Explain Violence Legitimisation?

This first section shows how the application of semantic analysis assists with detecting violence legitimisation in a hostile narrative. This section begins by connecting Galtung's view on the role of language in violence legitimisation with semantic analysis. This connection will show how semantic analysis assists with detecting violence legitimisation by detecting elevation and othering language clauses – objective 2 of the method. Based on this explanation, a technical review of NLP methods for semantically analysing text follows. Following this thesis's reoccurring theme, the section finds a disconnection between NLP literature and linguistic theory for semantic analysis. While semantic analysis applications in NLP generally treat text as unstructured data, established linguistic theory treats text as structured data. The subsequent section, therefore, connects theory and practice using pattern-based NLP to process natural language as structured data.

### 4.1.1 What is Semantic Analysis?

While not making explicit reference, Galtung and Njshimura (1983) draw upon semantic analysis to introduce the role of language in violence legitimisation. Galtung and Njshimura (1983) provide a comparative analysis of Indo-European, Japanese and Chinese languages to reveal what they describe as the 'social cosmology' of each culture. Social cosmology refers to the 'deep structure' or 'deep culture' of latent or implicit assumptions within a language community. Concerning group belonging, Galtung and Njshimura (1983) explains that 'language becomes a symbol of social attribution, of belongingness as well as relationship in a more vertical sense' (Galtung & Njshimura, 1983, p. 20). Martin's (2018) perspective from evolutionary psychology supports this view whereby language and other expressions of art are identity markers for group identity (Martin, 2018, p. 103).

'Semantics is the branch of linguistics devoted to the investigation of linguistic meaning' (Chierchia & McConnell-Ginet, 2000a, p. 1). As explained by both Matthews (2003) and Burton-Roberts (1997), linguists generally treat 'a language' as a system of rules combining words or set of grammatically correct sentences (Burton-Roberts, 2016, pp. 284–285; Matthews, 2003, pp. 83–88). Of the infinite combinations of words, 'grammar' comprises the rules constraining how they are linked to convey meaning. Thereafter, sentences are structured sequences of words deemed acceptable to a particular language. Semantic analysis is then the analysis of text as a linguistic system to reveal insights about the person or people using a particular language. Each objective for the method in Figure 29 is the insight hostile narrative analysis seeks to reveal about assertions of hostility in a text.

For the semantic analysis of syntactic structures and sentences, Noam Chomsky's 1957 'Syntactic Structures' has become a foundation of modern linguistics. Chierchia and McConnell-Ginet (2000) describe Chomsky's grammatical framework as a 'set of abstract devices, rule systems and principles that serve to characterise formally various properties of a well-formed sentence of that language' (Chierchia & McConnell-Ginet, 2000b, p. 1). In his thesis, Chomsky creates a framework for parsing text using a 'form of grammar associated with the theory of linguistic structure based upon constituent analysis' (Chomsky, 1957, p. 29). Chomsky's framework became part of a broader effort to develop a theory of generative grammar, which is a set of rules to generate (or describe) all possible sentences of a language.

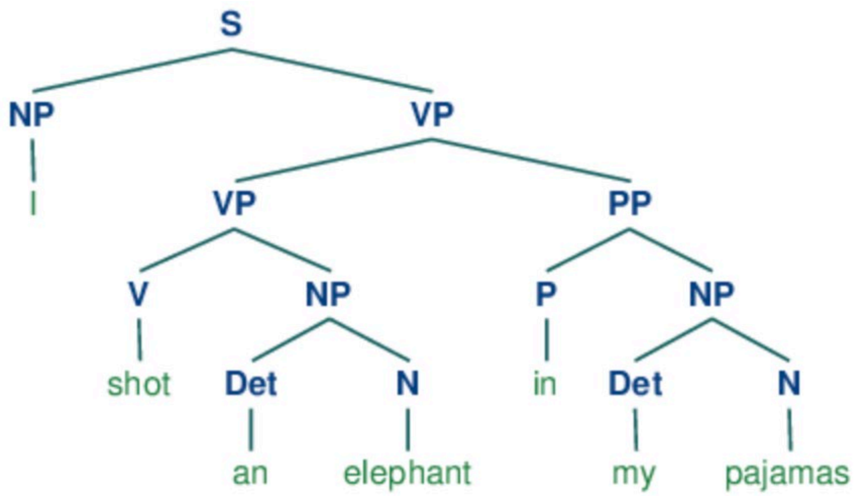



Figure 31. A tree diagram of phrase structure grammar.

Chomsky's approach to parsing text has become known as phrase structure grammar, with phrases being the constituent components of a sentence. Figure 31 shows the tree diagram of a sentence, 'I shot an elephant in my pajamas [sic]' (Bird et al., 2009, p. 294). In this sentence (S), the first noun phrase (NP) is the pronoun 'I', and the verb phrase (VP) is 'shot an elephant'. The verb phrase also has a prepositional phrase (PP), 'in my pyjamas'. The head of the first noun phrase is 'I', and the head of the first verb phrase is 'shot'. This formalism of phrase structure grammar is known as Chomsky Normal Form (CNF); other subsequent formalisms include Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994) and Lexical-Functional Grammar (LFG) (Bresnan, 1982).

In contrast to parsing a text by phrases, an alternative approach known as dependency grammar is about parsing a text by language clauses. As Nivre observes, 'the starting point of the modern theoretical tradition of dependency grammar is usually taken to be the work of the French linguist Lucien Tesnière', published posthumously in 'Elements of Syntactic Structure' (Nivre, 2010). In contrast to parsing a sentence by phrases, Tesnière (1959) uses individual words as the nodes of a dependency grammar framework. Tesnière placed the verb as the governing node of a sentence upon which other words are dependent in a subject->predicate->object relationship (Tesnière, 1959, p. 98). The spaCy library follows this dependency grammar model.

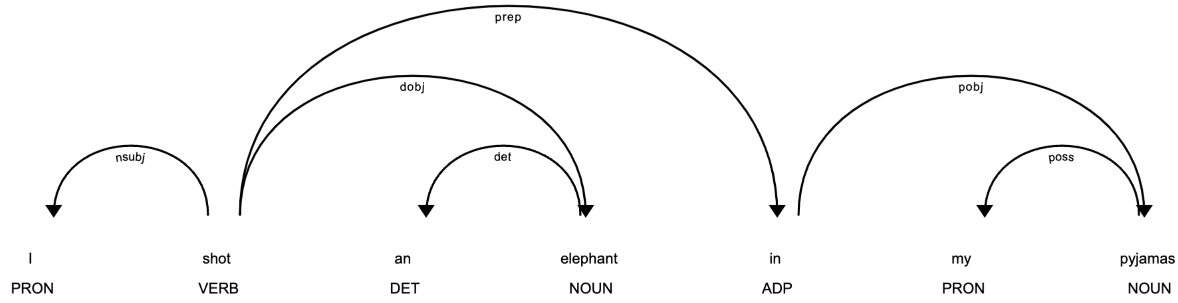



Figure 32. A dependency parse using dependency grammar.

In contrast to a tree diagram, dependency grammar represents a sentence as a directed graph, more commonly known as a dependency parse, as shown in Figure 32. Dependency parsing involves labelling the relationships between words, thereby assigning grammatical structure to a sentence (Patel & Arasanipalai, 2021, p. 14). Figure 32 uses the same sentence as in Figure 31 to show how words become a graph's nodes, and the dependencies between each word are the edges[59]. In this parse, the verb (VERB) 'shot' is the predicate and sentence root, the pronoun (PRON) 'I' is the nominal subject (nubj), and the noun (NOUN) 'elephant' is the direct object (dobj) or the root. 'Elephant' is then the head of the noun phrase 'an elephant', where 'an' is a determiner (DET) of 'elephant'. The adposition (ADP) 'in' begins the preposition (prep) that modifies the root. The preposition comprises another noun phrase, 'my pyjamas', that serves as a prepositional object (pobj) to the root, where 'my' is a possession modifier (poss).

This thesis chooses dependency grammar to model language due to its consistency with other relevant theories from narrative analysis. Dependency grammar is consistent with Todorov and Weinstein (1969), who model a language clause's structure as a subject and object linked by a predicate term (Todorov & Weinstein, 1969). Labov and Waletzky (1997) also use the same structure and refer to the predicate as a clause's head (Labov & Waletzky, 1997, p. 22). As is explained in further work in the next chapter, there is also a degree of consistency between language clauses and structures of knowledge graphs. Dependency grammar, therefore, becomes the basis for pattern-based NLP and processing text by language clauses.

To understand how language clauses feature in elevation and othering, consider the following sentence from the Race and People chapter of Mein Kampf.

[59] Generated using the spaCy python library for NLP.

> *The Aryan himself was probably at first a nomad and became a settler in the course of ages.*

- Clause 1: was(Aryan, Nomad)
- Clause 2: became(Aryan, settler in the course of ages)

In isolation, this two-clause sentence contains no apparent features of hostility. It asserts that Hitler's ingroup of Aryans were firstly Nomads and then settlers over time. In isolation, this sentence is benign; its interaction with a subsequent sentence, however, promotes antisemitism by dehumanising Jews as parasites in comparison to Hitler's Aryan ingroup.

> *The Jew has never been a nomad, but always a parasite, battening on the substance of others.*

- Clause 3: never_been(Jew, Nomad)
- Clause 4: been(Jew, parasite)
- Clause 5: batten(Jew, substance of others)

Through the linking term 'nomad', clause 3 disassociates Jewish people from Hitler's ingroup, thereby establishing Jews as an outgroup. Clause 2 then dehumanises Jews to the status of a parasite. This dehumanisation inappropriately makes Jews akin to fungi, bacteria, and viruses in the narrative truth of those who subscribe to *Mein Kampf*. Where 'battening' means 'to grow prosperous, especially at the expense of another'[60], clause three extends the dehumanisation in clause two by implying Jewish existence is at the expense of others. As is explained in the subsequent sections of this chapter, this clausal analysis forms the basis for each objective of the hostile narrative analysis method in Figure 29. Clauses 1 and 3 relate to objective 2, while clauses one, four and five relate to objective 3. The assertions of these clauses contrast the ingroup and outgroup to create the Self-other gradient between each.

Modelling language clauses, nevertheless, is limited by what Manning and Schütze (1999) explain as the problem of ambiguity in language (Manning & Schütze, 1999, p. 17), which Bird *et al*. (2009) refer to as 'Ubiquitous Ambiguity' (Bird et al., 2009, p. 293). This thesis explains this ambiguity using both denotative meaning and grammatical relations. Accordingly, the meaning of the sentence 'I shot an elephant in my pyjamas' is based on what the verb 'shot' denotes. If it denotes discharging a firearm, the sentence implies 'I killed an elephant'. Conversely, if it denotes the action of taking a photograph, the sentence implies, 'I took a photograph of an elephant'. The

[60] Marriam Webster (n.d.) Batten (verb), retrieved on 17th Feb 2023

meaning of this sentence also depends upon the preposition's relation to other words. Where the preposition modifies 'shot', the sentence implies 'I shot the elephant while I was wearing my pyjamas'. As Bird *et al.* (2009) explain, however, this preposition can also modify 'elephant', which conjures the image of an elephant wearing a set of pyjamas while being shot. This ambiguity is a problem for human verification within an explanatory dialogue. Having introduced how semantic analysis features in understanding violence legitimisation, the section now continues with a short review of how it features in NLP literature.

### 4.1.2 How Does Semantic Analysis Feature in NLP Literature?

While semantic analysis from linguistics treats text as structured data, semantic analysis in natural language processing tends to treat text as unstructured data. Any treatment of text as structured data in NLP tends to be theoretical. Manning and Schütze (1999) present several statistical methods for the 'probabilistic parsing' of a sentence, whether through phrase structure or dependency grammar (Manning & Schütze, 1999, p. 456). Bird *et al.* (2009), a primary source for the nltk python toolkit for NLP, provide the dependency parses in Figure 31 and Figure 32. Another textbook from Bengfort *et al*. (2018) defines semantics as decoding meaning from natural language using the subject->predicate->model as a template for constructing ontologies using such resources as Wikipedia or DBPedia. The spaCy library used in this chapter also provides a dependency matcher to parse a text using language patterns.

Beyond theory, the practical applications of semantic analysis tend to treat text as unstructured data by using statistical methods to infer meaning. Chapter 2 reviewed word embeddings to explain how they can represent a word's denotative meaning but struggle with connotative meaning. In the review of text classification for social science applications, Chapter 2 also explains how the processing of text by word co-occurrence skews the orator's intended message. More broadly, Landauer and Dumais (1997) introduced Latent semantic analysis (LSA), which has become a popular method in addition to text classification. 'LSA is a fully automatic mathematical and statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse' (Landauer et al., 1998, p. 263). LSA, however, still relies upon word co-occurrence by applying inductive reasoning to 'construe the semantic similarity between two words in terms of semantic space: the smaller the distance, the greater the similarity' (Landauer & Dumais, 1997, p. 215). 'Semantic space' refers to the number of words between two words of interest and applies the distributional hypothesis that Chapter 2 critically analysed.

Nevertheless, there is emerging literature about applying semantic web technologies to NLP that treats text as structured data. As such, the application of ontologies and knowledge graphs features in computational journalism to augment knowledge in text (Castells et al., 2004; Fernández et al., 2010; Rospocher et al., 2016; Rudnik et al., 2019; Vossen et al., 2016). To semantically parse a text using grammatical relations, GraphBrain from Menezes and Roth (2019) aims to facilitate automated meaning extraction and text understanding, as well as the exploration and inference of knowledge. To meet this aim, Graphbrain uses a 'Semantic Hypergraph' based on the subject->predicate->object model of a language clause[61]. This use of graphs to extract knowledge seems to be an emerging approach which this thesis seeks to develop upon in further work.

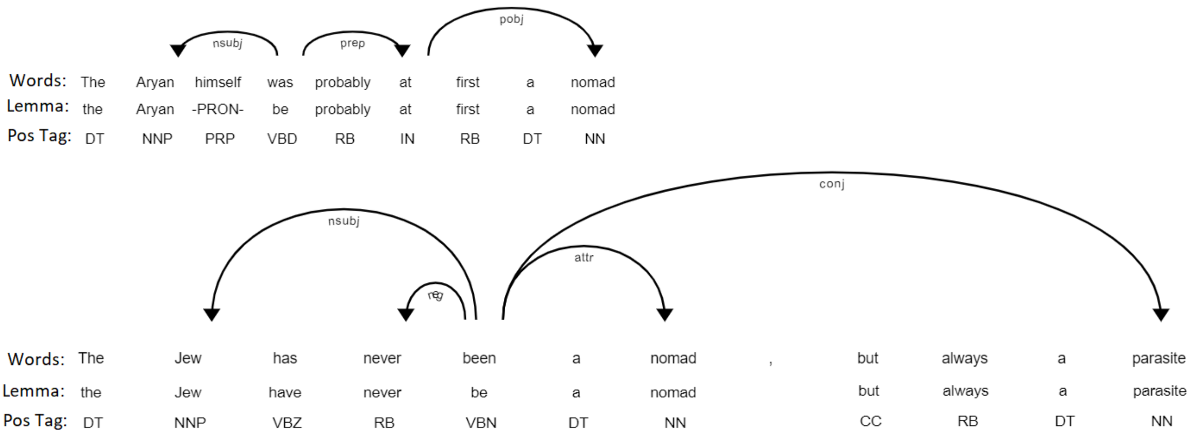



Figure 33. Dependency parse of the *Mein Kampf* statements.

Figure 33 shows the dependency parses for the *Mein Kampf* clauses in the previous section using the spaCy library. As a functional task, spaCy's algorithm uses statistical models and word embeddings to predict the grammatical attributes of each word. There are 23 models for different world languages[62]; the English model used in this chapter uses the OntoNotes 5, a large, annotated corpus of 2.9 million words comprising telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs, religious texts (Weischedel et al., 2012). SpaCy augments these texts with WordNet, a lexical database of English words. Since most of OntoNotes' texts are from news and broadcast news, this model strongly represents the geopolitical vocabulary of the hostile narrative corpus. Other genres and dialects require more bespoke models.

The parts-of-speech (POS) tags use the Universal Dependencies (UD)[63] framework to label the grammatical function of each word. The edge labels between each word use the ClearNLP schema

---

[61] GraphBrain (2020) The Semantic Hypergraph, retrieved on 17th Feb 2023

[62] Explosion.ai (2022) Language Support, retrieved on 17th Feb 2023

[63] Universal Dependencies (2021), retrieved on 17th Feb 2023

developed by Emory NLP[64] to show the syntactic relations. As the figure shows, the subject (Aryan, Jew) and object (Nomad, parasite) have a POS label of either noun (NN) or proper noun (NNP). For grammatical relations, the subject has a nominal subject (nsubj) label, and the object is either a prepositional object (pobj) or part of a conjunction (conj). In each case, the predicate (was, been) is a verb of a different tense (VDB, VBN), and a preposition (prep) follows a verb. Additionally, through lemmatisation, the verbs 'was' and 'been' resolve to their head of 'be'. Also, note the negation (neg) label, which enables a detection system to account for negated relations that are not detected when processing text by word co-occurrence.

In narrative analysis theory, the verb, or clause head defines the meaning of a clause. While the UD framework is a resource for labelling verbs, VerbNet from Schuler (2006) provides a lexical resource for interpreting their semantic meaning (Schuler, 2006). To explain, consider the clause be(Jew, parasite). The predicate 'be' is in the 'representation' class on VerbNet – along with 'denote', 'mean', 'represent', 'signify' and 'symbolise' – with the semantic role of 'signify'[65]. 'Be' could be replaced with another word of the representation class, and the clause would broadly have the same meaning. In this respect, the clause proposes that the subject 'signifies' the object. VerbNet, therefore, has the potential to infer a clause's meaning by what its head signifies.

This section has explained how semantic analysis applies to violence legitimisation through the analysis of language clauses to detect elevation and othering. Following a re-occurring theme of this thesis, the section has also explained an inconsistency between the linguistic and computational theories of semantic analysis. Where linguistic theory tends to derive meaning from linguistic structures, computational methods tend to ignore such structures to quantitatively process text as unstructured data. The section then continued by showing how parsing a text as structured data using language clauses can contribute to each objective of the hostile narrative method. The chapter continues by presenting a spaCy pipeline component for applying semantic analysis to text as structured data.

## 4.2 How can Pattern-based NLP Apply Semantic Analysis?

This section applies pattern-based NLP to semantic analysis using custom dependency parser. This customer dependency parser is for the computational methods of each objective of this hostile narrative method in Figure 29. To be consistent with the linguistic theory of semantic analysis, this

---

[64] EmoryNLP (2022) NLP4J, retrieved on 17th Feb 2023

[65] VerbNet (2022) representation-110.1, retrieved on 17th Feb 2023

parser processes text by language clauses. As such, the parser models language clauses using a set of patterns based on Grammar Patterns from the Collins Dictionary[66]. This section begins by introducing grammar patterns and how they apply to the custom dependency parser. The section then continues by showing how the custom dependency parser applies to a series of sentences from Bush's declarations of war. In doing so, the section reveals the broader text processing requirements for the other spaCy components to enable pattern-based NLP. These requirements augment quantitative methods of text classification with pattern-based approaches to develop hybrid NLP.

### 4.2.1 How Do Grammar Patterns Apply to Language Clauses for Pattern-based NLP?

Grammar patterns enable the modelling of language clauses for pattern-based NLP. Pattern grammar represents the theory behind investigating the grammatical structures of language, while grammar patterns describe those structures. Consistent with dependency grammar presented in the previous section, pattern grammar views language as a system of patterns or templates that can be combined in different ways to create meaning. Patterns are observed by regularly occurring linguistic items. 'When those items are identified as grammatical items, such as prepositions or clauses, they are described as the elements of a pattern' (Hunston, 2019, p. 1). Pattern grammar originated with the work of Hornby (1954) (Hornby, 1954). Francis, Hunston and Manning have since developed the field using corpus linguistics and recorded a range of grammar patterns in two reference works, 'Grammar Patterns 1: Verbs' and 'Grammar Patterns 2: Nouns and Adjectives' (Hunston et al., 1996, 1998).

Developed by Collins Dictionary, grammar patterns refer to the grammatical arrangement of words in phrases, clauses, and sentences. The Collins website provides a unique listing of all the English grammar patterns and all the words regularly used with a given pattern. Using corpus research, Collins' Lexicographers at COBUILD and experts at the University of Birmingham generated these patterns. The main types of patterns considered in this chapter are the 'simple pattern', the 'simple pattern with prepositions and adverbs' and 'complex patterns'.

*Simple:* {*we*}$_{subject}$ {*condemn*}$_{predicate}$ {*the Taliban regime*}$_{direct\ object}$.

The 'simple grammar pattern' type follows the subject-verb-object pattern. As shown in the clause, 'we condemn the Taliban regime', the subject of the sentence is the person or thing

---

[66] Collins (n.d) Grammar Patterns, retrieved on 17th Feb 2023

performing the action, the verb is the action being performed, and the object is the person or thing receiving the action.

> *Simple Preposition*: *And* {*on*} preposition {*behalf of the American people*} *prepositional object*, {*I*} *subject* {*thank*} *predicate* {*the world*} *direct object* {for} preposition {*its outpouring of support*} *prepositional object*

A ‘simple pattern with prepositions and adverbs’ is simple pattern with a prepositional phrase or an adverb group linked to the verb. In the example clauses here, ‘I thank the world for its outpouring of support’ is the simple pattern, and the preposition is ‘on behalf of the American people’. In effect, the simple pattern forms the base from more complex patterns are extended.

> *Complex Pattern:* {*They*} *subject* {*are*} *predicate* {*some of the murderers indicted for bombing American embassies in Tanzania and Kenya*}, *and* {*responsible for bombing the USS Cole*}

In complex patterns, the verb is followed by two elements. The verb is followed by two noun phrases, or the verb is followed by a noun phrase and an adjective phrase. In the example shown above, the verb, ‘are’, is followed by a conjunction of a noun phrase, ‘some of the murderers indicted for bombing American embassies in Tanzania and Kenya’ and the phrase, ‘responsible for bombing the USS Cole’.

As can be seen from these examples, the more complex patterns add various elements, whether prepositions or adverbs, to the base pattern. The patterns presented in this chapter similarly uses the simple pattern as the base, which is then extended for the more complex patterns. For pattern-based NLP, therefore, these patterns provide the grammatical relations that govern linguistic systems and help to interpret meaning.

There is a need to distinguish between the terms ‘lexical’ and ‘grammatical’ to understand the development of grammar patterns. Linguistics uses ‘lexical’ to describe the linguistic attributes of individual words and phrases of a language, as opposed to the grammar or structure of the language. The lexical properties of individual words in the custom dependency parser draw upon the UD framework introduced above. Grammatical refers to the relationships between words that conform to the rules of a particular language. The custom dependency parser uses grammar patterns to identify those relationships in language clauses. The interpretation of these clauses relates to how an orator links different words to convey meaning.

The interpretation of clauses also requires an understanding of the distinction between syntactic rules and patterns. Pattern grammar theory is usefully distinct from syntactic rules since patterns emphasise naturally occurring language rather than following an imposed order (Hunston & Francis, 2000, p. 15). The standard explanation for this distinction draws upon idioms, which typically present a figurative, non-literal meaning attached to a phrase. Idioms strongly feature in the hostile narrative corpus. As previously explained, both Bush and bin Laden use familial terms to construct a narrative truth of international relations. As such, the applications of rules would treat the United Kingdom and the United States as *literal* friends. In contrast, Pattern Grammar is more permissive by allowing more *figurative* interpretations of clauses. For Pattern Grammar, the meaning of friendship for a country differs from the meaning of friendship between people. As such, pattern grammar permits the analysis of narrative rather than literal truth.

The custom dependency matcher applies grammar patterns to using spaCy's DependencyMatcher pipeline component. Instead of parsing a text by word co-occurrence, the DependencyMatcher enables the detection of clauses by grammatical relations. The various pipeline components in Figure 30 tokenise and label words according to their lexical properties. The patterns also use Semregex[67] operators, developed by the Standford University NLP group, to match the relations between words in a dependency graph. A word's lexical attributes describe a linguistic system's nodes, while Semregex describes how these nodes connect.

| Dictionary | Node Name | Description |
|---|---|---|
| Anchor | RIGHT_ID | The name of the pattern's anchor node |
| | RIGHT_ATTRS | Token attributes to match the anchor node |
| | | |
| Node | LEFT_ID | The name of the left-hand node in the relation, which has been defined in the RIGHT_ID |
| | REL_OP | A semregex operator that describes how the two nodes are related |
| | RIGHT_ID | A unique name for the right-hand node relative to the anchor node |
| | RIGHT_ATTRS | The token attributes to match for the right-hand node |

Figure 34. SpaCy's structure for its dependency parser patterns.

[67] Stanford (n.d.) SemregexPattern, retrieved on 7th Jan 2023

Figure 34 shows the basic structure of a spaCy pattern for its dependency matcher pipeline component. spaCy describes this structure as ‘a list of dictionaries, with each dictionary describing a token to match and its relation to an existing token in the pattern’[68]. The anchor dictionary is the first in the list, followed by any number of connecting node dictionaries. For the patterns developed here, the anchor node is defined by verbs and is named ‘predicate’. The LEFT_ID for subsequent patterns is then also ‘predicate’ while the RIGHT_ID is either ‘subject’ or ‘object’ for simple clauses. The structure links the RIGHT_ID and LEFT_ID of the pattern through the common node name. Subsequent patterns may also contain verb prepositions; in such cases, the RIGHT_ID node is named as ‘preposition’ followed by a ‘prepositional object’. The primary semregex operator is ‘>’, which sets the direction from the anchor to the next node in the list regardless of whether the word appears to the left or right of the anchor.

This research has begun to create spaCy dependency patterns from grammar patterns to enable the proposed approach of parsing text by its grammatical relations. So far, the research has created 13 patterns, which are available online[69]. In creating these patterns, a decision was made to split a clause into two elements for the patterns. The first element is the base, which identifies a verb and its corresponding subject. The second element comprise the various objects linked to a verb, whether a direct object, preposition or adverb. The matcher then identifies these patterns in text and reconstructs the clause by linking the various objects to the predicate in the base pattern. This approach of splitting the clause in a pattern and then reconstructing it in the matcher best enables the parsing of the complex grammar patterns.

---

[68] Exploision.ai (n.d.) DependencyMatcher, retrieved on 7th Jan 2023

[69] Anning (2023) Lexical Objects, retrieved on 17th Feb 2023

```
patterns = [
  {
    "pattern_name": "BaseActive",
    "pattern":
      [
        {
          # the anchor of the pattern is the verb of a clause
          "RIGHT_ID": "PREDICATE",
          "RIGHT_ATTRS": VERB
        },
        {
          "LEFT_ID": "PREDICATE",
          "REL_OP": ">",
          "RIGHT_ID": "SUBJECT",
          "RIGHT_ATTRS": _subject
        },
      ],
  },
  {
    "pattern_name": "SimpleDirectObject",
    "pattern":
      [
        {
          # the anchor of the pattern is the verb of a clause
          "RIGHT_ID": "PREDICATE",
          "RIGHT_ATTRS": VERB
        },
        {
          "LEFT_ID": "PREDICATE",
          "REL_OP": ">",
          "RIGHT_ID": "DIRECTOBJECT",
          "RIGHT_ATTRS": _direct_object
        },
      ],
  }
]
```

Figure 35. Patterns for the primary language clause.

Accordingly, Figure 35 shows the two patterns developed by this research for the simple grammar pattern. The explanation of this pattern links to the pattern structure shown in Figure 34. The anchor element of the pattern is the verb of a clause. ‘RIGHT_ID’ gives a name to the anchor and ‘RIGHT_ATTRS’ points to the POS tag. In this case both patterns name the anchor as ‘PREDICATE’, while the variable ‘VERB’ contains a python dictionary for the verb POS tag. The predicate of each pattern refers to the same verb thereby enabling the reconstruction of the clause from these two patterns in the matcher object.

Referring back to Figure 34 again, the node element of these patterns contains the object linked to the anchor. For both patterns, 'LEFT_ID' refers to the same 'RIGHT_ID' given to the anchor, in this case, 'PREDICATE'. 'REL_OP' contains the Semregex operator that signals the direction from 'LEFT_ID' to 'RIGHT_ID'. In this case, the direction is from the clause's predicate to either the subject or object. The BaseActive pattern then points to the _subject variable, which is a dictionary containing the dependency tag, 'nsubj' for a clause's subject. This variable is given the name 'SUBJECT' by the corresponding 'RIGHT_ID' dictionary. The SimpleDirectObject pattern points from the predicate to the object of a clause. In this case, 'RIGHT_ID' names the object, and the _direct_object variable contain the 'dobj' dependency label for a direct object. For a more complex clause, the preposition pattern and others are linked to this same predicate. As previously stated, after identifying these patterns in text, the matcher reconstructs the full clause by linking the patterns that point to the same verb. The following section shows how these patterns apply to Bush's declaration of war.

### 4.2.2 How do Grammar Patterns Apply to Bush's Declaration of War?

This subsection applies the custom dependency parser to selected sentences from Bush's declaration of war to show how grammar patterns apply to each objective in Figure 29. This subsection summarises an open-sourced notebook[70] that shows the code for applying the custom parser. The selected sentences each show how Bush sought to identify and other his outgroup. Each part of this section shows the sentence text, and the DataFrame output is a screenshot from the notebook. The accompanying explanation compares the parser's output with an assumed human interpretation. The explanations show how parsing text as structured data using these clauses addresses the problems with processing text as unstructured data identified in Chapter 2. The findings of this subsection then inform the processing requirements for the proposed pipeline in the subsequent section and further work in the next chapter.

*Sentence 1: On September the 11th, enemies of freedom committed an act of war against our country.*

| | RULE | PREDICATE | SUBJECT | DIRECTOBJECT | PREPOSITION | PREPOSITIONALOBJECT |
|---|---|---|---|---|---|---|
| 0 | SimpleDirectObject | committed | enemies of freedom | an act of war against our country | | |
| 1 | SimpleNounPreposition | committed | enemies of freedom | | On | September the 11th |

Figure 36.The *Simple* grammar pattern.

The first sentence, shown in

[70] Anning (2023) Developing Grammar Patterns, retrieved on 17th Feb 2023

Figure 38, is the narrative clause in which Bush asserts that enemies of freedom attacked the United States of America. The relevant grammar pattern to this sentence is a complex preposition, where the verb is followed by a noun phrase and then by a prepositional phrase or adverb. However, this pattern's constituent elements are a SimpleDirectObject and SimpleNounPreposition. The direct object is 'act of war against our country', and the preposition is 'September the 11th'. The subject then becomes the agent of this clause, while the direct object and preposition become attributes of the verb.

> *Sentence 2: The evidence we have gathered all points to a collection of loosely affiliated terrorist organisations known as al Qaeda.*

| | RULE | PREDICATE | SUBJECT | PREPOSITION | PREPOSITIONALOBJECT |
|---|---|---|---|---|---|
| **0** | SimpleNounPreposition | points | The evidence we have gathered | to | a collection of loosely affiliated terrorist organisations known as al Qaeda |

Figure 37. The *SimpleNounPreposition* grammar pattern.

This second sentence is the first to attribute al Qaeda to the act of war Bush mentions in the first sentence. As such, this sentence applies to objective 2 of the method as an identifier of Bush's outgroup. As seen from the output, the parser applies the SimpleNounPreposition pattern, albeit the constituent noun phrases are complex. Bush identifies his outgroup in the noun phrase, 'a collection of loosely affiliated terrorist organisations known as al Qaeda'. A human would resolve this noun phrase as a group with the attributes of terrorist and name al Qaeda. Interestingly, the attribution of this clause relates to the previous sentence in which Bush claims, 'Americans are asking: Who attacked our country?'. The attribution then requires the resolution of 'al Qaeda' to the pronoun 'who' of this question. The message Bush communicates, therefore, requires an understanding of how these sentences interact.

> *Sentence 3: This group and its leader, named Usama bin Laden, are linked to many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan.*

| | RULE | PREDICATE | SUBJECT | PREPOSITION | PREPOSITIONALOBJECT |
|---|---|---|---|---|---|
| **0** | SimpleNounPreposition | linked | This group | to | many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan |
| **1** | SimpleNounPreposition | linked | its leader -- a person named Usama bin Laden -- | to | many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan |

Figure 38. The *SimpleNounPreposition* grammar pattern

Sentence 3 is the first to show how the parser resolves conjunctions. The sentence further identifies Bush's outgroups by linking two other organisations to his already-named outgroup of al Qaeda. The sentence makes this link through two conjunctions where the head refers to two different entities, 'this group' and 'its leader – a person named Usama bin Laden'. The sentence links each entity to 'many other organisations in different countries, including the Egyptian Islamic Jihad and Islamic Movement of Uzbekistan'. As explained in Chapter 2, standard pre-processing practices would not detect these two conjunctions. Processing this sentence using language clauses, therefore, begins to represent the message Bush sought to communicate more accurately than the standard practices reviewed in Chapter 2.

*Clause 1: {[al Qaeda]}* $_{\textit{subject}}$ *{is linked to}* $_{\textit{predicate}}$ *{many other organisations in different countries}* $_{\textit{object.}}$

*Clause 2: {[al Qaeda's] leader}* $_{\textit{subject}}$ *{--}* $_{\textit{predicate}}$ *{a person named Usama bin Laden}* $_{\textit{object}}$.

*Clause 3: {[Usama bin Laden]}* $_{\textit{subject}}$ *{is linked to}* $_{\textit{predicate}}$ *{many other organisations in different countries}* $_{\textit{object}}$.

*Clause 4: {many other organisations in different countries}* $_{\textit{subject}}$, *{including}* $_{\textit{predicate}}$ *{the Egyptian Islamic Jihad}* $_{\textit{object}}$ *and {the Islamic Movement of Uzbekistan}* $_{\textit{object.}}$

Figure 39. Complex clauses

While the patterns developed so far focus on verbs, the reality of this sentence requires the additional development of patterns for noun phrases and the resolution of named entities to abstract representations. Figure 39 shows how a human would subconsciously resolve the four clauses of this sentence. In context, they would resolve the named entity of 'al Qaeda' to the noun phrase 'this group' and the pronoun 'its' for clauses one and two. They would then resolve the named entity 'Usama bin Laden' to the noun phrase, '[al Qaeda's] leader'. Clause four is a hyponymic noun phrase; hyponymic relations exist when a hypernym term classifies a series of hyponyms. The clause uses 'including' to classify the two named organisations as hyponyms of the hypernym, 'many organisations in different countries'. As such, development of noun phrase grammar patterns is also required.

*Sentence 4: They are some of the murderers indicted for bombing American embassies in Tanzania and Kenya, and responsible for bombing the USS Cole*

| | RULE | PREDICATE | SUBJECT | ATTRIBUTE |
|---|---|---|---|---|
| 0 | SimpleAttribute | are | They | some of the murderers indicted for bombing American embassies in Tanzania |
| 1 | SimpleAttribute | are | They | responsible for bombing the USS Cole |

Figure 40. The *SimpleAttribute* Pattern

Sentence 4 applies to objective 2 of the hostile narrative method as an othering clause for Bush's outgroup. As with sentence three above, understanding this sentence requires the resolution of 'they' to the appropriate named entity. In context, a human would resolve the pronoun 'they' to terrorists, which in turn refers to 'al Qaeda', 'the Taliban', 'the Egyptian Islamic Jihad' and the 'Islamic Movement of Uzbekistan'. While the parser resolves the two span conjunctions, the noun phrase conjunction, 'some of the murderers indicted for bombing American embassies in Tanzania and Kenya' is unresolved. Further work will resolve this second type of conjunction. As an othering phrase, Bush asserts the attributes of 'murderer' to the terrorists and the responsibility for bombing the USS Cole. A machine understanding of why this is an othering phrase must know that being a murderer and bombing a US warship connotes the behaviour of an outgroup. The required information to do so, nevertheless, is not in the text.

> *Sentence 5: The enemy of America is not our many Muslim friends; it is not our many Arab friends.*

| | RULE | PREDICATE | SUBJECT | ATTRIBUTE |
|---|---|---|---|---|
| **0** | SimpleAttribute | is not | The enemy of America | our many Muslim friends |
| **1** | SimpleAttribute | is not | it | our many Arab friends |

Figure 41. A negated clause

This final sentence is vital to correctly parse as it contains a negation; processing by word co-occurrence could wrongly assign Muslims and Arabs as an outgroup. In this sentence, Bush asserts that Muslims and Arabs are not the enemies of America; yet this outgroup term co-occurs with these named entities. To process a text by word co-occurrence may incorrectly associate 'the enemy of America' with 'Muslims' and 'Arabs'. Conversely, Figure 41 shows how the custom parser detects negation by including the negation modifier with the verb.

This section has shown how applying semantic analysis using grammar patterns more accurately reflects Bush's intended message and has shown some of the processing requirements to enable hybrid NLP. The custom dependency parser has applied these grammar patterns using pattern-based NLP to extract language clauses from Bush's declaration of war. Each clause then contains an assertion about his outgroup that Bush made to legitimise his 'War on Terror'. In comparison to a human interpretation of these sentences, this section has also revealed processing requirements for the proposed pipeline in the next section. Primarily, a human interpretation of these sentences resolves the abstract terms to the named entities to which they refer; any NLP claiming to understand a hostile narrative must do the same. The chapter now continues by summarising discovery work with spaCy components for the computational methods of hybrid-NLP for each objective in Figure 29.

## 4.3 What Are the Computational Methods of Hostile Narrative Analysis?

This final section records discovery work to develop computational methods for analysing hostile narratives using the hybrid-NLP. This discovery work is organised around each objective of the method in Figure 29 and is available online[71]. The purpose of each objective is as follows:

- Objective 1 is about text pre-processing for the subsequent objectives. These pre-processing tasks generally draw upon the quantitative method of text classification to label words by their grammatical, semantic, and syntactic properties.
- Objective 2 is about detecting social categorisation in natural language whereby an orator categorises named entities according to whether they are an ingroup or outgroup.
- Objective 3 then seeks to detect the clauses an orator uses to create a Self-other gradient between the groups identified in objective 2. The use of religion and ideology are a central feature of elevation and othering in these clauses.
- Objective 4 relies upon a quantitative method to score the Self-other gradient. The words associated with a group in a clause are scored to generate a value for elevation and othering.

The outputs of each objective then provide inputs to a rigorous explanatory dialogue to promote a human understanding of hostile narratives.

[71] Anning (2023) Hostile Narrative Analysis GitHub Repository, retrieved on 17th Feb 2023

### 4.3.1 What are the Pre-Processing Requirements for The Computational Methods of Hostile Narrative Analysis?

| Objective | Description | Pipeline Component | Output |
|---|---|---|---|
| **Pre-processing** | | | |
| **Obj 1.1** | Tokenisation - splitting a text into meaningful segments known as a token | spaCy tokeniser | A spaCy Doc object containing tokens |
| **Obj 1.2** | POS Tagging - Assign part-of-speech tags. | spaCy tagger | Tokens labelled with pos tags |
| **Obj 1.3** | Parser – assign dependency labels | spaCy dependency parser | Tokens with dependency labels |
| **Obj 1.4** | Lemmatizer – add a word lemma to each token | spaCy lemmatizer | Tokens with added lemmas as labels |
| **Obj 1.5** | Named entity recognition and disambiguation – detect and label named entities | spaCy named entity recognition<br><br>spaCy entity linker | Tokens with assigned entity labels |
| **Obj 1.6** | Named Concept Recognition – label the concepts of a text | Custom component | Tokens with assigned concept labels |
| **Obj 1.7** | Coreference resolution – assign named entities to pronouns and noun phrases. | Coreference resolution component | Resolution of noun phrases to named entities. |

Figure 42. The pre-processing requirement for the hostile narrative analysis method.

This first subsection addresses the pre-processing requirements for the hostile narrative analysis method. In addition to addressing the pre-processing problems identified in Chapter 2, Objective 1 must also address the abstract representation of groups in a text to reveal the target of violence in a hostile narrative. Figure 42 shows the pre-processing requirements identified thus far, and the section continues by explaining each.

#### 4.3.1.1 Objective 1.1 – 1.4: Labelling the Lexical and Grammatical Attributes of Words

Sentence 2

| | Lexical Unit | word |
|---|---|---|
| 0 | Noun | evidence |
| 1 | Pronoun | we |
| 2 | Verb | gathered |
| 3 | Noun | points |
| 4 | Noun | collection |
| 5 | Verb | affiliated |
| 6 | Noun | organizations |
| 7 | Verb | known |
| 8 | Proper Noun | al Qaeda |

Sentence 2 Noun Chunks

| | Noun Chunk |
|---|---|
| 0 | The evidence |
| 1 | we |
| 2 | all points |
| 3 | a collection |
| 4 | loosely affiliated terrorist organizations |
| 5 | al Qaeda |

Sentence 3

| | Lexical Unit | word |
|---|---|---|
| 0 | Noun | group |
| 1 | Pronoun | its |
| 2 | Noun | leader |
| 3 | Noun | person |
| 4 | Verb | named |
| 5 | Proper Noun | Usama bin Laden |
| 6 | Verb | linked |
| 7 | Noun | organizations |
| 8 | Noun | countries |
| 9 | Verb | including |
| 10 | Proper Noun | the Egyptian Islamic Jihad |
| 11 | Proper Noun | the Islamic Movement of Uzbekistan |

Sentence 3 Noun Chunks

| | Noun Chunk |
|---|---|
| 0 | This group |
| 1 | its leader |
| 2 | a person |
| 3 | Usama bin Laden |
| 4 | many other organizations |
| 5 | different countries |
| 6 | the Egyptian Islamic Jihad |
| 7 | the Islamic Movement of Uzbekistan |

Figure 43. The lexical units of sentences 1 and 2.

Objectives 1.1 to 1.4 generate a machine-readable Doc object from a text input. The constituent elements of this object then contain the lexical and grammatical attributes of each word. Figure 43 depicts the lexical units of sentences two and three whereby spaCy labels each word as either a noun, proper noun, pronoun or verb[72]. For Objective 1.1, tokenisation is the process of splitting text into minimal meaningful units such as words, punctuation marks [and] symbols' (Patel & Arasanipalai, 2021, p. 13). For Objective 1.2, part-of-speech tagging is about assigning an attribute

[72] Anning (2022) Semantic Labelling, retrieved on 17th Feb 2023

label to each token to indicate a word's grammatical function in a sentence (Bengfort et al., 2018, p. 44). For objective 1.3, dependency labelling assigns attributes to each token according to a word's grammatical function in a sentence. Lemmatisation refers to representing a word using its conical head, known as a lemma, and is used in NLP to generalise words by their different tenses or spellings. The noun chunks shown in Figure 43 are multi-word noun phrases that abstractly represent named entities. The spaCy components for these tasks operate to state-of-the-art standards and are not assessed further here.

The machine-readable objects spaCy's pipeline produces are Tokens, Docs and Spans. The tokeniser converts individual lexical items into a Token object, whether a word, punctuation symbol or whitespace[73], which contains the item's lexical attributes. A Doc object is a sequence of Token objects representing the complete text input [74]. A Span object is then a slice of the Doc object, which, in turn, comprises a sequence of Tokens. Returning to noun chunking in Chapter 2, a noun chunk is a Span object comprising several in-sequence Tokens. That the Doc and Span comprise a sequence of Tokens is important to understand. As the analysis of processing conjunctions in Chapter 2 explains, treating words as a sequence constrains the accurate representation of an orator's original message. This constraint of sequentially treating lexical items features later in this section.

#### 4.3.1.2 Objective 1.3: Custom Noun Chunks


Figure 44. A comparison of spaCy's noun chunker with the customised chunker.

Figure 44 compares spaCy's in-built noun chunker with a custom noun chunker developed for this research. The custom noun chunker responds to the problems identified in Chapter 2 by correctly pre-processing prepositional noun phrases and conjunctions. As a reminder, the problem is that spaCy's noun chunker does not parse prepositional phrases, thereby skewing an orator's intended

---

[73] Explosion.ai (2023) Token, retrieved on 17th Feb 2023

[74] Explosion.ai (2023) Doc, retrieved on 17th Feb 2023

use of words. The development of this custom noun chunker used example prepositional phrases from the hostile narrative corpus that spaCy would not otherwise detect, for example, 'war on terror' or 'weapons of mass destruction'. Against the test data, the spaCy's noun chunk tokeniser yielded an accuracy of 50% because it does not detect the prepositional phrases; in contrast, the custom component yielded a 44% improvement at 94% accuracy[75]. The inaccuracies are more about whether a word is correctly labelled with the necessary POS and dependency tags than any fault with the pattern. Nevertheless, this improvement is because the custom chunker correctly parses prepositional phrases to better represent an orator's intended use of words.

The example sentences in Figure 44 are drawn from the previous section and contain a series of now correctly chunked prepositional phrases. For example, spaCy chunks 'a collection of loosely affiliated terrorist organisations' as 'a collection' and 'loosely affiliated terrorist organisations'. In contrast, the custom chunker more accurately chunks this phrase as the preposition, 'a collection of loosely affiliated terrorist organisations'. Equally, spaCy incorrectly chunks 'the enemy of America' as 'the enemy' and 'America'; the custom chunker correctly chunks this noun preposition as 'the enemy of America'. While the custom chunker works for noun phrases, the sequential treatment of Token objects by spaCy constrains the correct chunking of conjunctions. This improvement in tokenising noun chunks means the pre-processed Doc reflects the orator's original message more accurately than the established practices identified in Chapter 2.

The conjunction chunks for sentence 4 are 'some of the murderers indicted for bombing American embassies in Tanzania and Kenya' and 'responsible for bombing the USS Cole'. The first chunk, however, contains a further conjunction that should be split as 'some of the murderers indicted for bombing American embassies in Tanzania' and 'some of the murderers indicted for bombing American embassies in Kenya'.

Figure 40 shows how the custom dependency parser can correctly parse conjunctions when they are separate noun chunks. The output, nevertheless, is a string object rather than a Span object. The span slice for the first conjunction would start with 'some' and end with 'Tanzania', while the slice for the second begins with 'some' and ends with 'Kenya'. Slicing the Doc object does not permit the exclusion of 'Tanzania and' for the second chunk. Treating words as sequential means conjunctions within a noun chunk are incorrectly processed.

[75] Anning (2023) Tokenising Noun Chunks, retrieved on 17th Feb 2023

#### 4.3.1.3 Objective 1.6: Named Entity Recognition and Resolution

The second review assesses the accuracy of spaCy's language model for NER and a spaCy add-on for entity resolution[76]. SpaCy labels named entities using the OntoNotes 5 NER schema, which has 18 entity types (Weischedel et al., 2012). SpaCy's classification algorithm uses contextual information about a word to assign the label. As such, the algorithm differentiates the named entity 'US' from the pronoun 'us'. Entity-Fishing[77] is a spaCy add-on for the entity fishing tool[78] for named entity resolution. The algorithm uses Wikidata[79] to provide an ID that uniquely identifies a named entity. Of particular utility for explanatory dialogues, entity-fishing also provides a confidence score for the ID, the associated Wikidata description for the named entity, and alternate IDs. These additional outputs offer a rationale within an explanatory dialogue for assessment by human judgement.

| | Named Entity | Label | OntoNotes Description | Wikidata ID | Nerd Score | Normal term |
|---|---|---|---|---|---|---|
| 0 | al Qaeda | ORG | Companies, agencies, institutions, etc. | Q34490 | 0.8919 | Al-Qaeda |
| 1 | Usama bin Laden | PERSON | People, including fictional | Q1317 | 0.9258 | Osama bin Laden |
| 2 | the Egyptian Islamic Jihad | ORG | Companies, agencies, institutions, etc. | Q310214 | 0.3634 | Egyptian Islamic Jihad |
| 3 | the Islamic Movement of Uzbekistan | ORG | Companies, agencies, institutions, etc. | None | NaN | None |
| 4 | al Qaeda | ORG | Companies, agencies, institutions, etc. | Q34490 | 0.8919 | Al-Qaeda |
| 5 | Afghanistan | GPE | Countries, cities, states | Q889 | 0.3812 | Afghanistan |
| 6 | Taliban | ORG | Companies, agencies, institutions, etc. | Q42418 | 0.6837 | Taliban |

Figure 45. The named entities and Wikidata IDs for the test sentences.

Figure 45 shows how spaCy and Entity-Fishing detect and resolve the named entities from the test sentences. Named entity recognition (NER) and resolution enable objective 2 to detect groups in a text for subsequent classification as an ingroup or outgroup. Named entity recognition is the 'task of classifying tokens of interest in a sequence of tokens into specific entity types, such as a person, an organisation, or a location' (Patel & Arasanipalai, 2021, p. 62). Entity resolution is about assigning a unique identifier to different mentions of the same named entity across a text. Named entity resolution means an individual ID identifies the same entity. For example, 'the United States of America' as a named entity is also represented by the terms 'USA', 'US' and the 'United States', therefore, each should attract the same unique ID.

[76] Anning (2023) Named Entity Recognition, retrieved on 17th Feb 2023

[77] Luccaterre (2022) spacyfishing, retrieved on 17th Feb 2023

[78] Yas1994 (2022) entity-fishing, retrieved on 17th Feb 2023

[79] Wikimedia Foundation (n.d.) wikidata, retrieved on 17th Feb 2023

The named entity column shows the original text. The label and description column each show the spaCy label and Ontonotes description. Note how the labels do not introduce biases; for example, ‘al Qaeda’ is labelled as an organisation, not a terrorist group. The Wikidata ID column shows the unique identifier from Wikidata, while the Nerd Score is the confidence level for that identifier. The ‘normal term’ is the normative reference for the original text. Notably, while Wikidata recognises ‘Usama bin Laden’ and ‘the Egyptian Islamic Jihad’, it does not recognise ‘the Islamic Movement of Uzbekistan’. A review evaluated the label assigned to each named entity using external sources, and an Islamic academic was consulted for bin Laden’s text to ensure accuracy. In response, these corrections informed the development of a customer component, with 65 corrections for bin Laden and 31 for Bush[80].

#### 4.3.1.4 Objective 1.5: Named Concept Recognition

| **text** | our | enemy | is | a | radical | network | of | terrorists | , | and | every | government | that | supports | them |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **lemma** | our | enemy | be | a | radical | network | of | terrorist | , | and | every | government | that | support | they |
| **head** | enemy | is | is | network | network | is | network | of | is | is | government | is | supports | government | supports |
| **ent type** | | | | | | | | | | | | | | | |
| **concept** | SELF | ADVERSARY | | | | | | CRIMEGROUP | | | | GPEGROUP | | | |
| **attribute** | ingroup | outgroup | | | | | | outgroup | | | | identity | | | |
| **ideology** | social | military | | | | | | security | | | | geopolitics | | | |

Figure 46. Applying named concept recognition.

Under the idea of named concept recognition, Figure 46 shows the output of a developing pipeline component for labelling a word’s connotative meaning. Named concept recognition draws upon quantitative coding to label nouns associated with religion and ideology. ‘Codes are commonly created prior to data collection…concepts and hypotheses are most often developed in advance, and categories and their codes are derived deductively from theory or borrowed from the extant literature’ (Benaquisto, 2008, p. 85). Accordingly, the coding schema for this research was developed through several reviews of the data to deduce the underlying contexts. The schema groups synonymous concepts from Bush and bin Laden’s speeches into eight different group contexts: social, medical, academic, religious, political, economic, security and military[81] Each context then contains various code words to signal different group attributes. Nevertheless, further research is required to assess how well this schema generalises beyond the hostile narrative corpus. As will be explained in the next subsection, this schema enables the

---

[80] Anning (2023) Named Entity Corrections, retrieved on 17th Feb 2023

[81] Anning (2023) Named Concept Recognition, retrieved on 17th Feb 2023

identification of ingroups and outgroups in objective 2 and the identification of elevation and othering in objective 3 of the method.

```
They are some of the murderers indicted for bombing American embassies in Tanzania and Kenya, and responsible for bombing the USS Cole
They are some of the CRIMEGROUP indicted for MILACTION American GPEENTITY in Tanzania and Kenya, and responsible for MILACTION the USS Cole


The evidence we have gathered all points to a collection of loosely affiliated terrorist organisations known as al Qaeda.
The evidence SELF have gathered all points to a collection of loosely affiliated CRIMEGROUP SOCIALGROUP known as al Qaeda.


Our enemy is a radical network of terrorists, and every government that supports them
SELF ADVERSARY is a radical network of CRIMEGROUP , and every GPEGROUP that supports them


The enemy of America is not our many Muslim friends; it is not our many Arab friends
The ADVERSARY of America is not SELF many Muslim AFFILIATE ; it is not SELF many Arab AFFILIATE
```

Figure 47. Applying the connotative meaning of words to the example sentences.

Figure 46 shows how each concept provides labels to indicate a word’s connotative meaning. The pronoun ‘our’ connotes an ingroup, while the noun ‘enemy’ and the adjective ‘terrorist’ connotes an outgroup. ‘Government’ connotes a more neutral meaning of a group of people. How the orator links these words then assists with understanding the sentence’s meaning. The possessional modifier between ‘our’ and ‘enemy’ suggests the ingroup has an outgroup possession. The clause head ‘is’ then asserts two attributes onto the outgroup. The first attribute is the noun preposition ‘radical network of ADVERSARY’, and the second is ‘every GPEENTITY that supports them’. Figure 47 then shows a similar application of connotative meaning for the other example sentences. The schema used in this chapter is a simple JSON object, whereas a more production-ready version requires knowledge graphs. The next chapter proposes this knowledge graph development as further work.

#### 4.3.1.5 Objective 1.7: Coreference Resolution

While the preceding objectives prepare data for processing, coreference resolution directly addresses abstract representations of groups. Coreference resolution is about identifying all the pronouns and noun phrases that refer to the same named entity. As explained in Chapter 2, abstract representations are a particular feature of othering in a hostile narrative. Resolving these representations to their referent named entity then identifies the target of hostility. For objective 2, coreference resolution assists with identifying the noun phrase an orator uses to represent their ingroup or outgroup. For objective 3, coreference resolution detects the elevation and othering statements that refer to a named entity in the subject of a clause.

The following assessment summarises the performance of three coreference resolution algorithms for the spaCy pipeline. The first algorithm is ‘coreferee’ from explosion.ai; the second is 'neuralcoref', developed by Huggingface[82]; the third is 'coref', provided by AllenNLP[83]. While 'neuralcoref' draws upon neural networks only, both 'coref' and 'coreferee' also include a mixture of rules and machine learning. The assessment reviews each algorithm's output using the example sentences above to assess how well they resolve the noun phrases referring to the groups in Bush's text. The assessment shows that co-reference resolution cannot resolve hyponymic relations in a text, such as resolving all or Bush's named outgroups to the noun phrase 'terrorist'.

| | Sentence 1 |
|---|---|
| **All Algorithms** | The evidence we have gathered all points to a collection of loosely affiliated terrorist organizations known as al Qaeda. |

| | Sentence 2 |
|---|---|
| **Coreferee** | This group and This group and its leader -- a person named Usama bin Laden -- leader -- a person named Usama bin Laden -- are linked to many other organizations in different countries , including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan |
| **Coref** | This group and Al Qaeda leader -- a person named Usama bin Laden -- are linked to many other organizations in different countries , including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan |
| **AllenNLP** | a collection of loosely affiliated terrorist organizations known as al Qaeda and a collection of loosely affiliated terrorist organizations known as al Qaeda's leader -- a person named Usama bin Laden -- are linked to many other organizations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan. |

| | Sentence 4 |
|---|---|
| **Coreferee** | Americans hate our freedoms |
| **Coref** | Their, leaders hate our freedoms. |
| **AllenNLP** | Terrorists hate Americans's freedoms. |

| | Sentence 5 |
|---|---|
| **Coreferee** | These terrorists kill not merely to end lives , but to disrupt and end a way of life. |
| **Coref** | These terrorists These terrorists kill not merely to end lives , but to disrupt and end a way of life. |
| **AllenNLP** | terrorists kill not merely to end lives, but to disrupt and end a way of life. |

Figure 48. Coreference outputs of sentences 1 and 2.

Figure 48 summarises the assessment with each algorithm's coreference outputs for sentences 1, 2, 4 and 5. Most notably, Coreferee incorrectly resolves 'they' in sentence four with 'Americans'. In context, Bush barely mentions his outgroups in the section this sentence is drawn from; however, 'Americans' is most proximate to 'they'. This is a misrepresentation of Bush's text.

Six sentences separate sentences one and two with only one other mention of 'al Qaeda' between them. Coreferee fails to make this connection and misrepresents sentence two. Coref, on the other hand, makes the connection between 'al Qaeda' and the pronoun 'its' and produces a plausible output. AllenNLP successfully connects both 'this group' and 'its' to the entire noun phrase from sentence 1, 'a collection of loosely affiliated terrorist organisations known as al Qaeda'. Coref, and AllenNLP at least for the sentences reviewed here, therefore, have identified Bush's outgroup to each elevation and othering statement for objective 2.

[82] HuggingFace (2021) neuralcoref, retrieved on 17th Feb 2023

[83] AllenNLP (2021) coref, retrieved on 17th Feb 2023

While the AllenNLP and Coref algorithms generate plausible outputs, this assessment shows how these algorithms are limited by the number of entities they can resolve to a noun phrase. Where a human subconsciously resolves the four terrorist organisations to ‘they’ and ‘these terrorists’, coreference algorithms only link a single reference. In effect, ‘terrorists’ serves as a hyponym for each named entity that these co-reference algorithms fail to detect. A similar problem likely exists with bin Laden’s equivalent outgroup representation, ‘Jewish-crusade alliance’. As such, the algorithms work well for shorter texts, but manual intervention through an explanatory dialogue is required to connect these hyponymic relations over longer narratives. Having reviewed the pre-processing requirements, the section continues with the analytical objectives for identifying the ingroup and outputs along with the detection and analysis of language clauses.

### 4.3.2 How Does Pattern-Based NLP Enhance Explanatory Dialogues About Hostile Narratives?

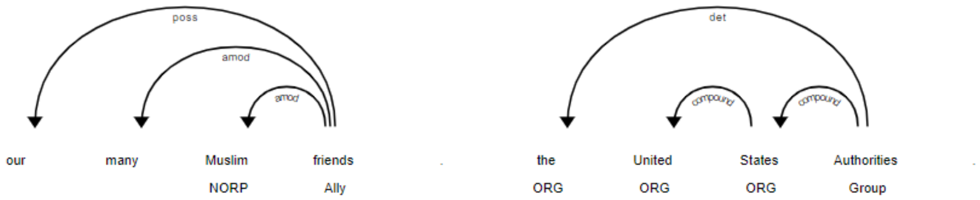
| | RULE | PREDICATE | SUBJECT | DIRECTOBJECT | PREPOSITION | PREPOSITIONALOBJECT | ATTRIBUTE |
|---|---|---|---|---|---|---|---|
| 0 | SimpleDirectObject | committed | enemies of freedom | an act of war against our country | | | |
| 1 | SimpleNounPreposition | committed | enemies of freedom | | On | September the 11th | |
| 2 | SimpleDirectObject | hate | Terrorists | our freedoms -- our freedom of religion, our freedom of speech, our freedom to vote and assemble and disagree with each other | | | |
| 3 | SimpleNounPreposition | linked | Al Qaeda | | to | many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan | |
| 4 | SimpleNounPreposition | linked | Al Qaeda's leader -- a person named Usama bin Laden -- | | to | many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan | |
| 5 | SimpleAttribute | is not | The enemy of America | | | | our many Muslim friends |
| 6 | SimpleAttribute | is not | The enemy of America | | | | our many Arab friends |
| 7 | SimpleAttribute | is | Our enemy | | | | a radical network of terrorists, |
| 8 | SimpleAttribute | is | Our enemy | | | | every government that supports them |
| 9 | SimpleAttribute | are | Terrorists | | | | some of the murderers indicted for bombing American embassies in Tanzania |
| 10 | SimpleAttribute | are | Terrorists | | | | responsible for bombing the USS Cole |

Figure 49. A parse of selected sentences from Bush

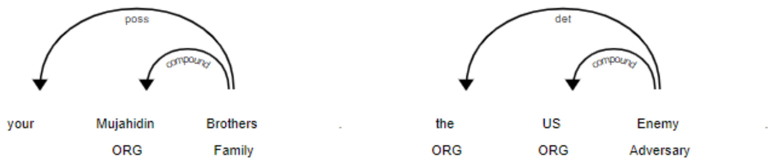
The first objective is about pre-processing, this next subsection covers the analytical objectives and shows how pattern-based NLP enhances explanatory dialogues about hostile narratives with rationales to explain elevation and othering in text. Figure 49 shows a representation of selected clauses from Bush’s text. The clauses are derived from grammar patterns, and each contain an assertion made by Bush. He asserts the enemy as ‘a radical network of terrorists’ and ‘every government that supports them’. In effect, each assertion is an elevation or othering statement to create the necessary Self-other gradient for legitimising violence. This section reviews how hybrid-NLP enables the generation of a rationale to explain the hostility of each assertion.

This section reviews three additional language patterns for making these assertions: framing, naming and hypernymy. Each subsection contains an explanation of each pattern and the corresponding computational methods. The output of each computational method is a visual depiction of each clause and the accompanying rationale. Beneath each word is either the named entity or concept label. The rationale applies these labels in a standardised framework and relates the pattern to a relevant methodological objective from the second section. Preliminary testing of the hypernymy pattern using a custom component yields an f1-score of 0.72[84]. The test assessed the hypernyms identified by the custom component against manually generated hypernyms from the hostile narrative corpus. Improvements to this score are possible with improvements to the pattern. Compared to processing text by word co-occurrence, these subsections suggest the improved explanatory value of treating text as structured data using hybrid NLP.

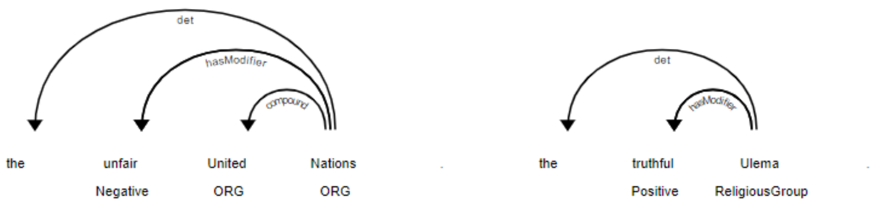
#### 4.3.2.1 The Framing Pattern




```
'Muslim friends' has a 'ingroup' classification where 'friends' is an 'Ally' phrase from the 'Social' context.
'the United States Authorities' has a 'neutral' classification where 'Authorities' is an 'Group' phrase from the 'Political'
context.
```




```
'Mujahidin Brothers' has a 'ingroup' classification where 'Brothers' is an 'Family' phrase from the 'Social' context.
'the US Enemy' has a 'outgroup' classification where 'Enemy' is an 'Adversary' phrase from the 'Military' context.
```




```
'United Nations' is othered by the 'adjectival modifier (amod)' term 'unfair' from the 'Social' context.'
'Ulema' has a 'elevation' classification where 'truthful' is an 'Positive' phrase from the 'Social' context.
'Ulema' is elevated by the 'adjectival modifier (amod)' term 'truthful' from the 'Social' context.'
```

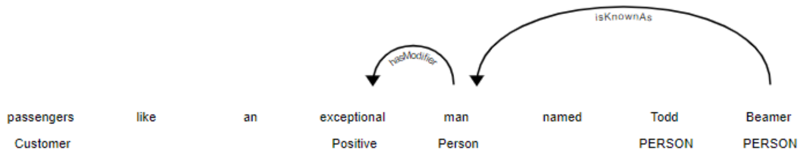
Figure 50. Output for the framing pattern.

The framing pattern detects modifier relationships between entities and concepts.

[84] Anning (2023) Detecting the Ingroup and Outgroup of a text, retrieved on 17th Feb 2023

Figure 50 shows parses for four wrongly classified named entities in ‘Experiment 3: Detect the Ingroup and Outgroup’ from Chapter 2. Bush’s incorrectly scored named entities were ‘Muslim’ (-0.83), ‘the United States Authorities’ (-0.64), and bin Laden’s are ‘Mujahidin Brothers’ (-0.58) and ‘US Enemy’ (+0.42). The rationale for the first four parses uses concepts of ‘friend’, ‘family’, or ‘adversary’ to identify each orator's ingroup and outgroups. Equally, the final two parses show how bin Laden uses *adjectival modifiers* (amod) for elevation and othering. In contrast to the counter-intuitive results in Experiment 3, these results more accurately reflect each orator's intended message.

#### 4.3.2.2 The Naming Query




```
'passengers' has a 'neutral' classification where 'passengers' is an 'Customer' phrase from the 'Commercial' context.
'exceptional man' has a 'elevation' classification where 'exceptional' is an 'Positive' phrase from the 'Social' context.
'Todd Beamer' is referred to as 'an exceptional man', which is a 'elevation' term from the 'Social' context.
```

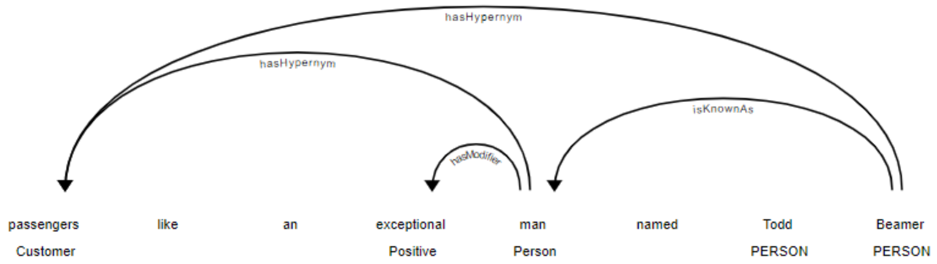

```
'terrorist organizations' has a 'outgroup' classification where 'terrorist' is an 'Criminal' phrase from the 'Security' context.
'al Qaeda' is referred to as 'loosely affiliated terrorist organizations', which is a 'outgroup' term from the 'Security' context.
```

Figure 51. Output for the naming pattern.

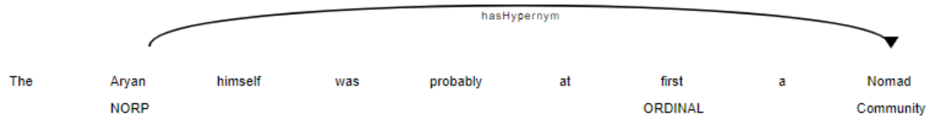
The naming pattern shown in Figure 51 is about parsing noun phrases. This parsing of noun phrases also uses the subject->predicate->object model for parsing natural language. With ‘known’ or ‘named’ as the predicate, 20 similar terms from VerbNet could be added to scale the query[85]. Applying this naming pattern to Todd Beamer and al Qaeda is particularly interesting. As a passenger on United Airlines Flight 93, which al Qaeda hijacked, Bush lionises Todd Beamer for his attempt to regain the plane. In contrast, Bush declares war against al Qaeda in response. The same query applies to each clause, but the rationale framework generates different outputs in accordance with the named concepts.

[85] VerbNet (2021) dub-29.3, retrieved on 17th Feb 2023

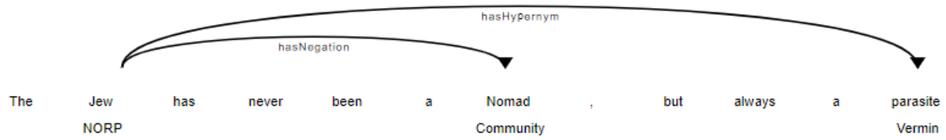
#### 4.3.2.3 The Hypernymy Patterns




```
'passengers' has a 'neutral' classification where 'passengers' is an 'Customer' phrase from the 'Commercial' context.
'exceptional man' has a 'elevation' classification where 'exceptional' is an 'Positive' phrase from the 'Social' context.
'Todd Beamer' is referred to as 'an exceptional man', which is a 'elevation' term from the 'Social' context.
'Todd Beamer' is referred to as 'passengers', which is a 'neutral' term from the 'Commercial' context.
```




```
'Aryan' is referred to as 'a Nomad', which is a 'neutral' term from the 'Social' context.
'Nomad' has a 'neutral' classification where 'Nomad' is an 'Community' phrase from the 'Social' context.
```




```
'Jew' is referred to as 'a parasite', which is a 'outgroup' term from the 'Health' context.
WARNING: the term 'parasite' referring to 'Jew' is from the 'Vermin' category and is often used in genocidal language.
'Jew' is disassociated from 'a Nomad', which is a 'neutral' term from the 'Social' context.
'Nomad' has a 'neutral' classification where 'Nomad' is an 'Community' phrase from the 'Social' context.
'parasite' has a 'outgroup' classification where 'parasite' is an 'Vermin' phrase from the 'Health' context.
```

Figure 52. Output for the hypernymy pattern.

For detecting the classification of a named entity in a narrative clause, the hypernym pattern applies Hearst Patterns. Hearst (1992) introduced hypernymy as a ‘way to discover a hyponymic lexical relationship between two or more noun phrases in a naturally occurring text’ (Hearst, 1992, p. 539). Hearst patterns identify the relationship between hypernyms and hyponyms using the subject->predicate->object model. The *Mein Kampf* sentence section 1 of this chapter and shown in Figure 52 contain hypernyms for othering. In clause 1, the term nomad is the hypernym for categorising Aryans, while parasite as an outgroup term is the hypernym of clause 4. Clause 2 contains a negation term which disqualifies it as a hyponymic relation. To explain the more general application of Hearst Patterns, the predicate for clause one could change to ‘be an example of’ or ‘like other’ and Aryans would still qualify as Nomads. Using Hearst Patters to link named entities to hypernyms that represent either an ingroup or outgroup, therefore, provides a way to detect their group classification in a text.

Figure 52 shows the results for applying the hypernymy pattern to the Todd Beamer clause pattern and the two *Mein Kampf* statements. The Todd Beamer rationale explains how he was a passenger as a hyponym of the hypernym term, 'passenger'. In the first *Mein Kampf* statement, the term 'nomad' is a hypernym for categorising 'Aryans'. The second clause disassociates Jews from Nomads through a negation modifier. Word co-occurrence is unlikely to detect this negation. Additionally, the 'warning' rationale responds to this clause being a particular feature of genocidal narratives and would fall within Matsuda's original application of hate speech. The algorithm is detecting how the orator uses the word rather than its proximate location to other words. Suggesting general applicability, replacing 'Jews' with 'Muslims' in this clause would mean Islamophobia.

The scoring schema for objective 3 of the methodology requires the linking of subjects of a clause to each other and to resolve the hyponymic relations. Figure 49 represents the beginning of how to link these narrative elements in a text. The next step, as identified in further work, is a proposal to link them using knowledge graphs. The queries over these graphs can then assign a score to each entity based on the attributes an orator asserts onto them.

## 4.4 Discussion

In response to the research question, this chapter has shown how applying semantic analysis through hybrid-NLP can generate rationales that explain hostility in natural language as inputs into an explanatory dialogue about hostile narratives. The hybrid approach to the computational methods of hostile narrative analysis draws upon pattern-based NLP to parse the language clauses of a text. The quantitative methods draw upon machine learning for text pre-processing to label a word's grammatical, semantic and lexical properties. The qualitative aspect first draws upon cultural violence theory as a basis for an explanation. Hybrid NLP parses language clauses to detect an ingroup and outgroup and how they are elevated and othered across a narrative.

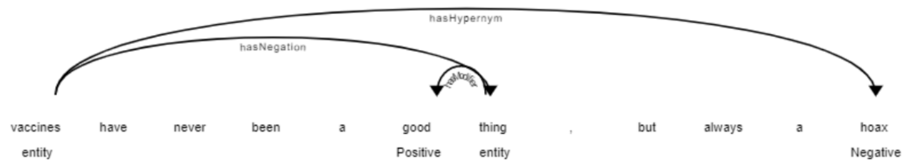
The pipeline component presented in this chapter can be developed further with the continued creation of grammar patterns. Developing grammar patterns means recording the many permutations of language clauses, which is time-consuming and has been the subject of technical development. Roller et al. (2018) use word embeddings to infer hyponymic relations (Roller et al., 2018); Issa et al. (2018) use dependency relations (Issa et al., 2018). Word embeddings infer dependency relations by the co-occurrence of words, which does enable scalable systems, but does not account for information provided by a predicate head, therefore, narrative information would be missed. Standford CoreNLP also provides a dependency parser[86]; however, unlike

[86] CoreNLP (2021) Dependency Parser Demo, retrieved on 17th Feb 2023

spaCy's version, CoreNLP does not link words using the pattern grammar model. Pattern grammar, nevertheless, suggests the further development of patterns is a bounded task with a finite possibility of patterns. The most feasible way to develop these patterns is to manually develop them using grammar patterns.

The generalisability of the proposed methods for this pipeline depends on overcoming a series of limitations in the approach. In the first instance, labelling words of these clause relies upon text classification, which has high accuracy in spaCy models, but there is still the potential for inaccuracies. Secondly, ubiquitous ambiguity also presents a problem for how to label the grammatical relations between words. As explain in Chapter 4, different grammatical relations can present very different meanings in a clause. In the example given, either the orator was wearing pyjamas while shooting an elephant, or the elephant was wearing pyjama while being shot. While there is an obvious error in this latter image of an elephant wearing pyjamas, there are assumed more subtle examples where the error is much less certain. Thirdly, modelling language clauses assumes well written English. This assumption applies to political speeches and published works, but as some of the tweets in Chapter 5 show, these patterns are unlikely to detect poorly written English, which is assumed to be more typical for social media texts. Finally, these patterns would require constant review and maintenance in a production system. Each of these limitations would have to be addressed as part of the continued development.

Motivating the continued development of grammar patterns and labelling schema is the potential for new use cases for hostile narrative analysis in future research. While this paper has focused on narrative clauses, modelling them does enable the analysis of hostile narrative structures. Just as Propp (1968), Labov and Waletzky (1997), and Riessman (2005) reviewed folktales and personal narratives to determine the structure, a modified review of the declarations of war analysis presented in this paper could reveal common syntagmatic structures. The clause and narrative structures should, in turn, apply to analysing hostile narratives on online platforms. These ideas are explored in further work in the next chapter.




```
'vaccines' has a 'neutral' classification where 'vaccines' is an 'entity' phrase from the 'Health' context.
'vaccines' is disassociated from 'a good thing', which is a 'elevation' term from the 'Social' context.
'good thing' has a 'elevation' classification where 'good' is an 'Positive' phrase from the 'Social' context.
'hoax' has a 'othering' classification where 'hoax' is an 'Negative' phrase from the 'Social' context.
```

Figure 53. Applying grammar patterns to the domain of pandemic response.

Beyond the analysis of hostile narrative, Figure 53 shows the application of a hypernymy grammar pattern to a made-up, but plausible, sentence about the Covid-19 pandemic. The labels for 'vaccine' and 'hoax' have been added to the named concept recognition schema. The syntactic and rationale pattern displayed here is the same as for the second *Mein Kampf* sentence from Figure 52, but the output changes according to how words are labelled. This example suggests pattern-based NLP can be broadly applied to different domains. This more general application then motivates the continued development of hybrid-NLP for social science applications.

## 4.5 Conclusion

Concerning the research hypothesis, this chapter presents a proof of concept to show how incorporating qualitative methods into NLP improves the meaningful analysis of hostile narratives. The qualitative methods start with the methodology presented in the previous chapter. This methodology provides a framework to explain how an orator may seek to legitimise violence. This methodology is validated in how it features in Bush's and bin Laden's declarations of war. The method presented in this chapter is then derived from the methodology. The computational methods for each objective apply hybrid-NLP by augmenting quantitative approaches to NLP with semantic analysis. Hybrid NLP then detects the language clauses that identify ingroups and outgroups along with elevation and othering statements for each methodological objective. In contrast to the exclusively quantitative methods reviewed in Chapter 2, this more hybrid approach enables the generation of rationales to explain why a text may be hostile. While these rationales might be questioned, the point is that they can be either accepted, rejected or modified as part of an explanatory dialogue.

Most of all, this chapter shows how to connect social scientific theories with NLP. Pattern based NLP is an application of semantic analysis using grammar patterns. The parsing of text by language clauses is consistent with both linguistic and narrative theory. Of particular interest, recall Van Dijk's (1983) from the introduction who suggests narratives are a social database of groups. The social database in Figure 49 begins to represent the social database of Bush's narrative and those who subscribes to the war on terror. Moreover, the assertions contained in the clause this database captures represent Galtung's idea of social cosmology, or belief systems Bush sought to promote. Queries into this database has the potential to provide new insights into the social databases of group narratives in social science applications.

# Chapter 5 Developing Hostile Narrative Analysis to Tackle Online Abuse

This thesis begins with a research hypothesis that questions the effectiveness of exclusively quantitative approaches to NLP in social science applications. For the benefit of testing this hypothesis, social science applications generate outputs with broad, sometimes opposing, degrees of interpretation by audiences. In contrast, the outputs of functional applications for which quantitative methods are effective have narrower degrees of interpretation. As a focus for this thesis, the introduction problematises Explainable AI by situating the application of social science applications in explanatory dialogues. The assessment of quantitative or hybrid approaches is then about how they contribute to a human understanding of hate speech, sentiment analysis or hostile narrative analysis. Presently, quantitative methods dominate the application of NLP to social science applications.

The thesis introduction also explained a somewhat problematic tension between the use of quantitative and qualitative methods in research. This tension arises from the expectations of each approach. Researchers generally expect high degrees of objectivity from quantitative methods to give a sense of certainty and predictability to research outputs. Conversely, qualitative methods are about capturing people's subjective experiences. As such, qualitative methods produce much less certain results than quantitative methods because people are inherently unpredictable. Hall's (1974) encoding and decoding theory also explains how the outputs of social science applications are very subjective. The research questions for exploring this hypothesis, therefore, question the appropriateness of a purely quantitative approach for subjectively defined NLP applications.

This final chapter begins by explaining this thesis's policy context concerning emerging UK legislation to tackle online abuse and recent announcements for developing the AI industry. The chapter then continues by responding to each research question, how they support the research hypothesis and their implications concerning the national context. The response to these questions explains how tackling online abuse requires a hybrid approach to NLP to facilitate productive dialogue between content moderators, the users of online platforms and the OFCOM regulator. Before concluding, the chapter proposes further work for developing the methodology and computational methods presented in this thesis. This further work is about applying narrative theory to the methodology and applying knowledge graphs to a production-ready NLP pipeline for analysing hostile narratives.

The research hypothesis and questions from the introduction are reproduced here.

> **Research Hypothesis**: Integrating qualitative methods with NLP improves the meaningful analysis of hostile narratives.
>
> **RQ1**: To what extent do quantitative methods in NLP 'understand' social science applications?
>
> **RQ2**: How can integrating Peace Research and NLP enable the meaningful analysis of hostile narratives?
>
> **RQ3**: How does augmenting quantitative NLP methods with qualitative approaches enable the meaningful analysis of hostile narratives?

The chapter now continues by introducing the policy context.

## 5.1 What is the Policy Context for Developing Hostile Narrative Analysis?

The policy context for developing hostile narrative analysis is in the proposed Online Safety Bill and developing the UK's AI industry. According to a UK Government guide, the Online Safety Bill 'is a new set of laws to protect children and adults online' by making 'social media companies more responsible for their users' safety on their platforms'[87]. The general idea is to keep internet-based services free of illegal and harmful content while defending Free Speech. Currently, online platforms only become liable for problematic content after being made aware of it. However, the Bill's current version obligates internet service providers to make 'proactive use of technological tools, where appropriate, to identify, flag, block or remove illegal or harmful content[88]'.

The necessity of this Bill is in no doubt. Chapter 3 explains the rather unfortunate relationship between the propagation of hate speech and the growth of the internet. The Bill's white paper additionally explains how 'terrorist groups use the internet to spread propaganda designed to radicalise vulnerable people, and distribute material designed to aid or abet terrorist attacks'[89]. Indeed, Paul Dunleany, a teenager, was jailed in Nov 2020 for preparing acts of terrorism after joining a neo-Nazi group called Feuerkrieg Division (FKD) through an online chat group. According to the head of the West Midlands Counter Terrorism Unit, 'this boy had an unhealthy interest in

---

[87] UK Government (2022) A guide to the Online Safety Bill, retrieved on 17th Feb 2023

[88] UK Government (2020) Online Harms White Paper, retrieved on 17th Feb 2023

[89] UK Government (2020) Online Harms White Paper, retrieved on 17th Feb 2023

other attacks across the world and he knew exactly what online platforms to join to share his extreme views'[90]. Moreover, Kingdon's (2017) semiotic analysis of over 100 hours of Islamic State propaganda explains the role of online platforms in radicalisation. Her research 'identified the themes of seduction, grievance, utopia, military warfare, and theatrical displays of violence, all of which serve as powerful recruitment strategies' (Kingdon, 2017, p. 3).

In addition to the problems of online radicalisation, the Bill cites the problem of online hate speech. As such, the white paper also includes provisions for online hate crimes that demonstrate 'hostility on the grounds of an individual's actual or perceived race, religion, sexual orientation, disability or transgender identity'[91]. A 2022 report by Tell MAMA, an independent, non-governmental organisation which works on tackling anti-Muslim hatred, reveals the role of online platforms in hate crime (TellMAMA, 2022). The report is a case study of Andrew Leak, who threw firebombs at an immigration processing centre in Dover[92]. Firstly, the report shows Leak's extensive engagement with Far Right, anti-Muslim, homophobic, transphobic and conspiracy theory content on both Facebook and Twitter. Secondly, the report provides examples of times when Leak posted violent messages on online platforms. Whether radicalisation or hate crime, the propagation of online propaganda leads to the potential for real-world violence.

Examples of Leak's violent messages in the TellMAMA report highlight some of the main problems of detecting online hate reviewed in Chapter 2. Some messages contain benign content, while others are more obviously hateful. In a May 2022 tweet, Leak commented, 'If you own a jet ski we are starting up we are starting up our own patrols join us'. In isolation, this message contains no apparent indicators of hatefulness; in context, it refers to patrolling the English Channel for immigrants crossing in small boats. Responding to stories about refugees in another tweet, he asserts, 'What else did they expect keep poking keep poking and you will be poked back'. To understand why this message is hateful is to resolve 'they' to refugees and that 'poking' refers to the idiom of 'don't poke the bear'. This idiom imagines the probably violent consequence of literally poking a bear. In effect, his victim blaming messages suggests refugees seeking asylum are stoking problems and should expect a violent response. Both these messages are examples of 'dog whistles' that even humans, let alone machines, may struggle to decode.

---

[90] BBC (2020) Rugby teenager Paul Dunleavy jailed for terror offences, retrieved on 17th Feb 2023

[91] UK Government (2020) Online Harms White Paper, retrieved on 17th Feb 2023

[92] BBC (2022) Dover migrant centre attack: Firebomber died of asphyxiation, inquest told, retrieved on 17th Feb 2023

Other examples of Leak's messages are much more overtly hateful. In one post, he asserts, 'We in the UK have a serious problem with Pakistani Muslim grooming gangs', adding, 'this is part of their culture it is racism the reason is do with race it's because they're only raping white Christian girls [sic]'. In another post, he develops this message by asserting, 'all Muslims are guilty of grooming…they only rape non-Muslims and only rape white Christian girls'. Each post asserts a hateful stereotype of child abuse in Muslim communities. The first comment asserts 'Muslim grooming gangs' as a threat to both Leak's ingroup of the UK and to 'White Christian girls' as members of his ingroup. Leak's hateful comments, nevertheless, are not just restricted to Muslims. In a homophobic message, he asks people to 'just remember gay men abuse young boys'. These more overtly hateful messages combine with the dog whistles to create a pernicious nationalist narrative against imagined threats to the UK.

The proactive detection of such harmful content has implications for how the Bill is regulated. In cases where online platforms fail in their duty of care to users, the planned Bill empowers communications watchdog Ofcom to fine a social media company up to £18m, or 10% of its annual turnover if that is higher. To enforce such fines, Ofcom will require a clear definition of what constitutes harmful content otherwise they risk lengthy litigation with online platforms. Therefore, the technical element of detecting harmful content will require clear algorithmic explanations as to why a particular message should be prohibited.

Nonetheless, a Big Brother Watch report observes how the Bill takes a rather technical view of detecting harmful content with algorithms.

> *The Bill makes repeated references to different types of 'technology' that regulated services may use to guarantee compliance with their relevant duties. This is often an endorsement of algorithmic content moderation tools which surveil users' online activity and make blunt, inaccurate and often biased judgements on the permissibility of online expression proactive detection of harmful content*[93]*.*

The findings of this thesis that are explained in the next section very much support this observation. In addition to the Online Safety Bill, developing UK's Artificial Intelligence industry provides a broader policy context.

---

[93] Big Brother Watch (2022) Big Brother Watch's Briefing on the Online Safety Bill for House of Commons Report Stage, retrieved on 17th Feb 2023

UK government policy has centred on mathematics to grow the UK's AI industry. In Oct 2017, Professor Dame Wendy Hall and Jérôme Pesenti published a government-commissioned review for growing the UK's AI industry, which made 18 recommendations over four thematic areas. The thematic area of interest for this thesis is about improving the supply of skills. For this recommendation, the report explains how Professor Sir Adrian Smith of the Alan Turing Institute studied the feasibility of compulsory mathematics study for all pupils to 18 to 'improve the foundations for skills to develop, understand, and work with AI' (Hall & Pesenti, 2017, p. 52). In April 2018, the government and the UK's AI ecosystem responded to Hall's and Pesenti's report with a £1 billion AI Sector Deal to boost the UK's global position as a leader in developing AI technologies. This deal includes a promise to 'invest an additional £406 million in maths, digital and technical education, helping to address the shortage of science, technology, engineering and maths (STEM) skills'[94].

This mathematically focused agenda has continued. In Jan 2023, the UK's Prime Minister, Rishi Sunak, announced plans to move towards some form of compulsory maths education for children up to 18 years old[95]. Professor Mark Girolami, Chief Scientist at The Alan Turing Institute, responded to the announcement by suggesting 'a good understanding of mathematics is important to improve and accelerate progress in the data sciences such as artificial intelligence and machine learning'[96]. Placing mathematics at the centre of a STEM agenda places quantitative methods at the centre of developing the AI industry. Yet, as the following section explains, the findings of this thesis question the relevance of taking a purely mathematical view of developing the AI industry: developing social science applications requires a more sociotechnical approach.

## 5.2 Quantitative Methods Fail to Provide Meaningful Inputs into Explanatory Dialogues about Hate Speech

Chapter 2 addresses RQ2 by assessing the extent to which quantitative methods in NLP 'understand' social science applications like hate speech detection and sentiment analysis. The research question uses the word 'understand' because NLP literature generally claims NLP algorithms can understand natural language. This question also addresses the null hypothesis for

[94] UK Gov (2019) AI Sector Deal, retrieved on 17th Feb 2023

[95] Sunak (2023) PM speech on building a better future: 4 January 2023, retrieved on 17th Feb 2023

[96] Girolami (2030) The Alan Turing Institute responds to the Prime Minister's plans to ensure all pupils in England study mathematics until the age of 18, retrieved on 17th Feb 2023

the research. Addressing the null hypothesis using research question then assesses the commonly used quantitative method of text classification. As is now explained, the findings of Chapter 2 reject this null hypothesis to find that quantitative methods fail to provide meaningful inputs into explanatory dialogues about hate speech.

The chapter begins by introducing a theoretically perfect text classifier that represents a machine understanding of natural language. Introducing this classifier starts with an early text classification paper by Maron (1961) that records a series of statistical methods to classify a text's aboutness according to the occurrence of words. The chapter then explains how contemporary machine learning algorithms similarly use statistical methods to classify a text's meaning, albeit these methods are now significantly more advanced. The algorithm's output is then a function of the input's similarity to a (usually human) annotated dataset. Where the confusion matrix is the current method to assess the performance of a machine learning algorithm, the perfect classifier contains only true positives and true negatives.

The chapter then introduces Hall's theory of encoding and decoding to explain the more subjective element of this perfect classifier for social science applications. Hall's theory explains the orator-audience relationship in the interpretation of messages. Codes themselves are words and phrases that contain both denotative and connotative meanings. Denotative meaning is the generally accepted formal definition of words and phrases; connotative meaning is hidden in subtext and very much depends on the orator-audience relationship. The subjective element is the different interpretative positions an audience may take in response to the same message. In the dominant position, the orator and audience encode and decode a message with the same connotative meaning; in the oppositional position, and audience decodes connotative meaning in opposition to the orator; the negotiated position decodes connotative meaning with a range of interpretative frameworks. Where functional applications are characterised by the encoding and decoding of denotative meaning, the decoding of connotative meaning for different audiences characterises social science applications.

For this theoretically perfect classifier, annotated training data records connotative meaning. Humans generally apply annotations to training data in accordance with an annotation schema. The simplest example of an annotation schema for hate speech is {hateful, non-hateful}, and a simple schema for sentiment analysis is {positive, neutral, negative}. A coding process applies these annotations whereby humans annotate representative text of the domain the classifier seeks to understand. For example, hate speech detection generally uses social media data, and sentiment analysis often uses product reviews. In essence, humans interpret what the training data connotes and apply what they deem the most appropriate label. The algorithm then applies a label based on the similarity of an input to an annotated element of the training data. Interpretations of connotative meaning, therefore, are the degree of similarity between inputs and training data, which are determined by the co-occurrence of words between each.

Beyond theory, the chapter then continues with a practical assessment of how well text classification algorithms encode and decode social science applications. For encoding messages, the chapter finds that established pre-processing practices skew an orator's original message. The practice of stop-word removal removes linguistic information from a text that would otherwise be desirable to capture. For example, pronouns often refer to named entities in a text; pre-processing would not remove references to named entities, so why remove their referent pronouns? Additionally, removing adpositional stop-words – of, against, on – fundamentally alters the representation of prepositional noun phrases and conjunctions. Moreover, tokenisation misrepresents an orator's use of prepositional noun phrases and conjunctions. The resultant noun chunks or a pre-processed NLP object do not denote what orators seek to signify in their texts. Consequently, any application that uses stop word removal or does not account for prepositional noun phrases or conjunctions cannot claim to understand any form of natural language, let alone social science applications.

The chapter continues with an assessment of word embeddings and transformer architectures for encoding natural language. The chapter finds that while the encoding of words using word embeddings is effective for denotative meaning, connotative meaning is encoded with a consensus bias in the training data. For example, the Western consensus contained in the Google News corpus encodes al Qaeda and the Taliban with such negative connotations as 'terrorist'. This connotation, nevertheless, only applies to Western consensus; terrorist propaganda would take the oppositional position and encode al Qaeda and the Taliban with a more positive connotation. While the state-of-the-art has advanced significantly since word2vec, transformer architectures still process text by word co-occurrence, they still require humans to annotate connotative meaning and the training data contain (a most likely) Western consensus bias. As such, word embeddings understand Hall's dominant position where orators and audience share the same interpretation of connotative meaning; word embeddings, however, cannot claim to understand the negotiated or oppositional positions.

The experiments with Detoxify, a transformer-based classifier, and IBM's Watson, a commercial application, find problems with treating text as unstructured data for decoding a text. Treating text as unstructured data refers to processing a text by word co-occurrence rather than grammatical relations. The Detoxify experiment shows that decoding text by word co-occurrence often inappropriately changes the algorithm's outputs. For example, benign noun phrases like 'little black boys' and 'little black girls' inappropriately increase the toxicity score. In effect, the occurrence of the word 'black' increases the toxicity score regardless of context. The experiment also shows how combining a highly antisemitic sentence with a benign sentence reduces the toxicity score, despite how the combined text remains horrifically antisemitic. The extra words drown out the genuinely toxic words. The experiments with IBM Watson suggest these problems also exist in commercially available applications whose marketing material claims to understand natural language. The promise of giving machines the ability to understand humans using NLP does not stand up to the scrutiny of these tests.

From a methodological perspective, the chapter also finds that the treatment of text as unstructured data has not necessarily advanced the state-of-the-art since Maron's original text classification paper in 1961. Where Maron (1961) considers classifying documents by word occurrence, the review of word embeddings shows how the same approach still applies. The underlying premise of word embeddings is the distributional hypothesis whereby a word's meaning is encoded and decoded by word co-occurrence. Positional vectors also encode meaning by the relative position of words in a sentence but not their grammatical structure. Where Maron (1961) used co-occurrence to determine what a text is about, text classification uses co-occurrence to interpret a text's meaning. Therefore, applying co-occurrence to text processing has conflated two separate tasks of determining a text's aboutness and meaning. While the technical sophistication has undoubtedly advanced since Maron (1961), the actual method of processing a text by word co-occurrence remains consistent.

Chapter 2 also uses encoding and decoding to make a similar finding about the contrast between old methods and technical sophistication. Hall (1974) introduced this theory in response to Shannon's and Weaver's mathematical theory of communication that applies an assumption of fixed interpretation of meaning between an orator and their audience. In effect, this assumption assumes an orator and audience share the same interpretative framework of denotative and connotative meaning. Software code is an example of where Shannon's and Weaver's theory still applies. It has a rigid interpretative framework between a developer and a compiler. As explained, the assumption of fixed interpretation to humans only applies in the dominant position where orators and audiences share the same narrow interpretative framework. The oppositional position and any number of negotiated positions challenge the assumption of fixed interpretation. To develop NLP applications that can only 'understand' the dominant position, therefore, re-applies the assumption of fixed interpretation of codes that Hall (1974) has successfully challenged over several decades. The remainder of the research questions responds to RQ1 by exploring how augmenting quantitative methods with qualitative approaches may develop more meaningful ways to tackle hate speech.

## 5.3 Hostile Narrative Analysis Enables Meaningful Explanatory Dialogues About Violence Legitimisation

Chapter 3 responds to the RQ2 by using cultural violence from Peace Research to rethink hate speech detection as hostile narrative analysis. The chapter finds no generally accepted methodology for detecting hate speech due to the polysemy of how different people define hate speech itself. The chapter responds by using Galtung's theories of violence from Peace Research as the methodological basis for hostile narrative analysis. As such, the novel contribution of this chapter is the methodological framework of cultural violence. The chapter verifies this framework for at least Bush's and bin Laden's declarations of war. This methodological framework contributes to a meaningful analysis by enabling explanatory dialogues that promote a human understanding of violence legitimisation in a hostile narrative. Therefore, this thesis uses the methodological framework as a way to accept the research hypothesis by showing how integrating qualitative methods with NLP improves the meaningful analysis of hostile narratives.

The chapter introduces the growing investment in detecting hate speech but then explains how the field is beset with a series of acknowledged theoretical, definitional, and methodological problems. Matsuda (1989) initially conceived hate speech to advocate for criminalising a narrowly defined class of hateful language. She gave hate speech a precise definition that centres on using derogatory language to promote racial inferiority against a historically oppressed group. In advocating for criminalising particularly harmful types of hate speech, Matsuda employed the methodology of outsider jurisprudence. This methodology uses victim stories to explain the harm hate speech causes. As such, this outsider jurisprudence focuses on a victim's experience of harm to explain the need for change. This methodology, however, does not explain why a particular utterance is hateful; outsider jurisprudence only explains the need to criminalise hate speech.

Where hate speech originated in critical legal studies, it has since become a polysemous term through an interdisciplinary transition to the social, political and computer sciences. This polysemy arises from a legal and multiple ordinary meanings. The legal meaning refers to Matsuda's (1989) original definition, while the ordinary meaning largely depends on who uses the term. The progressive ordinary meaning focuses on the harm hateful language causes minority groups and is the focus of this thesis. Conversely, the reactionary ordinary meaning focuses on hate speech as a threat to Free Speech. Indeed, much like hate speech, reactionary politics has rendered Free Speech a polysemous term. This level of polysemy for legally precise terms, and the focus on subjective experiences of harm, means there is no formally accepted methodology to explain why a particular utterance is hateful.

The absence of a generally accepted definition and methodology for detecting hate speech is evident in the computer sciences. Where text classification is the generally accepted computational approach for detecting hatefulness, the previous section explains the shortcomings of processing text as unstructured data. This generally accepted approach, however, is not a formally agreed methodology when there is no agreed definition for what constitutes hate. Text classification is applied as a common method to different domains without consideration of the methodology. Moreover, to only classify hateful utterances does not account for ingroup elevation that the social sciences generally agree is an essential element of racism and other forms of hate. As a general observation, rather than making algorithms to fit the problem of detecting hate speech, developers make the problem fit the algorithm.

In response, the hostile narrative methodology presented in this chapter fills the methodological gap that has been created by the polysemy of hate speech. This methodological framework of cultural violence is how this thesis rethinks hate speech as hostile narrative analysis. To summarise the primary ideas of this methodology, social identity theory explains group formation and sources of intergroup differentiation. Cultural violence explains how an ingroup may seek to legitimise violence against an outgroup. For legitimising violence, religion and ideology provide the cognitive frameworks for elevating an ingroup and othering an outgroup, thereby creating a Self-other gradient between each. A hostile narrative communicates an orator's religious and ideological belief systems for elevation and othering. As such, this methodology applies to explanatory dialogues through the following underpinning hypothesis (reproduced from the introduction):

> *The steeper the Self-other gradient created by ingroup elevation and outgroup othering, the more legitimate violence against an outgroup becomes.*

The methodology complements Matsuda's legal definition of hate speech by distinguishing between unlawful hatefulness and hostility. The provenance for determining the lawfulness of hateful language draws upon historical texts like *Mein Kampf,* while hostility remains lawful, albeit problematic. Regarding hostility, cultural violence's focus on intent also gives this methodology a different epistemological position to hate speech. Where outsider jurisprudence focuses on a victim's experience of harm, hostile narrative analysis focuses on an orator's hostile intention. Following the underpinning hypothesis, the extent of hostility depends on the extent of elevation and othering in an orator's text. The provenance for determining hostility uses historical texts that sought to legitimise violence, such as Bush's and bin Laden's declarations of war. As such, in its legal definition, hate speech is victim-focused, while hostile narrative analysis is orator focused.

The hostile narrative methodology also draws upon the qualitative approaches of narrative and semantic analysis to detect elevation and othering. Narrative analysis applies by rethinking hate speech as a story that people use to legitimise harm against an outgroup. Following the principles of narratology, these stories exist as a narrative truth in the collective imagination of the ingroup who subscribe to them. Following the principles of personal experience narrative analysis, people use such stories to rationalise events. Yet, while the events may be real, they are not necessarily truthfully represented in a narrative. Moreover, a narrative's characters are more caricatures than real. As such, a historic truth gives way to an orator's narrative truth, and the story they portray is a potentially fictional representation of real-words events. The narrative truth of these stories legitimises harm by contrasting the imagined heroism of an orator's ingroup and the villainy of their outgroup, whether perceived or real.

The narrative truth of each hostile narrative reviewed in this thesis contains a similar plot for their violence legitimisation. In *Mein Kampf*, Hitler's Aryans are the racially pure heroes of the German people, whereas Jews are a racial poison. Hitler then lays the foundation for genocide by promoting an antisemitic conspiracy theory that Jews dominate world affairs. In Bush's declaration of war, Americans are the heroic protectors of political ideals, while terrorists are threat to these ideas, thereby legitimising his 'War on Terror'. Muslims are the heroes of bin Laden's Declaration of Jihad who protect the sacristy of Saudi Arabia. Bin Laden othered Americans as infidels to his religious ideals and used their presence in Saudi Arabia as a threat to legitimise his Jihad. In Andrew Leak's and Paul's Dunleany's nationalist narratives, immigrants, homosexuals, transexuals and 'Muslim grooming gangs' threaten the imagined greatness of their home country, Great Britain. Leask used this imagined threat to legitimise the firebombing of an immigration camp and his own suicide. At the highest levels of abstraction, each orator used a real or perceived threat to the imagined greatness of their ingroup to legitimise violence.

At lower levels of abstraction, the characters of each story define the hostile narrative genre. The imagined communities of geopolitics define the ingroup and outgroups of *Mein Kampf* for the antisemitic genre and of each declaration of war for the warfare genre. The narrative truth of each text present states and nations as characters in a story of a violent contest. In the more social justice focussed narratives, orators may frame their ingroup and outgroups by attributes of race for the racism genre, gender or sex for the sexism or transphobic genre, and sexuality for the homophobic genre. Matsuda's (1989) original meaning of hate speech categorises prohibited utterances of a hostile narrative, like characterising a particular group as parasites. As such, the methodology has general applicability to different genres of the progressive meaning of hate speech that this thesis supports.

Another defining feature of each genre is the role of ideology and religion in ingroup elevation and outgroup othering. A prominent feature of elevation and othering in *Mein Kampf*, is a perniciously compelling medical metaphor to create a sense of racial purity and impurity. Bush's War on Terror focuses on the political ideology contained in America's constitution to distinguish between a just ingroup and an unjust outgroup. Nevertheless, Chapter 2 explains how the medical metaphor exists in the broader War on Terror narrative that originated with Bush's texts. Leak's and Dunleany's nationalist narratives also focus on political ideology to create a sense of ingroup supremacy and outgroup inferiority. Bin Laden's Declaration of Jihad focuses on religion to distinguish between piety and impiety. In effect, religion and ideology serve as the cognitive frameworks to create a sense of moral distance between an orator's ingroup and outgroup. Where Chapter 3 provides the basis for qualitative approaches with the methodological framework, the following chapter then extends them by applying semantic analysis to the analysis of hostile narratives.

## 5.4 A Hybrid Approach to NLP Assists with Explaining Why a Narrative is Hostile

Chapter 4 addresses the third research question by using hybrid NLP to apply semantic analysis to the hostile narrative method. The novel contribution of this chapter is the hostile narrative analysis method derived from the methodological framework presented in Chapter 3. The computational methods for this method are a hybrid approach to NLP that augments text classification with pattern-based NLP. In contrast to treating text as unstructured data, pattern-based NLP treats text as structured data, which is consistent with the linguistic theory of semantic analysis from Chomsky (1959) and Tesnière (1959), as well as the narrative analysis theory of Todorov and Weinstein (1969) and Labov and Waletsky (1997). The quantitative element of the hybrid approach uses text classification to label a word's lexical attributes. The qualitative aspect applies semantic analysis to infer meaning from parsing language clauses using a pattern-based NLP. The parsing of language clauses then enables the generation of a rationale to inform explanatory dialogues about hostile narratives.

The chapter begins by developing how Galtung conceives the role of language in violence legitimisation. Much like linguistic theory, he views language as a system of interacting words. His research explains how analysing language reveals a group's 'social cosmology', which refers to how the underlying beliefs of a language community apply to violence legitimisation. The social cosmology of the Nazis in *Mein Kampf* has an inherent belief in the greatness of the Aryan race and the threat Jews, among many others, pose to this greatness. The social cosmology of Bush's declarations of war has a similar belief in American greatness that terrorism threatens. Bin Laden conversely believes in the greatness of Muslims to whom America poses a threat. The social cosmology of Leak's and Dunleany's nationalist narratives believes in the greatness of Great Britain, to which a range of scapegoats pose a threat. This social cosmology very much relates to the plot of each text that the hostile narrative method seeks to reveal.

The chapter then uses semantic analysis to infer meaning from processing text by language clauses. Chomsky (1959) and Tesnière (1959) present two models for modelling language clauses. Chomsky's model is called phrase structure grammar, while Tesnière is called dependency grammar. This thesis chooses Tesnière's model since his subject->predicate->object clausal structure is consistent with narrative analysis practices. This processing of text by clauses presents a further difference between hate speech detection and hostile narrative analysis. Hate speech detection tends to treat sentences in isolation, while hostile narrative analysis seeks to understand the interaction of clauses across a text's sentences.

The chapter then presents a spaCy pipeline component for applying semantic analysis using pattern-based NLP, which shows the value of treating text as structured data. The chapter presents a series of sentences from Bush's declaration of war and develops the corresponding patterns for detecting the constituent clauses. The example sentences from *Mein Kampf* in the previous section of the chapter show the importance of understanding how these clauses interact. These examples explain the interaction of two sentences that define Hitler's Jewish outgroup in contrast to his Aryan ingroup. The treatment of sentences in isolation misses the otherwise benign sentence that promotes an ingroup's greatness as a part of a narrative. Correspondingly, collating relevant clauses across a narrative enables the detection of the Self-other gradient in a text. Therefore, the value of treating text as structured data is in understanding how the interaction of elevation and othering clauses promote hostility.

The third section presents discovery work for each objective of the hostile narrative methodology. The pre-processing steps are required to enable the processing of text by language clauses. These pre-processing steps use text classification to label a word's lexical properties. This discovery work does find a limitation with co-reference resolution whereby co-reference algorithms cannot resolve multiple entities to pluralised nouns, like 'terrorists', and pronouns, like 'they'. The analytical steps in objectives 2-4 present ways to detect and analyse elevation and othering. The section shows how this approach generates a natural language rationale to interpret elevation and othering in a language clause. The development of a scoring schema for the Self-other gradient relies upon the application of knowledge graphs that the further work section proposes.

The difference between quantitative and hybrid-NLP is a shift from probabilistic to more deterministic reasoning. Probabilistic reasoning features strongly in the quantitative methods reviewed in Chapter 2. As this chapter has explained several times, the algorithmic outputs are a function of the similarity between an input and training data; similarity is based on the occurrence of words between each. This use of similarity, however, is inconsistent with how humans interpret language. Conversely, the patterns of pattern-based NLP are about grammatical relations in natural language. This more deterministic approach focuses on how an orator, rather than a community, grammatically links words to communicate meaning. As linguistic theory has explained over several decades, humans use grammatical relations to interpret meaning. The more deterministic approach, therefore, is more consistent with human interpretations of language and the epistemological focus on the perpetrator in cultural violence theory.

A general finding from this chapter and the others in the in adequacy of treating text as a sequence of words rather than a system. To treat a text as a sequence is a re-occurring theme in all current quantitative methods regardless of sophistication. Word embeddings apply the distributional hypothesis to create vector representations from the proximity of words. Positional vectors and bidirectionality does the same, just in a more sophisticated manner. The requirement to iterate over data in software is likely the reason NLP algorithms treat text as an iterable. The relative sophistication of NLP algorithms, however, does not necessarily generate more meaning insights, especially for small data. In the genocidal *Mein Kampf* clause, the closest noun to 'Jew' is 'Nomad' by five words and this relationship is negated, whereas the term the clause's subject and object are separated by nine words. Proximity does not imply meaning. The application of pattern grammar in this chapter has shown how parsing a language as a system, and not a sequence, better represent an orator's beliefs and better explain the implication of their expressed hostility.

This chapter has also shown the value of pattern-based NLP for parsing a small data set. The requirement for big data in labelling the grammatical properties of words. Whereas the patterns enable the analysis of small data to extract insights about the orator's assertions. Of particular interest, parsing *Mein Kampf* for the Detoxify experiment took nearly three hours and required an online connection. The application of this pipeline takes minutes and can be implemented directly on a laptop. Pattern-based NLP, therefore, generates insights that better represent the orator's intention and require much less processing power that large language models.

The result of this chapter points towards an NLP pipeline that creates a structured and searchable database of language clauses from a narrative. Each clause represents a particular assertion either in favour of the orator's ingroup or against an outgroup. Each narrative clause represents a particular event that the orator is responding to with their violence legitimisation. This database represents the social database contained in narratives that reveal the belief systems, or social cosmology, of the language community in question.

## 5.5 Quantitative approaches to Implementing the Online Safety Bill Provide Limited Explanatory Value

These problems with using quantitative methods to understand hate speech apply to the Online Safety Bill through explanatory dialogues about what constitutes harmful content. Without an underpinning sociological theory of sentiment or hate speech from literature or online platforms, the meaning of numerical scores makes no sense. In sentiment analysis, for example, what does the 0.15 difference between 'lies' (-0.81) and 'terror' (-0.96) explain when sentiment does not appear to have a unit of measurement? Moreover, on what authority does 'freedom' (+0.66) score more highly than 'unity' (+0.57) when these terms are often used as political ideas to motivate people, as is the case in Bush's text? Most of all, what does the similarity of co-occurring words between a text and training data explain? Without an underpinning theory, a unit of measurement or an obvious rationale for either sentiment analysis or hate speech detection, these numerical outputs provide little information about a text's hatefulness or sentiment for an explanatory dialogue. Numerical inputs to explanatory dialogues between annotators, users of online platforms and the moderators, or between an online platform and Ofcom, therefore, likely provide limited explanatory value.

Quantitative methods for NLP also lead to the high potential for false positives and negatives in proactively detecting harmful content. From a theoretical perspective, dog whistles and metaphors mean there is no fixed interpretation of what constitutes hateful content. Therefore, the potential for false positives or negatives with dog whistles and metaphors arises from bespoke annotation schemas in a platform's chosen algorithm or the regulator's interpretation of what constitutes harmful words. From a more technical perspective, the experiments show how even the most sophisticated technologies produce counterintuitive outputs. They show how false positives arise when such words as 'black' occur in otherwise benign sentences and how sentiment scores in commercially available algorithms add limited explanatory value. In effect, the algorithmic output and annotation decision are disconnected; the quantified output explains the similarity of an input to an annotated element of the training data, not why an annotator made a particular annotation decision.

This high potential for false positives and negatives with quantitative methods leads to unnecessary costs for both online platforms and regulators. The problem with false positives is the excessive time they take to review in content moderation and the disputes they may create between content moderators and end users; false negatives, however, are hidden. The experiments with *Mein Kampf* show how the number of words in a sentence can drown out horrifically hateful sentiment and, therefore, lead to a false negative. As such, quantitative methods can hide content that humans would interpret as obviously malicious. Where the Online Safety Bill obliges online platforms to detect harmful content proactively, therefore, quantitative methods will miss hateful sentiment in high-word content and create the potential for litigation with OFCOM. As such, purely quantitative methods to tackling online abuse will fail to achieve their specified task of proactively detecting online abuse.

In a move to the integration of more qualitative approaches, the methodological framework of cultural violence applies to the Online Safety Bill through explanatory dialogues about hostile narratives. Annotation, end-user engagement and regulatory scrutiny are examples of such dialogues. In each case, the methodology provides a way to explain how ingroup elevation and outgroup othering contained in a user’s narrative contributes to violence legitimisation. Where hate speech detection generally focuses on the hatefulness of individual messages, this methodology applies to a user’s entire contribution as a narrative. In doing so, online platforms and regulators can account for the interaction of otherwise benign messages, like dog whistles. The provenance for determining the level of hostility in a user’s narrative then draws upon historical texts used in this thesis or any others the platform or regulator deems appropriate.

The contributions of Chapter 4 apply to the Online Safety Bill for how online platforms can choose to proactively detect abusive language. This chapter has already explained the shortcomings of current approaches to processing text as unstructured data. The response is more hybrid approach to detecting abusive language by inferring meaning from the analysis of language clauses. The constituent pattern-based approach is consistent with linguistic and narrative theory, and the corresponding computational methods support the hostile narrative methodology. The resultant rationales have the potential to provide meaningful inputs to explanatory dialogues between moderators, end-users and regulators. The rationales are meaningful because they support the methodology presented in the previous chapter that applies to real-worlds examples of violence legitimisation. Moreover, they provide meaningful evidence for to moderate disputes and to educate people about why a narrative may be hostile.

The findings of this thesis then reveal a question about where the responsibility for tackling online abuse lies. The Big Brother Watch report explains how this responsibility seems to lie with online platforms who have the resources to develop sophisticated algorithms. Yet, the state-of-the-art suggests these algorithms use quantitative methods that have limited explanatory value; Big Brother goes as far as saying they are a blunt instrument. Moreover, the annotation schemas for developing these algorithms are likely based on an online platform's bespoke definition of hate speech rather than an empirically informed methodology as presented in this thesis. Bespoke definitions of hatefulness by an algorithm only likely ferment disagreement between proponents of the progressive meaning or reactionary meaning of hate speech. The responsibility for at least defining a methodology for detecting hate speech, therefore, should lie outside of online platforms and be testable by the concerned community. The methodology presented in this thesis, therefore, modestly removes the responsibility of defining hate speech from online platforms to make a contribution from Web Science for tackling online abuse.

# Chapter 6 Further Work

This thesis contributes hostile narrative analysis as a new field of study for detecting online abuse, thereby present a range of new research opportunities. This section covers continuing to conceptualise a hostile narrative as a story for the methodology and the development of new computational methods by combining semiotics with knowledge graphs. The application of knowledge graphs to the social database enables sophisticated queries about the narrative. This is a large task that requires the integration of look-up tables, like VerbNet, and the possibility of a new one for named concept recognition. The motivation, nevertheless, is the potential to reveal otherwise hidden insights about hostile belief systems in a language community for explanatory dialogues. Each of these insights are then accompanied by a rationale for an explanatory dialogue. The continuing development, therefore, seeks to enable the proactive detection of hostile narratives with an Explainable AI.

The most obvious starting point for further work is the gathering and development of criticisms of hostile narrative analysis. Since cultural violence has not featured in research much beyond Galtung (1989), there is not much critique to offer. There are other possible forms of critique arising from the different ideas and disciplines comprising the hostile narrative. Nevertheless, this thesis combines them in a unique way. Criticism, therefore, will be developed during continued development of both the methodology and method.

The most obvious technical starting point is to continue verifying the use of grammar patterns for the computational methods of hostile narrative analysis. Chapter 4 has to some extent shown the value of these patterns for analysing hostile narrative, but further verification is required. This further verification would comprise manually creating a test dataset of elevation and othering statements from the hostile narrative corpus. This test dataset then provides the means to evaluate how well each of the grammar patterns detect each elevation and othering statement. An assessment of how well the patterns generalise would experiments with elevation and othering statements from other hostile narrative genres. The results of these experiments would then inform how to refine these patterns and inform the limitations of this approach for a rigorous explanatory dialogue. Finally, this verification would also inform the development of knowledge graphs, as explained later in this section. This section now begins with further work to verify the methodology.

## 6.1 Verifying and Developing the Hostile Narrative Methodology

Verifying the methodology begins with a wider analysis of different genres of hostile narratives. While this paper has focused on language clauses, there is a wider potential for the analysis of narrative structures. A modified review of more hostile narrative genres using these structures should reveal more insight into violence legitimisation. For example, Propp (1968) analysed folktales to reveal a finite set of archetypical characters and plot devices. These characters and plot devices endure in modern storytelling such as cinema (Giswandhani, 2022; Saputra & Noverino, 2023). A similar study could be made of hostile narratives to reveal common character roles and plot devices. Jahan *et al*. (2021) attempt an equivalent computational analysis the ProppLearner corpus of using quantitative methods (Jahan et al., 2021). At the highest level of abstraction, the character roles in hostile narratives are friends and enemies, but there may be more interesting subclasses of each role.

While this thesis has focussed on the structure of narrative clauses, Labov and Waletsky (1997) proposed a common schema of five components for personal experience narratives. A preliminary evaluation of the hostile narrative texts suggests some correlation between violence legitimisation and the five components of personal experience narratives.

- **Orientation**: a series of clauses that 'orient the listener in respect to person, place, time and behavioural situation' (Labov & Waletzky, 1997, p. 27). Both Bush's and bin Laden's text orientate the audience towards the greatness of their ingroup through a series of elevation clauses.
- **Complication**: 'the main body of a narrative usually comprises a series of events that may be termed the complication or complicating action (Labov & Waletzky, 1997, p. 27). The primary complicating actions of both Bush's and bin Laden's narratives are the 9/11 attacks and the perceived occupation of Saudi Arabia respectively. These violent actions are then linked to a chain of events. In cases of structural violence, individual actions in isolation may be benign, but in aggregation they amount to violence.
- **Evaluation**: 'that part of the narrative that reveals the attitude of the narrator towards the narrative by emphasising the relative importance of some narrative units compared to others.' (Labov & Waletzky, 1997, p. 32). As such, an orator might portray a sense of ingroup victimhood in response to a violent action and an othering of the outgroup perpetrators. Evaluation, therefore, is an essential element of elevation an othering.
- **Resolution**: 'the resolution of the narrative is that portion of the narrative sequence that follows the evaluation' (Labov & Waletzky, 1997, p. 35). The resolution of the hostile narratives might be the violent action they seek to legitimise. For example, in in Bush uses

the threat of war to compel the Taliban to fulfil a series of demands. Bin Laden similarly compels America to leave Saudi Arabia. The ultimate resolution for each orator was the War on Terror and Jihad respectively.

- **Coda**: 'the coda is a functional device for returning the verbal perspective to the present moment' (Labov & Waletzky, 1997, p. 37). Examples of coda clause in the hostile narratives are not immediately obvious probably because they lead up to the resolution. Examples of coda clauses probably appear after the conflict is complete.

This narrative structure and the possible development of an updated version for hostile narratives can be verified with a coding review of a text that have been used in violence legitimisation. Such a coding review would either verify Labov's and Waletzky's schema, or an updated version will emerge. As with the case of personal experience narrative analysis, particular narrative clauses would identify the components of a hostile narrative schema. The ability to apply this schema to a range of texts then allows content moderators to connect narrative elements across user accounts and to historical texts. For examples, in a nationalist narrative, orientation clauses elevate the greatness of the orator's own country. In isolation these are benign; but when connected to evaluation statements about victimhood and the villainy of an outgroup their more pernicious nationalist intent becomes clear. A schema for the different components of a hostile narrative, therefore, has the potential to assist with explaining the hostility of that narrative. A more develop methodology then leads to further work in developing the method.

## 6.2 Towards a Production Ready Hostile Narrative Analysis NLP Pipeline

Chapter 4 demonstrates the value of treating text as structured data with example code for each objective. This section now proposes further work to create a hostile narrative pipeline using knowledge graphs. In accordance with the further work to develop the methodology above, the primary feature of a hostile narrative pipeline is to link interacting clauses in a text. As already stated in this thesis, text classification methods inappropriately tend to treat sentences in isolation. As is now explained, this section proposes the use of knowledge graphs to connect these clauses. As is also explained, knowledge graphs may also provide a way for NLP to interpret the connotative and denotative meaning of these clauses. This potential for interpreting meaning is derived from applying a further qualitative method of Semiotics. This section, therefore, introduces Semiotics and knowledge graphs as further work for developing a hostile narrative analysis NLP pipeline.

### 6.2.1 Reintroducing Semiotics

This thesis has already introduced Semiotics twice; as a reminder, semiotics is the study of signs, symbols, and communication systems, including the interpretation and use of meaning in different contexts. Semioticians are concerned with how meaning is created, conveyed, and understood in various forms of representation, including language, visual images, and cultural practices. Semiotics seeks to understand how meaning is constructed and used to convey messages and ideas, and how these meanings shape and are shaped by cultural and social norms. Semioticians study how meanings are made and how reality itself is represented (and indeed constructed) through signs and sign systems. Semantics is a sub-class of Semiotics in that is treats words as signs.

Semiosis is a process of meaning-making in which a sign, such as a word, image, or gesture, is interpreted in context to create a message (Eco, 1976a, p. 315). In semiotics, semiosis refers to the relationship between a sign and its meaning, as well as the numerous ways in which this relationship is established and understood. Semiosis, therefore, is concerned with processes that produce and interpret signs. Two models by Saussure and Pierce provide a way to understand this semiotic process of meaning making.

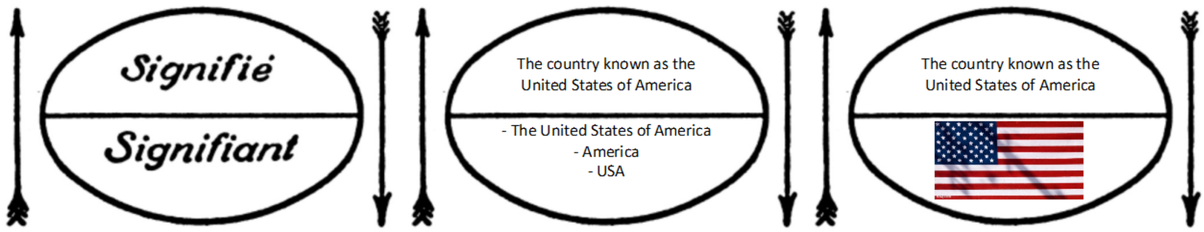



Figure 54. Saussure's model of the sign.

In his 1916 Course in General Linguistics, Swiss linguist Ferdinand de Saussure modelled a sign by two components he refers to as signifier (significant) and signified (signifié) (Saussure, 1916). While not necessarily cited in NLP literature, Figure 54 shows how this model is used in such tasks as named entity recognition. In his model, the signifier is a word or symbol that represents a signified concept or real-world entity. Figure 54 then shows how 'The United States of America', 'America' and 'USA' are three noun phrases to represent the country known as the United States of America. Recalling the named entity recognition and resolution task from Chapter 4, the same knowledge base identifiers denote how each of these noun phrases signify the same named entity. Similarly, the national flag of the United States, variously signified by the noun phrases, 'the Stars and Stripes' or the 'Star Spangled Banner' also signifies the country known as the United States of America. While the relationship between the signifier and signified are arbitrary, the idea they are inseparable is explained by Saussure's diagram, which shows how the sign denotes the object it represents.

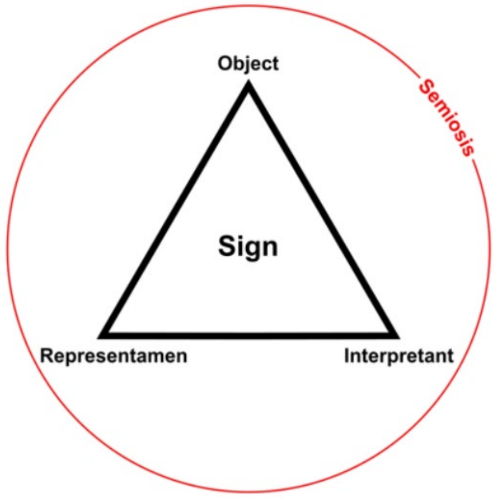



Figure 55. Peirce's model of a sign.

While Saussure's model enables the interpretation of a sign's denotative meaning, a second model by Peirce incorporates connotative meaning (X. Zhang & Sheng, 2017). Peirce's model, shown in Figure 55, comprises the representamen, the object and the interpretant. While not equivalent, the representamen and interpretant are analogous to the Saussure's signifier and signified respectively. There are also broader differences on the psychology of how these signs are constructed, but the raw models are dealt with here. While the representamen denotes the object, an interpretant is a sign's connotation, implication, or ramification. According to this model, Semiosis encodes both connotative and denotative meaning by combining the representamen, the object and interpretant into a single sign.

To explain how Peirce's model incorporates connotative meaning in language, consider the modelling of the following two sentences from Bush and bin Laden respectively.

> *Sentence 1: …{Americans}* $_{subject}$ *{are}* $_{predicate}$ *{generous}* $_{conjunction1}$ *and {kind}* $_{conjunction2}$*, {resourceful}* $_{conjunction3}$ *and {brave}* $_{conjunction4}$*.*

> *Sentence 2: {The American people}* $_{subject}$ *{are}* $_{predicate}$ *{the ones}* $_{object}$ *{who pay the taxes which fund}* $_{clause\ modifier}$ *{the planes that bomb us in Afghanistan}* $_{conjunction1}$*, {the tanks that strike and destroy our homes in Palestine}* $_{conjunction2}$*, {the armies which occupy our lands in the Arabian Gulf}* $_{conjunction3}$*, and {the feats which ensure the blockade of Iraq}* $_{conjunction4}$*.*

Each sentence is annotated according by their constituent clauses. Both subjects of each sentence, whether 'Americans' or 'American people', are representamen of the common named entity[97] of ‘American citizens’. The common predicate ‘are’ then asserts a series of contrasting noun phrase attributes to the same named entity. In sentence one, Bush asserts that Americans have the positive attributes of generosity, kindness, resourcefulness, and bravery. Sentence two from bin Laden, however, is an oppositional statement to Bush. In his more complicated sentence, bin Laden asserts a negative behaviour onto the American people. He accuses them of paying a tax that funds direct violence in Afghanistan, Palestine, the Arabian Gulf, and Iraq. In fact, ‘taxes’ is a second representamen to the named entity of ‘American taxes’ with the interpretant of funding direct violence. For the benefit of this explanation, nonetheless, this second sign for taxes is collapsed into ‘American citizens’.

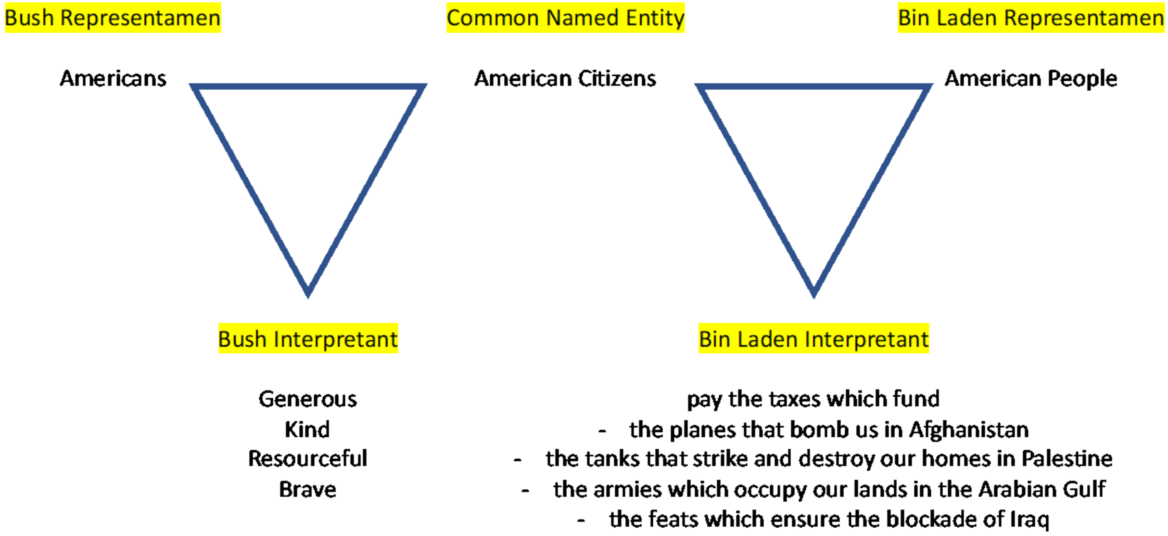



Figure 56. Applying Peirce’s sign model to sentence from Bush and bin Laden.

Figure 56 shows how semiosis encodes connotative meaning by combining the representamen, named entity and interpretant. Each statement is how each orator contests the prototypical American. Bush’s uses the word ‘Americans’ as a representamen to denote ‘American citizens, while bin Laden’s uses’ American people’. The interpretant, shows the connotative meaning each orator seeks to encode about American citizens. Bush’s combination of named entity, representamen and interpretant encodes a sign whereby American citizens connote positive attributes, while bin Laden’s sign for the same named entity connotes negative attributes. The positivity or negativity of each interpretant, nevertheless, is contingent on the audience’s position relative to each orator. A pro-American audience would likely decode Bush’s sign with the same interpretative framework, whereas an anti-American audience, such as al Qaeda, would decode Bush’s sign with the frame of reference asserted by bin Laden. Equally, a pro-War on Terror

---

[97] Note, the term, ‘named entity’ is used in place of Peirce’s use of ‘object’ to avoid confusion with the use of ‘object’ in a language clause.

audience may interpret bin Laden's sign positively, and vice versa for an anti-war audience. This semiotic model now provides a framework for developing a representative knowledge graph.

### 6.2.2 What is a Knowledge Graph?

The actual definition of a knowledge graph appears to be a contentious issue with no generally accepted version. This description, therefore, provides an overview from Hogan *et al.* (2022) who are a group of eminent knowledge graph experts. Hogan *et al*. (2022) define a knowledge graph as 'a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities' (Hogan et al., 2022, p. 2). They further define knowledge as *explicit knowledge*, which refers to something that is known and can be recorded in some way.

A relevant feature of knowledge graphs to semiotics is the application of entailment regimes. Entailment is a fundamental concept in logic, which describes the relationship between statements that hold true when one statement logically follows from one or more statements (Hogan et al., 2022, p. 13). Entailment regimes in a knowledge graph enable machines to reason over these statements. Deductive reasoning is about extracting explicit knowledge from a graph, whereas inductive reasoning employs a series of logic formalism to reveal hidden insight. As is now explained, entailment over a knowledge graph has the potential to mimic semiosis to draw out elevation and othering in a text.

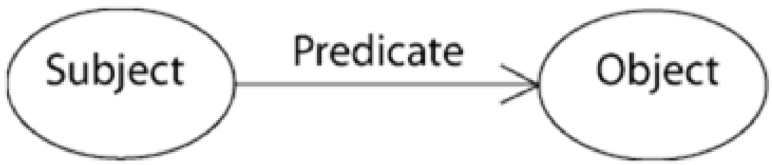



Figure 57. The RDF triple.

The basic graph structure uses the subject->predicate->object model, as shown in Figure 57. Computationally, W3C documentation refers to the model as a Resource Description Framework (RDF) triple whereby, 'some relationship, indicated by the predicate, holds between the things denoted by subject and object of the triple' (W3C, 2014). Referring back to the definition in Hogan et al. (2022), the subject and object are the nodes, and the link is the predicate. Each node then represents a real-world entity or concept, and the predicate represents the relationship between them. The logical formalisms for reasoning over a knowledge graph are expressed as triples and are constructed in an ontology. RDF is the syntax for formulating these ontologies. As will be explained, further work will explore how knowledge graphs may link different narrative elements to interpret the assertions contained in each clause. Further work will also explore the role of knowledge graphs in developing named concept recognition using Peirce's semiotic model.

### 6.2.3 How do Knowledge Graphs Apply to Natural Language?

Knowledge graphs apply to narrative analysis by modelling language clauses as an RDF triple. Accordingly, these triples are the statements over which an entailment regime will enable machine reasoning to parse a text. In applying RDF triples to language clauses, there is a need to differentiate between different uses of the words 'subject', 'predicate' and 'object' by linguists and W3C. Nomenclature for a modelling language clause and triple use the words 'subject', 'predicate' and 'object'. Despite using the same nomenclature, however, this subsection explains how the predicate link in a triple does not always map to the predicate of a language clause. Accordingly, there is a need to distinguish between a verb predicate and a triple predicate.

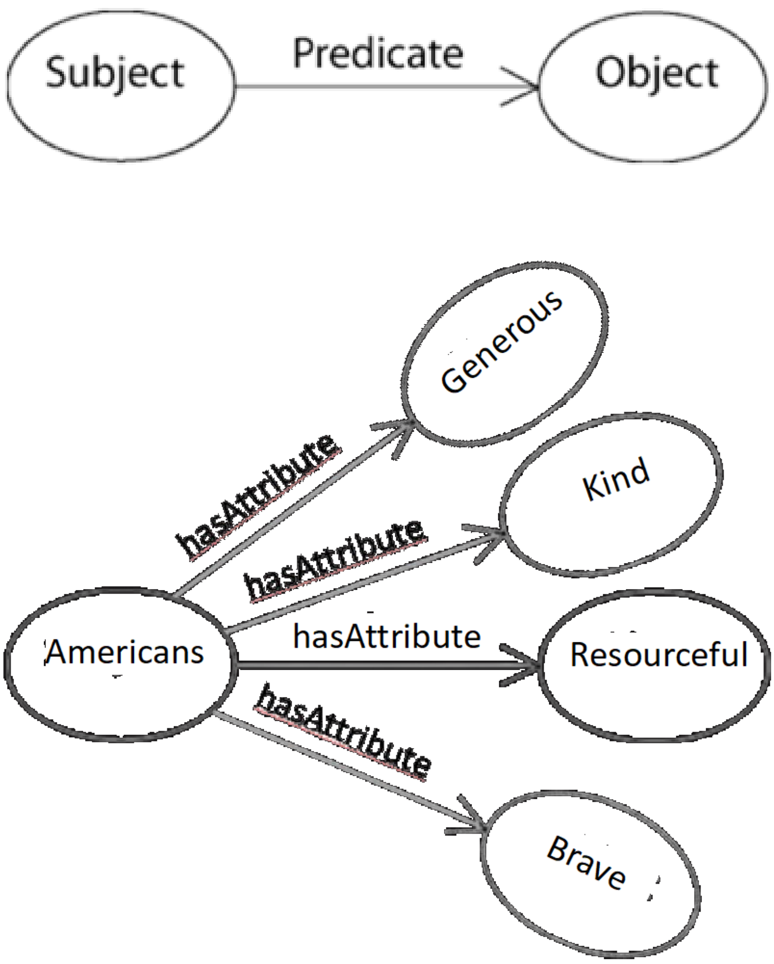



Figure 58. Mapping auxiliary verb clause to an RDF triple.

Figure 58 shows initial research into how RDF triples map to language clauses. The figure shows a mapping of the RDF triple to the clause 'American's are kind, resourceful and brave. In this clause, the verb 'are' is an auxiliary verb. In this clause the subject is the named entity of 'Americans', while the object is a conjunction of four adjectives, ‘generous’, ‘kind’, ‘resourceful’ and ‘brave’. The clause’s predicate is the auxiliary verb ‘are’, other such auxiliary verbs are ‘do’ and ‘have’[98]. The clauses assertion is to assign the adjectival attributes to the named entity of Americans. As such, the verb predicate, ‘are’, directly maps to the triple predicate link. Nevertheless, this direct mapping does not follow for other verb types.

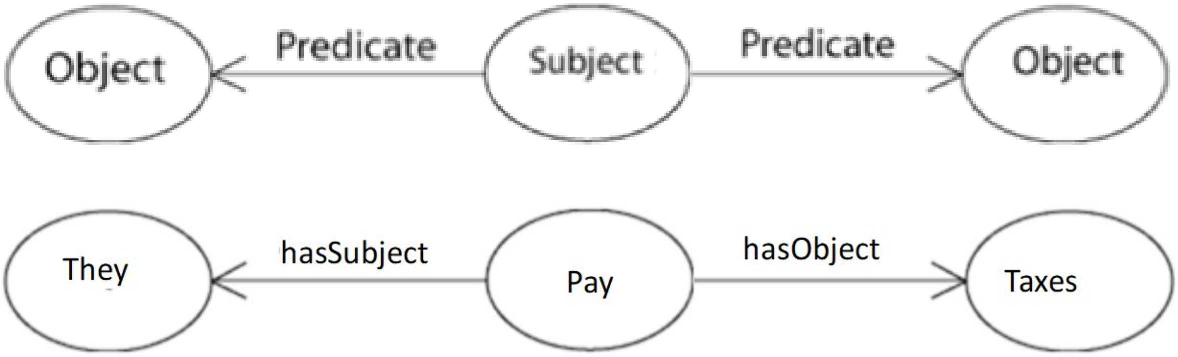



Figure 59. Mapping a verb clause to an RDF triple.

Figure 59 shows a more general verb construction for the clause, ‘they pay taxes’. In contrast to Figure 58, the verb predicate and triple predicate do not directly map. The actual clause comprises two triples that each link the subject and object to the verb. The triple predicate contains the semantic roles of the subject and object to the verb. In the first triple, the triple predicate contains the semantic role of ‘hasSubject’ between ‘they’ and ‘pay; the triple predicate of the second triple contains the semantic role of ‘hasObject’ between ‘pay’ and ‘taxes’. There is, therefore, the potential for confusion when the using the term ‘predicate’ with triple and language clauses. For non-auxiliary verbs, the clause predicate maps to the node of a triple.

Enabling entailment requires the connection of clauses with a common subject. As such, the clause, ‘Americans are generous and kind, resourceful and brave’ also links to ‘Americans showed a deep commitment to one another and an abiding love for our country’. Both clauses have the same subject, ‘Americans’. In doing so, the assertions Bush makes about his ingroup in these clauses are assigned to the same named entity. These assertions then contribute to the Self-other gradient score. Similarly, recall the clauses about Bush outgroups from section 2 in Chapter 4. The linking of these clauses asserts that each organisation is a hyponym of the concept ‘terrorist’. These clauses then assign negative attributes to the concept of terrorist. In the linguistic domain, semiosis explains how Americans inherit the positive attributes asserted by each clause. Equally, semiosis explains how each terrorist organisation inherits the negative attributes asserted by Bush. Reasoning over a knowledge graph can entail the same attributes to each named entity.

---

[98] Grammarly (2023) Auxiliary Verbs, retrieved on 16 Feb 2023

Moreover, entailment across a narrative collates the orator's assertions about their ingroup and outgroup to score the Self-other gradient. Entailment then enables the generation of a scoring schema for the Self-other gradient by collating the collective assertions an orator makes about their ingroup and outgroup.

The second role of knowledge graphs is the development of named concept recognition. Section 3 of Chapter 4 introduced named concept recognition for interpreting connotative meaning. Presently, the research behind this thesis has developed a simple schema using a JSON look up object. Returning to Figure 56, this schema would interpret, 'generous', 'kind', 'resourceful' and 'brave' as positive attributes. Similarly, the schema would negatively interpret 'bomb', 'destroy' and 'occupy'[99]. This schema is based on modelling groups using Martin's (2018) perspective from evolutionary psychology. Further work would explore whether the application of knowledge graphs to this schema would increase the level of insight into elevation and othering in a hostile narrative. This insight would give machines an ability to reason over the connotative and denotative meaning of a text.

### 6.2.4 How Do Ontologies Apply to Hostile Narrative Analysis?

Ontologies have the potential to create representations of hostile narrative to further develop upon the computational methods of hostile narrative analysis. In the context of knowledge graphs, an ontology refers to a formal representation of the concepts, relationships, and properties within a specified domain. 'Ontology' itself is 'the term used to refer to the shared understanding of some domain of interest' (Uschold & Gruninger, 1996, p. 96). Such a shared understanding requires commonly accepted ways to conceptualise the domain of interest. The conceptualisation of the domain is formalised in a generally accepted ontology to describe it constituent elements. These formal representations represent a community standard for the domain in question. Accordingly, the Linked Open Data (LOD) Cloud is one example of an open-source resource that ontologies for a variety of domains[100]. Ontologies, therefore, provide a structured and standardised way to define and organise knowledge within a knowledge graph.

An ontology typically consists of a set of classes, properties, and relationships that define the entities and their attributes in a particular domain. Classes represent concepts or categories of entities, while properties define the attributes or relationships between entities. Relationships

---

[99] Anning (2023) Named Concept Recognition, retrieved on 17th Feb 2023

[100] LOD Cloud (n.d.) The Linked Open Data Cloud, retrieved on 5th July 2023

specify how entities are connected or related to each other. The classes of a narrative are various concepts that words represent. This relationship between words and concepts is explained the Semiotic models presented in 6.2.1. Word properties are the grammatical properties, such as POS tags and dependency labels. The concept attributes are those which are expressed in a language clause. For example, the properties of 'Americans' in Figure 58 are 'generous', 'kind, 'resourceful', and 'brave'. The relationship between concepts are the underlying semantic relations are expressed by dependency labels. Accordingly, the necessary components for a hostile narrative ontology already feature in NLP technologies.

Meghini et al (2021) have begun conceptualising the narrative domain using the Narrative Ontology (NOnt). The domain of concern for their ontology is digital libraries in the cultural heritage domain. At the time of publishing, Mingei European project is validating NOnt for representing knowledge about Craft Heritage[101]. Of particular interest for developing an ontology for hostile narratives is the methodology used by Meghini et al. (2021). As shown here, the chapters of this thesis broadly follow this methodology (Meghini et al., 2021, p. 3). For the explanation of this methodology, each step is highlighted in bold, and the paragraph text is how this thesis relates to these steps.

**Creation of a conceptualisation of the domain, in which the issue is described and analysed in its main parts**. Both Meghini et al (2021) and this thesis define narratives as stories that seek to rationalise events (Meghini et al., 2021, p. 5). Chapter 3 then conceptualises the problem of analysing hostile narratives. This conceptualisation responds to the unhelpful polysemy of the term hate speech, and the absence of a generally agreed methodology for hate speech detection. The conceptualisation of hostile narratives is captured in the proposed methodological framework that centres on the detection of the self-other gradient in response to expressions of elevation and othering in text. The analysis of Bush's and bin Laden then validates this framework at least for the warfare genre of hostile narratives.

**Development of an ontology as the specification of the conceptualisation in terms of a logical theory**. Chapter 4 begins the development of an ontology by using grammar patterns to express the relationship between the concepts represented as words. Underpinning these grammar patterns are the POS and dependency labels of words. The clause itself is a logical expression that

[101] Mingei (2023) DigiTraining: Mingei Online Platform supports in representation of cultural heritage, retrieved on 5th July 2023

makes some sort of assertion, for example, the clause ‘American’s are generous, kind, resourceful and brave’ is a logical expression that asserts positive properties onto Americans. The coding schema introduced in section 4.3.1.4 for named concept recognition also forms part of this proposed ontology. For example, this schema asserts that ‘generous’, ‘kind’, ‘resourceful’ and ‘brave’ are positive attributes. This section deconflicts a clause predicate and a triple predicate for mapping clauses to the logical expressions of an ontology.

**Development of an inference engine for reasoning on knowledge bases conforming to the ontology**. The inference engine would be developed to support each objective of the method proposed in Chapter 4. For example, any assertion of ‘generous’, ‘kind’, ‘resourceful’ and ‘brave’ to named entity is an elevation statement for objective two of this method. It is developing the inference engine that the advantage of using ontologies over patterns become apparent. For example, the inference engine could detect elevation statements with a query that extracts named entities that are assigned with positive attributes. The improvement for the computational methods is in giving machines the ability to reason over a hostile narrative ontology to reveal new insights for new objectives of the method. Having a narrative expressed as an ontology enables the extendibility of the method. Developing queries for the ontology are less laborious that the development of language patterns and machine reasoning introduces the potential for some degree of automation to gain otherwise hidden insights.

**Implementation and evaluation of the ontology and of the inference engine, using Semantic Web technologies**. The implementation and evaluation of the ontology follow similar steps to the verification steps discussed in the discussion section of Chapter 4 and introduction to this chapter. There will also be similar limitations; in particular, representing text in an ontology assumes well written English for the consistent application of representing clauses as triples. This assumption, however, may not always apply; less well written English require bespoke formalisms.

The result of representing a hostile narrative as an ontology returns to Van Dijk’s (1983) idea of narratives as a social database and Galtung’s idea of social cosmology previously discussed in section 4.4. The discussion explains how Figure 49 represents a social database of Bush’s narrative and how each row of this database asserts a particular belief of Bush’s social cosmology. Using an ontology as an underpinning technology for the database in Figure 49 enables a more powerful way to represent and query hostile narratives. An ontology connects concepts across the rows and enables easier ways to query the database than the patterns presented in Chapter 4. An ontology for hostile narratives, therefore, enables the creation of queries for new objectives for the method presented in Chapter 4, and new insights into how hostile narratives feature in violence legitimisation.

# Chapter 7 Conclusion

The primary contribution of this thesis is hostile narrative analysis as a new way to tackle online abuse. Motivated to tackle online hostility, this contribution begins with the relatively new field of hate speech detection. Nevertheless, the technical development of hate speech detection is disconnected from an established methodology for detecting hate. As such, this disconnection questions the explanatory relevance of hate speech detection systems to empirical observations of hate. In response, hostile narrative analysis is founded on a novel methodology and method to connect theory and technical development. The methodology is derived from theories of intergroup relations and violence legitimisation and the method is consistent with established linguistic theory for interpreting meaning. This methodology and method have been verified for at least Bush’s and bin Laden’s declarations of war. With further verification, therefore, hostile narrative analysis provides improved explanatory value to empirical observations of hostility that current hate speech detection algorithms.

Where the primary contribution of this thesis, therefore, is hostile narrative analysis, the primary finding is the need to treat the detection of online abuse as a sociotechnical undertaking. To treat the detection of online abuse as a sociotechnical undertaking means applying social scientific theory to computational methods. As a sociotechnical thesis, Chapter 2 focuses on the more technical problems with hate speech detection, while Chapter 3 focuses on the more social elements. The hostile narrative methodology for detecting online abuse is grounded in the social scientific theory of intergroup relations from Tajfel and Turner along with Galtung’s cultural violence from Peace Research. The resultant computational methods are derived from technically applying these theories. This sociotechnical interaction of theory and computational methods, therefore, seeks to develop algorithms that are relevant to empirical observations of hostility.

Contrary to the UK’s AI growth strategy, these findings suggest a sole focus on growing mathematical competence will not develop operationally suitable NLP algorithms for social science applications. The overall problem of applying solely quantitative methods is to treat the detection of online abuse as a technical problem. The underlying mathematical methods of NLP are undoubtedly sophisticated, however, the high potential for false positives and negatives in the detection of online abuse suggests they fail to understand natural language adequately for meaningful explanatory dialogues. In particular, the experiments show that even the most sophisticated algorithms fail to distinguish between the opposing sentiments of *Mein Kampf* and *I Have a Dream* or the opposite positions of Bush’s and bin Laden’s declarations of war.

Distinguishing between these texts for a human is a simple task. Quantitative methods in NLP for social science applications, therefore, give only the illusion of understanding.

To treat the development of social science applications in NLP as a sociotechnical undertaking presents new collaborative opportunities to grow the UK's AI sector. A re-occurring theme in this thesis is the disconnection between long established social scientific methods and computer science. These methods have been tested an improved over decades to verify their explanatory value. The resultant sociotechnical approach to developing relevant methodologies and methods by connecting social scientific theories with computational methods should have applicability to developing other social science applications. Developing the methodologies and methods for these applications creates opportunities for collaboration between members of the social and computer sciences. While the AI growth strategy places emphasis on STEM subjects, sociotechnical development adds the arts to create the STEAM acronym that combines science, technology, engineering, the arts and maths. Perhaps the emphasis on compulsory maths education might be better placed on transdisciplinarity to promote AI in the UK.

In addition to developing upon hybrid NLP, this thesis also exposes a design choice in the develop of NLP systems between purely quantitative methods and augmenting them with look-up tables. As the current state-of-the-art, developers build large language models using quantitative methods and human annotation. Yet, these methods require a high sample of data for each label in a dataset, as a guide, spaCy advises around 1000 sentences per label. Yet the problem arises when something in the real-world changes, for example, the Taliban no longer being a proscribed terrorist organisation to become a government of state. In the spaCy labelling schema, they have changed from 'ORG' (organisation) to (GPE) geopolitical entity. Accordingly, this change requires a new annotation task. Look-up tables, however, can make this change instantaneously, but they require regular upkeep to maintain currency. The choice to use look up tables will very much depend on the use case and the domain in question.

This thesis makes another general finding about the necessity to conduct due diligence on terminology to enable effective transdisciplinarity research. This point might seem obvious, but this thesis has shown how words take on new meanings when transferred between disciplines. As explained in Chapter 2, the meaning of the term 'code' is specific to either the linguistic or computer science disciplines. As explained in Chapter 3, hate speech has lost its original meaning to become polysemous term through an interdisciplinary transition. Hate speech has a legal meaning in critical legal studies along with a progressive and reactionary meaning in the social and political science. The meaning of hate speech in computer science very much depends on the bespoke interpretations by developers. The mapping of language clauses to RDF triples also shows how the same word can have a different meaning between disciplines. Due diligence, therefore, enables effective collaboration by promoting common and precise definitions of terms.

The continued development of the hostile narrative analysis NLP pipeline places a synthesises of social scientific theory and technical aspects of NLP at the core of its sociotechnical approach. These challenges seem to arise from taking a ‘technical first’ approach before applying social theory. The continued development, therefore, begins with social theory to guide technical development. This approach is matched with a guiding philosophy beginning with first principles before employing advanced technologies. Additionally, the methodology, data sources, group schema and experiments are available in open source to enable reproducible research. This open approach seeks to provide a strong basis for social and technical communities to collaborate and develop more meaningful technologies.

# Bibliography

Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic Hate Speech Detection using Machine Learning: A Comparative Study. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491. www.ijacsa.thesai.org

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agarwal, S., & Chowdary, C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications*, *185*, 115632. https://doi.org/10.1016/J.ESWA.2021.115632

Akinsanya, A., & Bach, C. (2014). Narrative analysis: The Personal Experience Narrative approach. *ASEE 2014 Zone I Conference*, 159–168. https://www.asee.org/documents/zones/zone1/2014/Student/PDFs/21.pdf

Allwood, C. M., & Allwood, C. M. (2012). The distinction between qualitative and quantitative research methods is problematic. *Qual Quant*, *46*, 1417–1429. https://doi.org/10.1007/s11135-011-9455-8

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2018). The Enemy Among Us': Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web*, *13*(3), 39.

Alrehili, A. (2019). Automatic Hate Speech Detection on Social Media: A Brief Survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 1–6. https://doi.org/10.1109/AICCSA47632.2019.9035228

Anderson, B. (2006). *Imagined communities: reflections on the origin and spread of nationalism*. http://rebels-library.org/files/imagined_communities.pdf

Atteveldt, W. van, Velden, M. A. C. G. van der, & Boukes, M. (2021). The Validity of Sentiment Analysis:Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 1–20. https://doi.org/10.1080/19312458.2020.1869198

Augoustinos, Martha., & Reynolds, K. J. (2001). *Understanding prejudice, racism, and social conflict*. 362.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005

Benaquisto, L. (2008). Thematic Coding and Analysis. In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods: Vol. Volumes 1 & 2* (pp. 85–88). Sage. https://publish.uwo.ca/~pakvis/Codes_and_Coding.pdf

Bengfort, B., Bilbro, Rebecca., & Ojeda, Tony. (2018). *Applied text analysis with Python*. O'Reilly Media, Inc. https://books.google.co.uk/books?id=lK1fDwAAQBAJ&source=gbs_navlinks_s

Beniwal, R., & Maurya, A. (2021). Toxic Comment Classification Using Hybrid Deep Learning Model. *Sustainable Communication Networks and Application*, 461–473. https://doi.org/10.1007/978-981-15-8677-4_38

Bird, Steven., Klein, Ewan., & Loper, Edward. (2009). *Natural Language Processing with Python*. O'Reilly. https://books.google.co.uk/books/about/Natural_Language_Processing_with_Python.html?id=KGIbfiiP1i4C&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker?  Word Embeddings. *Advances in Neural Information Processing Systems*, *29*. http://cs.cmu.edu/directory/csd.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bresnan, J. (ed. ). (1982). *The Mental Representation of Grammatical Relations*. MIT Press.

Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Burnap, P., & Williams, M. L. (2015a). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2015b). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burton-Roberts, N. (2016). Analysing Sentences: An Introduction to English Syntax. In *Analysing Sentences: An Introduction to English Syntax: Fourth Edition* (4th ed.). Routledge. https://doi.org/10.4324/9781315646046

Cao, R., Lee, R. K. W., & Hoang, T. A. (2020). DeepHate: Hate Speech Detection via Multi-Faceted Text Representations. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 11–20. https://doi.org/10.1145/3394231.3397890

Capelos, T., & Katsanidou, A. (2018). Reactionary Politics: Explaining the Psychological Roots of Anti Preferences in European Integration and Immigration Debates. *Political Psychology*, *39*(6), 1271–1288. https://doi.org/10.1111/pops.12540

Carter, R., Collins, M., Davis, G., Hermansson, P., Lawrence, D., Mulhall, J., Murdoch, S., & Velleman, L. (2020). *State of Hate 2020*. https://www.hopenothate.org.uk/wp-content/uploads/2020/02/state-of-hate-2020-final.pdf

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., & Lorés, J. (2004). Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In Bussler C.J., Davies J., Fensel D., & Studer R. (Eds.), *The Semantic Web: Research and Applications. ESWS 2004. Lecture Notes in Computer Science,* (Vol. 3053, pp. 445–458). Springer, Berlin, Heidelberg. http://seweb.ii.uam.es/neptuno/

Chandler, D. (2005). Semiotics for Beginners. In *Retrieved March* (Vol. 2, Issue 1969). https://doi.org/10.1519/JSC.0b013e3181e7ff75

Chandler, D. (2017). *Semiotics: The Basics* (3rd ed., Issue 1969). https://www.routledge.com/Semiotics-The-Basics/Chandler/p/book/9781138232938

Chandra, M., Pailla, D., Bhatia, H., Sanchawala, A., Gupta, M., Shrivastava, M., & Kumaraguru, P. (2021, June). “Subverting the Jewtocracy’’’: Online Antisemitism Detection Using Multimodal Deep Learning.” *WebSci ’21*. https://doi.org/10.1145/3447535.3462502

Chase, S. E. (2011). Narrative Inquiry: Still a Field in the Making. In N. K. Denzin & Y. S. Loncoln (Eds.), *The Sage Handbook of Qualitative Research* (4th ed., pp. 421–435). Sage. https://www.google.co.uk/books/edition/The_SAGE_Handbook_of_Qualitative_Researc/AIR pMHgBYqIC?hl=en&gbpv=1&dq=The+Sage+handbook+of+qualitative+research+by+Denzin,+Norman+K.,,+Lincoln,+Yvonna+S&pg=PP1&printsec=frontcover

Chierchia, G., & McConnell-Ginet, S. (2000a). *Meaning and Grammar: An Introduction to Semantics* (2nd ed.). MIT Press. https://mitpress.mit.edu/books/meaning-and-grammar-second-edition

Chierchia, G., & McConnell-Ginet, S. (2000b). *Meaning and Grammar: An Introduction to Semantics.* (2nd ed.). MIT Press. https://doi.org/10.1016/0004-3702(93)90024-6

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, *14*(1), 322–352. https://doi.org/10.1007/s12559-021-09862-5

Chomsky, Noam. (1957). Syntactic Structures. *Language*, *33*(3), 375. https://doi.org/10.2307/411160

Coates, J. (1995). *Armed and dangerous: the rise of the survivalist right*. Hill and Wang.

Collier, P., & Hoeffler, A. (1998). On Economic Causes of Civil War. *Oxford Economic Papers*, *50*, 563–573. https://www.jstor.org/stable/3488674

Collier, P., & Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, *56*, 563–595. https://doi.org/10.1093/oep/gpf064

Collier, P., Hoeffler, A., & Rohner, D. (2009). Beyond Greed and Grievance: Feasibility and Civil War. *Oxford Economic Papers*, *61*(1), 1–27. https://doi.org/10.1093/oep/gpn029

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. https://www.perspectiveapi.com

D.C, A., S.O, A., E, O. I., & F.A, S. (2021). Text Classification Using Hybrid Machine Learning Algorithms on Big Data. In *International Journal of Trend in Research and Development* (Vol. 6, Issue 5). www.ijtrd.com

Delgado, R. (1993). The Inward Turn in Outsider Jurisprudence. *William & Mary Law Review*, *34*, 741–768.

Delgado, R., & Stefancic, J. (2017). *Critical Race Theory* (3rd ed.). University Press.

Denzin, N. K., & Lincoln, Y. S. (2018). Introduction: The Discipline and Practice of Qualitative Research. In *The Sage Handbook of Qualitative Research*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Arxiv.Org*.

Dijk, T. A. van. (1998). *Ideology: A Multidisciplinary Approach*. Sage.

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. http://amueller.github.

Dowlagar, S., & Mamidi, R. (2021). *CMSAOne@Dravidian-CodeMix-FIRE2020: A Meta Embedding and Transformer model for Code-Mixed Sentiment Analysis on Social Media Text*. http://arxiv.org/abs/2101.09004

Drakulich, K., Wozniak, K. H., Hagan, J., & Johnson, D. (2020). *Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics*. https://doi.org/10.1111/1745-9125.12239

Dunn, S., & Schuster, K. (2020). Research methods in the digital humanities: General introduction. In K. Schuster & S. Dunn (Eds.), *Routledge International Handbook of Research Methods in the Digital Humanities*. Routledge.

Durham, M. G., & Kellner, D. M. (Eds.). (2006). *Media and Cultural Studies: Key Works* (Revised). Blackwell Publishing.

Eck, K., & Hultman, L. (2007). One-Sided Violence Against Civilians in War: Insights from New Fatality Data. *Journal of Peace Research*, *44*(2), 233–246. https://doi.org/10.1177/0022343307075124

Eco, U. (1976a). The Subject of Semiotics. In *A Theory of Semiotics* (pp. 314–318). Indiana University Press. https://doi.org/10.1007/978-1-349-15849-2_5

Eco, U. (1976b). Theory of Codes. In *A Theory of Semiotics* (pp. 48–150). Indiana University Press. https://www.jstor.org/stable/j.ctt16xwcfd?refreqid=book-view%3A8fef78fda25b671bad3f8e189c9f8548

Eduardo Bonilla-silva, B., & Dietrich, D. (2011). The Sweet Enchantment of Color-Blind Racism in Obamerica. *The ANNALS of the American Academy of Political and Social Science*. https://doi.org/10.1177/0002716210389702

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fernández, N., Fuentes, D., Sánchez, L., & Fisteus, J. A. (2010). The NEWS ontology: Design and applications. *Expert Systems with Applications*, *37*(12), 8694–8704. https://doi.org/10.1016/j.eswa.2010.06.055

Fortuna, P. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6786–6794.

Galtung, J. (1969). Violence, Peace, and Peace Research. *Journal of Peace Research*, *6*(3), 167–191. https://about.jstor.org/terms

Galtung, J. (1990). Cultural Violence. In *Journal of Peace Research* (Vol. 27, Issue 3).

Galtung, J., & Fischer, D. (2013a). *Johan Galtung: Pioneer of Peace Research* (Vol. 5). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32481-9

Galtung, J., & Fischer, D. (2013b). *Johan Galtung: Pioneer of Peace Research* (H. G. Brauch, Ed.; 5th ed.). Springer.

Galtung, J., & Fischer, D. (2013c). Violence: Direct, Structural and Cultural. In *Johan Galtung: Pioneer of Peace Research* (pp. 35–40). https://doi.org/10.1007/978-3-642-32481-9_3

Galtung, J., & Njshimura, F. (1983). *Structure, Culture and Languages: An Essay Comparing the Indo-European, Chinese and Japanese Languages*. https://www.transcend.org/galtung/papers/Structure, Culture and Languages - An Essay Comparing Indo-European, Chinese and Japanese Languages.pdf

Gardner, M., Huang, K., Papalexakis, E., Fu, X., Talukdar, P., Faloutsos, C., Sidiropoulos, N., & Mitchell, T. (2015). *Translation Invariant Word Embeddings*. Association for Computational Linguistics.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media*.

Geva, M., Goldberg, Y., & Berant, J. (2019). *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets*.

Giswandhani, M. (2022). Yuni: Stereotype Representation Of Women. *Journal of Business, Social and Technology (Bustechno)*, *3*(1), 1–9. https://doi.org/10.46799/jbt.v3i1.63

Given, L. (2012). The SAGE Encyclopedia of Qualitative Research Methods. In L. M. Given (Ed.), *The SAGE Encyclopedia of Qualitative Research Methods: Vol. Volumes 1 & 2*. Sage.

Gleditsch, N. P., Nordkvelle, J., & Strand, H. (2014). *Peace research-Just the study of war?* https://doi.org/10.1177/0022343313514074

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring Hate Speech Detection in Multimodal Publications. *The IEEE Winter Conference on Applications of Computer Vision*, 1470–1478. https://www.hatebase.org/

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 609–614. https://github.com/gonenhila/gender_

Gregory, D. (2010). Seeing Red: Baghdad and the event-ful city. *Political Geography*, *29*(5), 266–279. https://doi.org/10.1016/j.polgeo.2010.04.003

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. https://doi.org/10.18653/v1/2021.eacl-main.114

Gurr, T. R. (2010). *Why Men Rebel*. Paradigm Publishers.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, *15*(5). https://doi.org/10.1371/journal.pone.0232525

Hall, S. (2006). Encoding and Decoding. In M. G. Durham & D. M. Kellner (Eds.), *Media and Cultural Studies: Key Works* (Revised, pp. 164–173). Blackwell Publishing.

Hall, W., & Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*.

Hammersley, M. (1992). Deconstructing the qualitative-quantitative divide. In *Mixing Methods: qualitative and quantitative research* (1st ed., pp. 39–55). Routledge.

Haney-López, I. (2014). *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. Oxford University Press.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *COLING-92*, 23–28. https://www.aclweb.org/anthology/C92-2082.pdf

Hinchman, L. P., & Hinchman, S. (Eds.). (1997). *Memory, Identity, Community: The Idea of Narrative in the Human Sciences - Google Books*. SUNY Press.

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. *Science Magazine*, *349*(6245), 261–266. http://science.sciencemag.org/

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. De, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1–37. https://doi.org/10.1145/3447772

Hogg, M. A. (2008). Social Categorization, Depersonalization, and Group Behavior. In *Blackwell Handbook of Social Psychology: Group Processes* (pp. 56–85). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470998458.ch3

Hogg, M. A. (2016). *Social Identity Theory* (pp. 3–17). https://doi.org/10.1007/978-3-319-29869-6_1

Hornby, A. S. (1954). *A Guide to Patterns and Usage in English*. Oxford University press.

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x

Hunston, S. (2019). Pattern Grammar. In C. A. Chapelle. (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–6). Wiley. https://doi.org/10.1002/9781405198431.wbeal0899.pub2

Hunston, S., & Francis, G. (2000). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. *Computational Linguistics*, *27*(2).

Hunston, S., Francis, G., & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1: Verbs*. HarperCollins.

Hunston, S., Francis, G., & Manning, E. (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. HarperCollins.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers*, *4*(8), 966–974.

Issa, A., Aldine, A., Harzallah, M., Berio, G., Béchet, N., Giuseppe, B., & Faour, A. (2018). Redefining Hearst Patterns by using Dependency Relations. In *KEOD* (pp. 146–153). https://doi.org/10.5220/0006962201480155

Jahan, L., Mittal, R., & Finlayson, M. (2021). Inducing Stereotypical Character Roles from Plot Structure. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 492–497. https://doi.org/10.18653/v1/2021.emnlp-main.39

Kahneman, D. (2011). *Thinking Fast and Slow*. Penguin. https://doi.org/0141033576

Kanchinadam, T., You, Q., Westpfahl, K., Kim, J., Gunda, S., Seith, S., & Fung, G. (2021). A Simple yet Brisk and Efficient Active Learning Platform for Text Classification. *ArXiv:2102.00426v1*. https://github.com/tkanchin/adaptive_reg_active_learning/

Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, *210*, 106458. https://doi.org/10.1016/j.knosys.2020.106458

Kapoor, N. (2021). Racism and the right: a response to Reactionary Democracy. *Ethnic and Racial Studies*, *44*(13), 2393–2398. https://doi.org/10.1080/01419870.2021.1896017

Kelman, H. C. (1981). Reflections on the History and Status of Peace Research. In *Source: Conflict Management and Peace Science* (Vol. 5, Issue 2).

Kingdon, A. R. (2017). Seductions of the Caliphate: A Cultural Criminological Analysis of Online Islamic State Propaganda. *Internet Journal of Criminology*.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (n.d.). *Skip-Thought Vectors*.

Knapp, K., Antos Volume, G., De Gruyter, M., Berlin, ·, & York, N. (2008). *Handbooks of Applied Linguistics Communication Competence Language and Communication Problems Practical Solutions*. http://dnb.d-nb.de.

Kosisochukwu, J. M., Gao, X., & Xue, B. (2020). In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 150–161. https://doi.org/10.18653/v1/P17

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. *SN Computer Science*, *2*(2). https://doi.org/10.1007/s42979-021-00457-3

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Labov, W. (1997). Some Further Steps in Narrative Analysis. *Journal of Narrative and Life History*, *7*(1–4), 395–415.

Labov, W., & Waletzky, J. (1997). Narrative Analysis: Oral Versions of Personal Experience. *Journal of Narrative & Life History*, *7*(1–4), 3–38.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Larsson, S. (2019). Narrative analysis and critical social work. In S. A. Webb (Ed.), *The Routledge Handbook of Critical Social Work* (1st ed., pp. 218–229). Routledge. https://doi.org/https://doi.org/10.4324/9781351264402

Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A Structured Self-attentive Sentence Embedding*. http://arxiv.org/abs/1703.03130

Lindgren, S. (2020). *Data theory: interpretive sociology and computational methods*. Polity.

Liu, Y., & Lapata, M. (n.d.). *Learning Structured Text Representations*.

Mahmood, K., & Afsar, A. (2016). Colligating "Terrorism": Syntax Of Terror Discourse In Post 9/11 Print Media. In *SECTION B Sci.Int.(Lahore)* (Vol. 28, Issue 4).

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/10.1016/J.JBUSRES.2020.07.045

Malik, J. S., Pang, G., & Hengel, A. van den. (2022). *Deep Learning for Hate Speech Detection: A Comparative Study*.

Malte, A., & Ratadiya, P. (2019). *Evolution of transfer learning in natural language processing*.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mansour, S. (2018). Social media analysis of user's responses to terrorism using sentiment analysis and text mining. *Procedia Computer Science*, *140*, 95–103. https://doi.org/10.1016/j.procs.2018.10.297

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417.

Martin, M. (2018). *Why we fight*. Hurst Publishers.

Matsuda, M. J. (1989a). Public Response to Racist Speech: Considering the Victim's Story. *Michigan Law Review*, *87*(8). https://doi.org/10.2307/1289306

Matsuda, M. J. (1989b). Public Response to Racist Speech: Considering the Victim's Story. In *Source: Michigan Law Review* (Vol. 87, Issue 8). https://about.jstor.org/terms

Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). Words that wound: Critical race theory, assaultive speech, and the first amendment. In *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. https://doi.org/10.4324/9780429502941

Matthews, P. H. (2003). *Linguistics: A Very Short Introduction*. Oxford University Press.

Meghini, C., Bartalesi, V., & Metilli, D. (2021). Representing Narratives in Digital Libraries: The Narrative Ontology. In *Semantic Web Journal* (Vol. 0, Issue 0). https://dlnarratives.eu/narratives.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://ronan.collobert.com/senna/

Mikolov, T., Chen, K., Sutskever, I., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013*, 746–751. http://research.microsoft.com/en-

Mondon, A., & Winter, A. (2020). *Reactionary democracy: how racism and the populist far right became mainstream*. Verso.

Mullah, N. S., & Zainon, W. M. N. W. (2021a). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Mullah, N. S., & Zainon, W. M. N. W. (2021b). *Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review | Enhanced Reader*. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3089515

Mullah, N. S., & Zainon, W. N. W. M. (2021c). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access Nanlir S*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Muller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3082972

Murdock, G. (2016). Encoding and Decoding. *The International Encyclopedia of Media Effects*, 1–11. https://doi.org/10.1002/9781118783764.wbieme0113

Nithyanand, R., Schaffner, B., & Gill, P. (2017). *Measuring Offensive Speech in Online Political Discourse*. http://arxiv.org/abs/1706.01875

Nivre, J. (2010). Dependency Parsing. *Language and Linguistics Compass*, *4*(3), 138–152. https://doi.org/10.1111/j.1749-818x.2010.00187.x

O’Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *Computer Law and Security Review*, *39*, 105474. https://doi.org/10.1016/j.clsr.2020.105474

Orlandi, T. (2021). Reflections on the Development of Digital Humanities. *Digital Scholarship in the Humanities*, *36*(Supplement_2), ii222–ii229. https://doi.org/10.1093/llc/fqaa048

Patel, A., & Arasanipalai, A. (2021). *Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand*. O’Reilly Media, Inc, USA.

Pemberton, A., Mulder, E., & Aarten, P. G. M. (2019). Stories of injustice: Towards a narrative victimology. *European Journal of Criminology*, *16*(4), 391–412. https://doi.org/10.1177/1477370818770843

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Arxiv.Org*.

Pettersson, T., & Eck, K. (2018). Organized violence, 1989-2017. *Journal of Peace Research*, *55*(4), 535–547.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation 2020 55:2*, *55*(2), 477–523. https://doi.org/10.1007/S10579-020-09502-8

Pollard, C. Jesse., & Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. 440.

Presser, L. (2009). The narratives of offenders. *Theoretical Criminology*, *13*(2), 177–200. https://doi.org/10.1177/1362480609102878

Presser, L. (2018). Inside story: How narratives drive mass harm. *Inside Story: How Narratives Drive Mass Harm*, 1–190.

Propp, V. I. (1968). *Morphology of the Folktale: Second Edition* (Vol. 9). University of Texas Press.

Randy, G., Ajay, C., Holzinger, Katharina Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The New 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction: Vol. 11015 LNCS* (pp. 1–10). https://doi.org/10.1007/978-3-319-99740-7_23

Resende, A., Railsback, D., Dowsley, R., Nascimento, A. C. A., & Aranha, D. F. (2021). Fast Privacy-Preserving Text Classification based on Secure Multiparty Computation. *ArXiv:2101.07365v1*. https://bitbucket.org/uwtppml/rustlynx/

Riessman, C. K. (2005). Narrative Analysis. In *Narrative, Memory & Everyday Life* (pp. 1–7). University of Huddersfield.

Riessman, C. K. (2008). Narrative Methods for the Human Sciences. *Narrative Inquiry*, *18*(2).

Rogers, M. M. (2021). Exploring the Domestic Abuse Narratives of Trans and Nonbinary People and the Role of Cisgenderism in Identity Abuse, Misgendering, and Pathologizing. *Violence Against Women*, *27*(12–13). https://doi.org/10.1177/1077801220971368

Roller, S., Kiela, D., & Nickel, M. (2018). Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *arXiv* (pp. 358–363). Association for Computational Linguistics.

Rospocher, M., Van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., & Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, *37–38*, 132–151. https://doi.org/10.1016/j.websem.2015.12.004

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2021). *Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks*. http://arxiv.org/abs/2112.07475

Rudnik, C., Teyssou, D., Ehrhart, T., Troncy, R., Ferret, O., & Tannier, X. (2019). Searching news articles using an event knowledge graph leveraged by Wikidata. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 1232–1239. https://doi.org/10.1145/3308560.3316761

Runciman, W. G. (Walter G. (1966). *Relative deprivation and social justice : a study of attitudes to social inequality in twentieth-century England*. 338.

Sabou, M., Bontcheva, K., & Scharl, A. (2012). Crowdsourcing research opportunities. *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 1–8. https://doi.org/10.1145/2362456.2362479

Saeed, W., & Omlin, C. (2021). *Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities*. https://arxiv.org/abs/2111.06420

Sandmann. (1995). Three Ifs and a Maybe: Mari Matsuda's Approach to Restricting Hate Speech. *Communication Studies*, *45*(3–4), 241.

Sang, Y., & Stanton, J. (2022). The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13192 LNCS*. https://doi.org/10.1007/978-3-030-96957-8_36

Saputra, A. G., & Noverino, R. (2023). The The Spheres Of Action And Narrative Function Analysis In Eternals Movie. *Journal International Seminar on Languages, Literature, Arts, and Education (ISLLAE)* , *5*(1).

Saussure, F. de. (1916). *Course in General Linguistics*.

Schafer, J. A. (2002). Spinning the web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture*, *9*(2), 69–88.

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. Association for Computational Linguistics. https://en.wikipedia.org/wiki/List_

Schuler, K. K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon* [PhD]. University of Pennsylvania.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioural and Brain Sciences*, *3*, 417–457. https://doi.org/10.1017/S0140525X00005756

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1). https://doi.org/10.1145/505282.505283

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 623–656.

Shmueli, B., Fell, J., Ray, S., & Ku, L.-W. (2021). *Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing*.

Simpson, Emile. (2012). *War From the Ground Up: Twenty-First Century Combat as Politics* (Vol. 20). C Hurst & Co Publishers Ltd.

Singer, J. W. (1984). The Player and the Cards: Nihilism and Legal Theory. *The Yale Law Journal*, *94*(1).

Spruin, E., Alleyne, E., & Papadaki, I. (2015). Domestic abuse victims' perceptions of abuse and support: a narrative study. *Journal of Criminological Research, Policy and Practice*, *1*(1). https://doi.org/10.1108/JCRPP-10-2014-0002

Squire, M. (2021). Monetizing Propaganda: How Far-right Extremists Earn Money by Video Streaming. *WebSci '21*, 158–167. https://doi.org/10.1145/3447535.3462490

Stappen, L., Baird, A., Schumann, L., & Schuller, B. (2021). *The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements*. http://arxiv.org/abs/2101.06053

Stephenson, C. (2020). *Peace Research / Peace Studies : A Twentieth Century Intellectual History* (Issue January). https://doi.org/10.1093/acrefore/9780190846626.013.273

Steuter, E., & Wills, D. (2010). 'The vermin have struck again': dehumanizing the enemy in post 9/11 media representations. *Media, War & Conflict*, *3*(2), 152–167. https://doi.org/10.1177/1750635210360082

Strachan, Hew., & Scheipers, Sibylle. (Eds.). (2011). *The Changing Character of War*. Oxford University Press.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tajfel, H., & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worschel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Brooks/Cole. http://www.ark143.org/wordpress2/wp-content/uploads/2013/05/Tajfel-Turner-1979-An-Integrative-Theory-of-Intergroup-Conflict.pdf

Tell MAMA. (2020). The Impact of the Christchurch Terror Attack. In *2019 Interim Report*. https://www.tellmamauk.org/wp-content/uploads/2020/03/The-Impact-of-the-ChristChurch-Attack-Tell-MAMA-Interim-Report-2019-PP.pdf

TellMAMA. (2022). *The Tangled Web of Far Right Anti-Muslim Hate*. www.facebook.com/tellmamauk

Tesnière, L. (1959). *Elements of Structural Syntax* (T. Osborne & S. Kahane, Eds.). John Benjamins Publishing Company. https://doi.org/10.1093/acprof:oso/9780199299829.003.0002

Titley, G. (2020). *Is free speech racist?* Polity Press.

Todorov, T., & Weinstein, A. (1969). Structural Analysis of Narrative. *NOVEL: A Forum on Fiction*, *3*(1), 70. https://doi.org/10.2307/1345003

Turner, J. C., & Reynolds, K. J. (2011). Self-Categorization Theory. In *Handbook of Theories of Social Psychology* (pp. 399–417). SAGE Publications Ltd. https://doi.org/10.4135/9781446249222.n46

Uschold, M., & Gruninger, M. (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, *11*(2), 93–136. https://doi.org/10.1017/S0269888900007797

Van Der Zwaan, J. M., Smink, W., Sools, A., Nl, M. S., Veldkamp, B., Wiegersma, S., & Nl, W. (2017, July 3). *Flexible NLP Pipelines for Digital Humanities Research*.

van Dijk, T. A. (1983). *Processes of Prejudice and the Roots of Racism: A Socio-Cognitive Approach* (No. 3).

Van Dijk, T. A. (2014). Stories and Racism. *Narrative and Social Control: Critical Perspectives*, 121–142. https://doi.org/10.4135/9781483345277.n6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., Tromble, R., Waseem, Z., & Hale, S. (2020). *Detecting East Asian Prejudice on Social Media*.

Vidgen, B., Margetts, H., & Harris, A. (2019). *How much online abuse is there? A systematic review of evidence for the UK Policy Briefing – Full Report*.

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: the Contextual Abuse Dataset. *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2289–2303. https://github.com/dongpng/cad_

Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2020). *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*. https://github.com/bvidgen/

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A. L., Palmero Aprosio, A., Rigau, G., Rospocher, M., & Segers, R. (2016). NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, *110*, 60–85. https://doi.org/10.1016/j.knosys.2016.07.013

W3C. (2014). RDF Schema 1.1. In *World Wide Web Consortium*. https://www.w3.org/TR/rdf-schema/

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *NLP + CSS 2016 - EMNLP 2016 Workshop on Natural Language Processing and Computational Social Science, Proceedings of the Workshop*. https://doi.org/10.18653/v1/w16-5618

Webber, C. (2007). Revaluating relative deprivation theory. *Theoretical Criminology*, *11*(1). https://doi.org/10.1177/1362480607072737

Webber, C. (2021). Rediscovering the Relative Deprivation and Crime Debate: Tracking its Fortunes from Left Realism to the Precariat. *Critical Criminology*. https://doi.org/10.1007/s10612-021-09554-4

Webel, C., & Galtung, J. (2007). Handbook of Peace and Conflict Studies. In *Handbook of Peace and Conflict Studies*. https://doi.org/10.4324/9780203089163

Weber, T. (2004). The impact of Gandhi on the development of Johan Galtung's peace research, Global Change. *Peace & Security*, *16*(1). https://doi.org/10.1080/1478115042000176166

Weinstein, J. (1991). a Constitutional Roadmap To the Regulation of Campus Hate Speech. *The Wayne Law Review*, *38*(163).

Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., Xue, N., Palmer, M., Hwang, J. D., Bonial, C., Choi, J., Mansouri, A., Foster, M., Hawwary, A.-A., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., & Houston, A. (2012). *OntoNotes Release 5.0*. http://www.bbn.com/NLP/OntoNotes

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. *NAACL-HLT*, 602–608. www.kaggle.com/c/jigsaw-toxic-comment-

Wieting, J., & Gimpel, K. (2017). *Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings*. http://arxiv.org/abs/1705.00364

Wright, R. (2010). *The Evolution of God: The Origins Of Our Beliefs*. Little Brown Book Group.

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-Based Bag-of-Words Model for Text Classification. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.2991074

Yang, D., Heaney, T., Tonon, A., Wang, L., Cudré-Mauroux, P., Yang DingqiYang, D., Terence Heaney TerenceHeaney, unifrch, Alberto Tonon AlbertoTonon, unifrch, Leye Wang, unifrch, & Cudré-Mauroux PhilippeCudre-Mauroux, P. (2018). *CrimeTelescope: crime hotspot prediction based on urban and social media data fusion*. *21*, 1323–1347. https://doi.org/10.1007/s11280-017-0515-4

Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and Characterizing Hate Speech on News Websites. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 125–134. https://doi.org/10.1145/3394231.3397902

Zhang, X., & Sheng, J. (2017). *A Peircean semiotic interpretation of a social sign*. https://doi.org/10.1016/j.annals.2017.04.002

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4). https://doi.org/10.1007/s13042-010-0001-0

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., & Pandey, H. M. (2021). Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Networks*, *133*, 40–56. https://doi.org/10.1016/j.neunet.2020.10.001

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, *6*(3), 275–290. https://doi.org/10.1016/J.ENG.2019.12.014

Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic Hate Speech Detection using Machine Learning: A Comparative Study. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agarwal, S., & Chowdary, C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications*, *185*, 115632. https://doi.org/10.1016/J.ESWA.2021.115632

Akinsanya, A., & Bach, C. (2014). Narrative analysis: The Personal Experience Narrative approach. *ASEE 2014 Zone I Conference*, 159–168.

Allwood, C. M., & Allwood, C. M. (2012). The distinction between qualitative and quantitative research methods is problematic. *Qual Quant*, *46*, 1417–1429. https://doi.org/10.1007/s11135-011-9455-8

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2018). The Enemy Among Us': Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web*, *13*(3), 39.

Alrehili, A. (2019). Automatic Hate Speech Detection on Social Media: A Brief Survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 1–6. https://doi.org/10.1109/AICCSA47632.2019.9035228

Anderson, B. (2006). *Imagined communities: reflections on the origin and spread of nationalism*.

Atteveldt, W. van, Velden, M. A. C. G. van der, & Boukes, M. (2021). The Validity of Sentiment Analysis:Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 1–20. https://doi.org/10.1080/19312458.2020.1869198

Augoustinos, Martha., & Reynolds, K. J. (2001). *Understanding prejudice, racism, and social conflict*. 362.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005

Benaquisto, L. (2008). Thematic Coding and Analysis. In L. M. Given (Ed.), *The Sage Encyclopaedia of Qualitative Research Methods: Vol. Volumes 1 & 2* (pp. 85–88). Sage.

Bengfort, B., Bilbro, Rebecca., & Ojeda, Tony. (2018). *Applied text analysis with Python*. O'Reilly Media, Inc.

Beniwal, R., & Maurya, A. (2021). Toxic Comment Classification Using Hybrid Deep Learning Model. *Sustainable Communication Networks and Application*, 461–473. https://doi.org/10.1007/978-981-15-8677-4_38

Bird, Steven., Klein, Ewan., & Loper, Edward. (2009). *Natural Language Processing with Python*. O'Reilly.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, *29*. http://cs.cmu.edu/directory/csd.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bresnan, J. (ed. ). (1982). *The Mental Representation of Grammatical Relations*. MIT Press.

Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Burnap, P., & Williams, M. L. (2015a). Cyber hate speech on twitter: An application of machine classification and statistical modelling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2015b). Cyber hate speech on twitter: An application of machine classification and statistical modelling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burton-Roberts, N. (2016). Analysing Sentences: An Introduction to English Syntax. In *Analysing Sentences: An Introduction to English Syntax: Fourth Edition* (4th ed.). Routledge. https://doi.org/10.4324/9781315646046

Cao, R., Lee, R. K. W., & Hoang, T. A. (2020). DeepHate: Hate Speech Detection via Multi-Faceted Text Representations. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 11–20. https://doi.org/10.1145/3394231.3397890

Capelos, T., & Katsanidou, A. (2018). Reactionary Politics: Explaining the Psychological Roots of Anti Preferences in European Integration and Immigration Debates. *Political Psychology*, *39*(6), 1271–1288. https://doi.org/10.1111/pops.12540

Carter, R., Collins, M., Davis, G., Hermansson, P., Lawrence, D., Mulhall, J., Murdoch, S., & Velleman, L. (2020). *State of Hate 2020*.

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., & Lorés, J. (2004). Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In Bussler C.J., Davies J., Fensel D., & Studer R. (Eds.), *The Semantic Web: Research and Applications. ESWS 2004. Lecture Notes in Computer Science,* (Vol. 3053, pp. 445–458). Springer, Berlin, Heidelberg.

Chandler, D. (2005). Semiotics for Beginners. In *Retrieved March* (Vol. 2, Issue 1969). https://doi.org/10.1519/JSC.0b013e3181e7ff75

Chandler, D. (2017). *Semiotics: The Basics* (3rd ed., Issue 1969).

Chandra, M., Pailla, D., Bhatia, H., Sanchawala, A., Gupta, M., Shrivastava, M., & Kumaraguru, P. (2021, June). "Subverting the Jewtocracy"': Online Antisemitism Detection Using Multimodal Deep Learning; *WebSci '21*. https://doi.org/10.1145/3447535.3462502

Chase, S. E. (2011). Narrative Inquiry: Still a Field in the Making. In N. K. Denzin & Y. S. Loncoln (Eds.), *The Sage Handbook of Qualitative Research* (4th ed., pp. 421–435). Sage.

Chierchia, G., & McConnell-Ginet, S. (2000a). *Meaning and Grammar: An Introduction to Semantics* (2nd ed.). MIT Press.

Chierchia, G., & McConnell-Ginet, S. (2000b). *Meaning and Grammar: An Introduction to Semantics.* (2nd ed.). MIT Press. https://doi.org/10.1016/0004-3702(93)90024-6

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, *14*(1), 322–352. https://doi.org/10.1007/s12559-021-09862-5

Chomsky, Noam. (1957). Syntactic Structures. *Language*, *33*(3), 375. https://doi.org/10.2307/411160

Coates, J. (1995). *Armed and dangerous: the rise of the survivalist right*. Hill and Wang.

Collier, P., & Hoeffler, A. (1998). On Economic Causes of Civil War. *Oxford Economic Papers*, *50*, 563–573.

Collier, P., & Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, *56*, 563–595. https://doi.org/10.1093/oep/gpf064

Collier, P., Hoeffler, A., & Rohner, D. (2009). Beyond Greed and Grievance: Feasibility and Civil War. *Oxford Economic Papers*, *61*(1), 1–27. https://doi.org/10.1093/oep/gpn029

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35.

D.C, A., S.O, A., E, O. I., & F.A, S. (2021). Text Classification Using Hybrid Machine Learning Algorithms on Big Data. In *International Journal of Trend in Research and Development* (Vol. 6, Issue 5).

Delgado, R. (1993). The Inward Turn in Outsider Jurisprudence. *William & Mary Law Review*, *34*, 741–768.

Delgado, R., & Stefancic, J. (2017). *Critical Race Theory* (3rd ed.). University Press.

Denzin, N. K., & Lincoln, Y. S. (2018). Introduction: The Discipline and Practice of Qualitative Research. In *The Sage Handbook of Qualitative Research*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Arxiv.Org*.

Dijk, T. A. van. (1998). *Ideology: A Multidisciplinary Approach*. Sage.

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*.

Dowlagar, S., & Mamidi, R. (2021). *CMSAOne@Dravidian-CodeMix-FIRE2020: A Meta Embedding and Transformer model for Code-Mixed Sentiment Analysis on Social Media Text*. http://arxiv.org/abs/2101.09004

Drakulich, K., Wozniak, K. H., Hagan, J., & Johnson, D. (2020). *Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics*. https://doi.org/10.1111/1745-9125.12239

Dunn, S., & Schuster, K. (2020). Research methods in the digital humanities: General introduction. In K. Schuster & S. Dunn (Eds.), *Routledge International Handbook of Research Methods in the Digital Humanities*. Routledge.

Durham, M. G., & Kellner, D. M. (Eds.). (2006). *Media and Cultural Studies: Key Works* (Revised). Blackwell Publishing.

Eck, K., & Hultman, L. (2007). One-Sided Violence Against Civilians in War: Insights from New Fatality Data. *Journal of Peace Research*, *44*(2), 233–246. https://doi.org/10.1177/0022343307075124

Eco, U. (1976a). The Subject of Semiotics. In *A Theory of Semiotics* (pp. 314–318). Indiana University Press. https://doi.org/10.1007/978-1-349-15849-2_5

Eco, U. (1976b). Theory of Codes. In *A Theory of Semiotics* (pp. 48–150). Indiana University Press.

Eduardo Bonilla-silva, B., & Dietrich, D. (2011). The Sweet Enchantment of Color-Blind Racism in Obamerica. *The ANNALS of the American Academy of Political and Social Science*. https://doi.org/10.1177/0002716210389702

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fernández, N., Fuentes, D., Sánchez, L., & Fisteus, J. A. (2010). The NEWS ontology: Design and applications. *Expert Systems with Applications*, *37*(12), 8694–8704. https://doi.org/10.1016/j.eswa.2010.06.055

Fortuna, P. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6786–6794.

Galtung, J. (1969). Violence, Peace, and Peace Research. *Journal of Peace Research*, *6*(3), 167–191. https://about.jstor.org/terms

Galtung, J. (1990). Cultural Violence. In *Journal of Peace Research* (Vol. 27, Issue 3).

Galtung, J., & Fischer, D. (2013a). *Johan Galtung: Pioneer of Peace Research* (Vol. 5). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32481-9

Galtung, J., & Fischer, D. (2013b). *Johan Galtung: Pioneer of Peace Research* (H. G. Brauch, Ed.; 5th ed.). Springer.

Galtung, J., & Fischer, D. (2013c). Violence: Direct, Structural and Cultural. In *Johan Galtung: Pioneer of Peace Research* (pp. 35–40). https://doi.org/10.1007/978-3-642-32481-9_3

Galtung, J., & Njshimura, F. (1983). *Structure, Culture and Languages: An Essay Comparing the Indo-European, Chinese and Japanese Languages*.

Gardner, M., Huang, K., Papalexakis, E., Fu, X., Talukdar, P., Faloutsos, C., Sidiropoulos, N., & Mitchell, T. (2015). *Translation Invariant Word Embeddings*. Association for Computational Linguistics.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media*.

Geva, M., Goldberg, Y., & Berant, J. (2019). *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets*.

Giswandhani, M. (2022). Yuni: Stereotype Representation Of Women. *Journal of Business, Social and Technology (Bustechno)*, *3*(1), 1–9. https://doi.org/10.46799/jbt.v3i1.63

Given, L. (2012). The SAGE Encyclopedia of Qualitative Research Methods. In L. M. Given (Ed.), *The SAGE Encyclopedia of Qualitative Research Methods: Vol. Volumes 1 & 2*. Sage.

Gleditsch, N. P., Nordkvelle, J., & Strand, H. (2014). *Peace research-Just the study of war?* https://doi.org/10.1177/0022343313514074

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring Hate Speech Detection in Multimodal Publications. *The IEEE Winter Conference on Applications of Computer Vision*, 1470–1478.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 609–614.

Gregory, D. (2010). Seeing Red: Baghdad and the eventful city. *Political Geography*, *29*(5), 266–279. https://doi.org/10.1016/j.polgeo.2010.04.003

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. *EACL 2021 - 16th Conference of the European*

*Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. https://doi.org/10.18653/v1/2021.eacl-main.114

Gurr, T. R. (2010). *Why Men Rebel*. Paradigm Publishers.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of pre-processing on text classification using a bag-of-words representation. *PLoS ONE*, *15*(5). https://doi.org/10.1371/journal.pone.0232525

Hall, S. (2006). Encoding and Decoding. In M. G. Durham & D. M. Kellner (Eds.), *Media and Cultural Studies: Key Works* (Revised, pp. 164–173). Blackwell Publishing.

Hall, W., & Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*.

Hammersley, M. (1992). Deconstructing the qualitative-quantitative divide. In *Mixing Methods: qualitative and quantitative research* (1st ed., pp. 39–55). Routledge.

Haney-López, I. (2014). *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. Oxford University Press.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *COLING-92*, 23–28.

Hinchman, L. P., & Hinchman, S. (Eds.). (1997). *Memory, Identity, Community: The Idea of Narrative in the Human Sciences*. SUNY Press.

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. *Science Magazine*, *349*(6245), 261–266.

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. de, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1–37. https://doi.org/10.1145/3447772

Hogg, M. A. (2008). Social Categorization, Depersonalization, and Group Behavior. In *Blackwell Handbook of Social Psychology: Group Processes* (pp. 56–85). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470998458.ch3

Hogg, M. A. (2016). *Social Identity Theory* (pp. 3–17). https://doi.org/10.1007/978-3-319-29869-6_1

Hornby, A. S. (1954). *A Guide to Patterns and Usage in English*. Oxford University press.

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x

Hunston, S. (2019). Pattern Grammar. In C. A. Chapelle. (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–6). Wiley. https://doi.org/10.1002/9781405198431.wbeal0899.pub2

Hunston, S., & Francis, G. (2000). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. *Computational Linguistics*, *27*(2).

Hunston, S., Francis, G., & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1: Verbs*. HarperCollins.

Hunston, S., Francis, G., & Manning, E. (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. HarperCollins.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers*, *4*(8), 966–974.

Issa, A., Aldine, A., Harzallah, M., Berio, G., Béchet, N., Giuseppe, B., & Faour, A. (2018). Redefining Hearst Patterns by using Dependency Relations. In *KEOD* (pp. 146–153). https://doi.org/10.5220/0006962201480155

Jahan, L., Mittal, R., & Finlayson, M. (2021). Inducing Stereotypical Character Roles from Plot Structure. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 492–497. https://doi.org/10.18653/v1/2021.emnlp-main.39

Kahneman, D. (2011). *Thinking Fast and Slow*. Penguin. https://doi.org/0141033576

Kanchinadam, T., You, Q., Westpfahl, K., Kim, J., Gunda, S., Seith, S., & Fung, G. (2021). A Simple yet Brisk and Efficient Active Learning Platform for Text Classification. *ArXiv:2102.00426v1*.

Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, *210*, 106458. https://doi.org/10.1016/j.knosys.2020.106458

Kapoor, N. (2021). Racism and the right: a response to Reactionary Democracy. *Ethnic and Racial Studies*, *44*(13), 2393–2398. https://doi.org/10.1080/01419870.2021.1896017

Kelman, H. C. (1981). Reflections on the History and Status of Peace Research. In *Source: Conflict Management and Peace Science* (Vol. 5, Issue 2).

Kingdon, A. R. (2017). Seductions of the Caliphate: A Cultural Criminological Analysis of Online Islamic State Propaganda. *Internet Journal of Criminology*.

Knapp, K., Antos Volume, G., de Gruyter, M., Berlin, ·, & York, N. (2008). *Handbooks of Applied Linguistics Communication Competence Language and Communication Problems Practical Solutions*.

Kosisochukwu, J. M., Gao, X., & Xue, B. (2020). In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 150–161. https://doi.org/10.18653/v1/P17

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. *SN Computer Science*, *2*(2). https://doi.org/10.1007/s42979-021-00457-3

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Labov, W. (1997). Some Further Steps in Narrative Analysis. *JOURNAL OF NARRATIVE AND LIFE HISTORY*, *7*(1–4), 395–415.

Labov, W., & Waletzky, J. (1997). Narrative Analysis: Oral Versions of Personal Experience. *Journal of Narrative & Life History*, *7*(1–4), 3–38.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Larsson, S. (2019). Narrative analysis and critical social work. In S. A. Webb (Ed.), *The Routledge Handbook of Critical Social Work* (1st ed., pp. 218–229). Routledge. https://doi.org/https://doi.org/10.4324/9781351264402

Lindgren, S. (2020). *Data theory: interpretive sociology and computational methods*. Polity.

Mahmood, K., & Afsar, A. (2016). Colligating "Terrorism": Syntax of Terror Discourse in Post 9/11 Print Media. In *SECTION B Sci.Int.(Lahore)* (Vol. 28, Issue 4).

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/10.1016/J.JBUSRES.2020.07.045

Malte, A., & Ratadiya, P. (2019). *Evolution of transfer learning in natural language processing*.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mansour, S. (2018). Social media analysis of user's responses to terrorism using sentiment analysis and text mining. *Procedia Computer Science*, *140*, 95–103. https://doi.org/10.1016/j.procs.2018.10.297

Mäntylä, M. v., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417.

Martin, M. (2018). *Why we fight*. Hurst Publishers.

Matsuda, M. J. (1989a). Public Response to Racist Speech: Considering the Victim's Story. In *Source: Michigan Law Review* (Vol. 87, Issue 8). https://about.jstor.org/terms

Matsuda, M. J. (1989b). Public Response to Racist Speech: Considering the Victim's Story. *Michigan Law Review*, *87*(8). https://doi.org/10.2307/1289306

Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). Words that wound: Critical race theory, assaultive speech, and the first amendment. In *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. https://doi.org/10.4324/9780429502941

Matthews, P. H. (2003). *Linguistics: A Very Short Introduction*. Oxford University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://ronan.collobert.com/senna/

Mikolov, T., Chen, K., Sutskever, I., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013*, 746–751.

Mondon, A., & Winter, A. (2020). *Reactionary democracy: how racism and the populist far right became mainstream*. Verso.

Mullah, N. S., & Zainon, W. M. N. W. (2021a). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Muller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3082972

Murdock, G. (2016). Encoding and Decoding. *The International Encyclopedia of Media Effects*, 1–11. https://doi.org/10.1002/9781118783764.wbieme0113

Nithyanand, R., Schaffner, B., & Gill, P. (2017). *Measuring Offensive Speech in Online Political Discourse*. http://arxiv.org/abs/1706.01875

Nivre, J. (2010). Dependency Parsing. *Language and Linguistics Compass*, *4*(3), 138–152. https://doi.org/10.1111/j.1749-818x.2010.00187.x

O'Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *Computer Law and Security Review*, *39*, 105474. https://doi.org/10.1016/j.clsr.2020.105474

Orlandi, T. (2021). Reflections on the Development of Digital Humanities. *Digital Scholarship in the Humanities*, *36*(Supplement_2), ii222–ii229. https://doi.org/10.1093/llc/fqaa048

Patel, A., & Arasanipalai, A. (2021). *Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand*. O'Reilly Media, Inc, USA.

Pemberton, A., Mulder, E., & Aarten, P. G. M. (2019). Stories of injustice: Towards a narrative victimology. *European Journal of Criminology*, *16*(4), 391–412. https://doi.org/10.1177/1477370818770843

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Arxiv.Org*.

Pettersson, T., & Eck, K. (2018). Organized violence, 1989-2017. *Journal of Peace Research*, *55*(4), 535–547.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation 2020 55:2*, *55*(2), 477–523. https://doi.org/10.1007/S10579-020-09502-8

Pollard, C. Jesse., & Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. 440.

Presser, L. (2009). The narratives of offenders. *Theoretical Criminology*, *13*(2), 177–200. https://doi.org/10.1177/1362480609102878

Presser, L. (2018). Inside story: How narratives drive mass harm. *Inside Story: How Narratives Drive Mass Harm*, 1–190.

Propp, V. I. (1968). *Morphology of the Folktale: Second Edition* (Vol. 9). University of Texas Press.

Randy, G., Ajay, C., Holzinger, Katharina Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The New 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction: Vol. 11015 LNCS* (pp. 1–10). https://doi.org/10.1007/978-3-319-99740-7_23

Resende, A., Railsback, D., Dowsley, R., Nascimento, A. C. A., & Aranha, D. F. (2021). Fast Privacy-Preserving Text Classification based on Secure Multiparty Computation. *ArXiv:2101.07365v1*.

Riessman, C. K. (2005). Narrative Analysis. In *Narrative, Memory & Everyday Life* (pp. 1–7). University of Huddersfield.

Riessman, C. K. (2008). Narrative Methods for the Human Sciences. *Narrative Inquiry*, *18*(2).

Rogers, M. M. (2021). Exploring the Domestic Abuse Narratives of Trans and Nonbinary People and the Role of Cisgenderism in Identity Abuse, Misgendering, and Pathologizing. *Violence Against Women*, *27*(12–13). https://doi.org/10.1177/1077801220971368

Roller, S., Kiela, D., & Nickel, M. (2018). Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *arXiv* (pp. 358–363). Association for Computational Linguistics.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., & Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, *37–38*, 132–151. https://doi.org/10.1016/j.websem.2015.12.004

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2021). *Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks*. http://arxiv.org/abs/2112.07475

Rudnik, C., Teyssou, D., Ehrhart, T., Troncy, R., Ferret, O., & Tannier, X. (2019). Searching news articles using an event knowledge graph leveraged by Wikidata. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 1232–1239. https://doi.org/10.1145/3308560.3316761

Runciman, W. G. (Walter G. (1966). *Relative deprivation and social justice : a study of attitudes to social inequality in twentieth-century England*. 338.

Sabou, M., Bontcheva, K., & Scharl, A. (2012). Crowdsourcing research opportunities. *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 1–8. https://doi.org/10.1145/2362456.2362479

Saeed, W., & Omlin, C. (2021). *Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities*. https://arxiv.org/abs/2111.06420

Sandman, W. (1995). Three ifs and a maybe: Mari Matsuda's approach to restricting hate-speech. *Communication Studies*, *45*(3–4), 241.

Sang, Y., & Stanton, J. (2022). The Origin and Value of Disagreement Among Data Labellers: A Case Study of Individual Differences in Hate Speech Annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13192 LNCS*. https://doi.org/10.1007/978-3-030-96957-8_36

Saputra, A. G., & Noverino, R. (2023). The Spheres of Action and Narrative Function Analysis in Eternals Movie. *Journal International Seminar on Languages, Literature, Arts, and Education (ISLLAE)*, *5*(1).

Saussure, F. de. (1916). *Course in General Linguistics*.

Schafer, J. A. (2002). Spinning the web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture*, *9*(2), 69–88.

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. Association for Computational Linguistics.

Schuler, K. K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon* [PhD]. University of Pennsylvania.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioural and Brain Sciences*, *3*, 417–457. https://doi.org/10.1017/S0140525X00005756

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1). https://doi.org/10.1145/505282.505283

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 623–656.

Shmueli, B., Fell, J., Ray, S., & Ku, L.-W. (2021). *Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing*.

Simpson, Emile. (2012). *War From the Ground Up: Twenty-First Century Combat as Politics* (Vol. 20). C Hurst & Co Publishers Ltd.

Singer, J. W. (1984). The Player and the Cards: Nihilism and Legal Theory. *The Yale Law Journal*, *94*(1).

Spruin, E., Alleyne, E., & Papadaki, I. (2015). Domestic abuse victims' perceptions of abuse and support: a narrative study. *Journal of Criminological Research, Policy and Practice*, *1*(1). https://doi.org/10.1108/JCRPP-10-2014-0002

Squire, M. (2021). Monetizing Propaganda: How Far-right Extremists Earn Money by Video Streaming. *WebSci '21*, 158–167. https://doi.org/10.1145/3447535.3462490

Stappen, L., Baird, A., Schumann, L., & Schuller, B. (2021). *The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements*. http://arxiv.org/abs/2101.06053

Stephenson, C. (2020). *Peace Research / Peace Studies: A Twentieth Century Intellectual History Peace Research / Peace Studies: A Twentieth Century Intellectual History the Forerunners of Peace Research* (Issue January). https://doi.org/10.1093/acrefore/9780190846626.013.273

Steuter, E., & Wills, D. (2010). 'The vermin have struck again': dehumanizing the enemy in post 9/11 media representations. *Media, War & Conflict*, *3*(2), 152–167. https://doi.org/10.1177/1750635210360082

Strachan, Hew., & Scheipers, Sibylle. (Eds.). (2011). *The Changing Character of War*. Oxford University Press.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tajfel, H., & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worschel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Brooks/Cole.

Tell MAMA. (2020). The Impact of the Christchurch Terror Attack. In *2019 Interim Report*.

TellMAMA. (2022). *The Tangled Web of Far Right Anti-Muslim Hate*.

Tesnière, L. (1959). *Elements of Structural Syntax* (T. Osborne & S. Kahane, Eds.). John Benjamins Publishing Company. https://doi.org/10.1093/acprof:oso/9780199299829.003.0002

Titley, G. (2020). *Is free speech racist?* Polity Press.

Todorov, T., & Weinstein, A. (1969). Structural Analysis of Narrative. *NOVEL: A Forum on Fiction*, *3*(1), 70. https://doi.org/10.2307/1345003

Turner, J. C., & Reynolds, K. J. (2011). Self-Categorization Theory. In *Handbook of Theories of Social Psychology* (pp. 399–417). SAGE Publications Ltd. https://doi.org/10.4135/9781446249222.n46

van der Zwaan, J. M., Smink, W., Sools, A., Nl, M. S., Veldkamp, B., Wiegersma, S., & Nl, W. (2017). *Flexible NLP Pipelines for Digital Humanities Research*.

van Dijk, T. A. (1983). *Processes of Prejudice and the Roots of Racism: A Socio-Cognitive Approach* (No. 3).

van Dijk, T. A. (2014). Stories and Racism. *Narrative and Social Control: Critical Perspectives*, 121–142. https://doi.org/10.4135/9781483345277.n6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., Tromble, R., Waseem, Z., & Hale, S. (2020). *Detecting East Asian Prejudice on social media*.

Vidgen, B., Margetts, H., & Harris, A. (2019). *How much online abuse is there? A systematic review of evidence for the UK Policy Briefing – Full Report*.

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: the Contextual Abuse Dataset. *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2289–2303.

Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2020). *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*.

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A. L., Palmero Aprosio, A., Rigau, G., Rospocher, M., & Segers, R. (2016). NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, *110*, 60–85. https://doi.org/10.1016/j.knosys.2016.07.013

W3C. (2014). RDF Schema 1.1. In *World Wide Web Consortium*.

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *NLP + CSS 2016 - EMNLP 2016 Workshop on Natural Language Processing* *and Computational Social Science, Proceedings of the Workshop*. https://doi.org/10.18653/v1/w16-5618

Webber, C. (2007). Revaluating relative deprivation theory. *Theoretical Criminology*, *11*(1). https://doi.org/10.1177/1362480607072737

Webber, C. (2021). Rediscovering the Relative Deprivation and Crime Debate: Tracking its Fortunes from Left Realism to the Precariat. *Critical Criminology*. https://doi.org/10.1007/s10612-021-09554-4

Webel, C., & Galtung, J. (2007). Handbook of Peace and Conflict Studies. In *Handbook of Peace and Conflict Studies*. https://doi.org/10.4324/9780203089163

Weber, T. (2004). The impact of Gandhi on the development of Johan Galtung's peace research, Global Change. *Peace & Security*, *16*(1). https://doi.org/10.1080/1478115042000176166

Weinstein, J. (1991). a Constitutional Roadmap to the Regulation of Campus Hate Speech. *The Wayne Law Review*, *38*(163).

Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., Xue, N., Palmer, M., Hwang, J. D., Bonial, C., Choi, J., Mansouri, A., Foster, M., Hawwary, A.-A., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., & Houston, A. (2012). *OntoNotes Release 5.0*.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. *NAACL-HLT*, 602–608.

Wright, R. (2010). *The Evolution of God: The Origins of Our Beliefs*. Little Brown Book Group.

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-Based Bag-of-Words Model for Text Classification. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.2991074

Yang, D., Heaney, T., Tonon, A., Wang, L., Cudré-Mauroux, P., Yang DingqiYang, D., Terence Heaney TerenceHeaney, unifrch, Alberto Tonon AlbertoTonon, unifrch, Leye Wang, unifrch, & Cudré-Mauroux PhilippeCudre-Mauroux, P. (2018). *CrimeTelescope: crime hotspot prediction based on urban and social media data fusion*. *21*, 1323–1347. https://doi.org/10.1007/s11280-017-0515-4

Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and Characterizing Hate Speech on News Websites. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 125–134. https://doi.org/10.1145/3394231.3397902

Zhang, X., & Sheng, J. (2017). *A Peircean semiotic interpretation of a social sign*. https://doi.org/10.1016/j.annals.2017.04.002

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4). https://doi.org/10.1007/s13042-010-0001-0

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., & Pandey, H. M. (2021). Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Networks*, *133*, 40–56. https://doi.org/10.1016/j.neunet.2020.10.001

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modelling, Learning, and Reasoning. *Engineering*, *6*(3), 275–290. https://doi.org/10.1016/J.ENG.2019.12.014