University of Southampton

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Embracing Emojis in Sarcasm Detection to Enhance Sentiment Analysis

by

Malak Abdullah Alsabban

ORCID ID 0000-0002-2668-0206

Thesis for the degree of Doctor of Philosophy in Computer Science

March 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Doctor of Philosophy

Embracing Emojis in Sarcasm Detection to Enhance Sentiment Analysis

by

Malak Abdullah Alsabban

People frequently share their ideas, concerns, and emotions on social networks, making sentiment analysis on social media increasingly important for understanding public opinion and user sentiment. Sentiment analysis provides an effective means of interpreting people's attitudes towards various topics, individuals, or ideas.

This thesis introduces the creation of an Emoji Dictionary (ED) to harness the rich contextual information conveyed by emojis. It acts as a valuable resource for deciphering the emotional nuances embedded in textual content, contributing to a deeper understanding of sentiment. In addition, the research explores the complex domain of sarcasm detection by proposing a novel Sarcasm Detection Approach (SDA). This approach identifies sarcasm by analysing conflicts between textual content and the accompanying emojis.

The thesis addresses key challenges in sentiment analysis by evaluating and comparing emoji dictionaries and sarcasm detection approaches to enhance sentiment classification. Extensive experimentation on diverse datasets rigorously assesses the effectiveness of these methods in improving sentiment analysis accuracy and sarcasm detection performance, particularly in emoji-rich datasets. The findings highlight the crucial role of emojis as contextual cues, underscoring their value in sentiment analysis and sarcasm detection tasks.

The outcomes of this thesis aim to advance sentiment analysis methodologies by offering insights into preprocessing strategies, leveraging the expressive potential of emojis through the Emoji Dictionary (ED), and introducing the Sarcasm Detection Approach (SDA). The research demonstrates that integrating emojis through these tools substantially enhances both sentiment analysis and sarcasm detection. By utilizing these tools, the study not only improves model performance but also opens avenues for further exploration into the nuanced complexities of digital communication.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

DECLARATION OF AUTHORSHIP

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

Advancing Sentiment Analysis: Embracing Emojis, Sarcasm Detection, and Preprocessing Strategies with VADER and BERT on X Data

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. None of this work has been published before submission;

Signature:	Date:
Malak Alsabban	07/03/2025

Acknowledgements

First and foremost, I would like to praise and thank Allah. I am grateful for the strength, guidance, and countless blessings that I have received throughout this journey, and for that, I am eternally thankful.

I am profoundly thankful to my esteemed supervisors, Prof. Dame Wendy Hall and Prof. Mark Weal. Their invaluable mentorship, unwavering support, and insightful wisdom have been pivotal in shaping both this research and my academic journey. Their guidance has inspired me to pursue excellence with dedication and integrity, nurturing my development as a researcher. I am deeply grateful to Southampton University's Electronics and Computer Science Department for providing an intellectually stimulating environment that has significantly contributed to my academic journey. The department's educational atmosphere has enriched my doctoral experience.

My heartfelt appreciation extends to the Saudi Arabian government and Majmaah University for their generous financial support. This support was not just a contribution towards my educational expenses but a crucial investment in my potential as a researcher and academic.

To my parents, Abdullah Alsabban and Dalal Khawandana, my pillars of strength and love, I owe everything. Your unconditional love, endless sacrifices, and the values you've instilled in me from my first steps have been the guiding light of my life. Your belief in me and unwavering support have emboldened me to chase my dreams, no matter the hurdles. You have gifted me with the greatest treasures—my siblings. Their support and encouragement have been a source of strength and motivation throughout this journey. Their belief in my abilities and their constant cheer have been invaluable. Together, we have shared the joy of every achievement and the challenge of every obstacle, making our family bond the bedrock of my success.

My dear husband, Salem Shebaili, your enduring patience, profound understanding, and boundless love have been my haven. The steadfast support and sacrifices you've made resonate deeply within me, and I am endlessly grateful for all you do for our family. To my beloved children, Dalal and Abdullah, you are the heart of my universe. Your laughter and joy ignite my determination, shining a light on the beauty that surrounds us. Your presence turns every challenge into a worthwhile endeavour and imbues every achievement with deeper meaning.

This expression of gratitude goes beyond mere words, representing a heartfelt appreciation for everyone's contributions. With the highest praise reserved for Allah, I extend my sincerest thanks to everyone who has played a part in this journey.

List of Abbreviations

SA	Sentiment analysis
ED	Emoji Dictionary
SDA	Sarcasm Detection Approach
Prec.	Precision
Rec.	Recall
F1	F1-score
Orig.	Original
rm	remove
Op.	Opinion
mod	Modified
Mtd.	Method
Negative	NEG
Neutral	NEU
Positive	POS

Chapter 1 Introduction

Today, in the digital and interconnected world, an enormous amount of data is being produced every day. The explosion of social media applications gives users opportunities to share whatever they want in all data forms (videos, photos, and texts), and text is the most popular form. Social media data has become a valuable resource for researchers seeking to understand human behaviour, emotions, and opinions across a wide range of topics. One way to explore this is through sentiment analysis (SA) of social data in different domains. This data plays a vital role in all disciplines. Data is collected and analysed to help in making better decisions, either for businesses, the government, or individuals, by providing evidence and insights.

One of the tools used is sentiment analysis (SA), which is a form of natural language processing (NLP) that is used to determine people's feelings from their texts. People's emotions are complex; they are changing regarding the surrounding circumstances, and their emotions are affecting their behaviours and actions. Sentiment is defined as "what one feels about something," "personal experience, one's own feeling", "an attitude toward something," or "an opinion" (Farhadloo & Rolland, 2016). Sentiment analysis is not a trivial task; there are a lot of challenges and problems that need to be addressed, like sarcasm detection, negation handling, and spam detection. Notably, sarcasm detection has emerged as a key challenge, particularly in social media where short and informal text complicates accurate analysis (Pokhriyal & Jain, 2024). Sentiment analysis is used in many domains for many purposes, such as politics (Arista et al., 2024), crisis management (Villasor & Baradillo, 2024), customer experience (Arifiansyah, 2024), and business insights (Muntinova, 2023). There are different contexts in which sentiment analysis is used, like monitoring services (Md Saad et al., 2023), predicting students' performance (Obeleagu et al., 2019), and court records (Lovell et al., 2023).

There are two common approaches that are used for sentiment analysis purposes: lexicon-based and machine learning-based techniques (Patil & Gupta, 2015). These approaches differ in how they process text and generate sentiment scores. Understanding which approach to use depends on the context of the analysis and the need for accurate results. While lexicon-based approaches rely on pre-defined word dictionaries to measure sentiment, machine learning models (such as BERT) utilise vast datasets to learn sentiment patterns automatically. The accuracy of this study is measured by automatically attributing sentiment that corresponds to that recognized by humans.

Choosing the right preprocessing steps is critical for both approaches, as preprocessing ensures that the data is in the correct format for analysis. Preprocessing tasks, such as removing URLs, mentions, emojis, and stop words, can affect the accuracy of a sentiment

classifier. The different preprocessing methods will be used for different purposes. For example, if tracking the source of the content is a priority, URLs may not be relevant to retain, whereas if retaining information about interactions or relationships between users is important, mentions may be valuable to keep. Different preprocessing pipelines are required depending on the dataset, and selecting the best pipeline can result in improved performance for both lexicon-based and machine learning-based models (Biradar, 2024).

Emojis represent a unique aspect of social media data which people usually use to express their sentiments. Emojis carry emotional and contextual weight in posts, and the Statista website reported that the percentage of posts containing emojis from July 2016 to July 2021 reached 20.69% (Dixon, 2023). Given this substantial usage, removing emojis during preprocessing could negatively impact the analysis. Removing emojis is often one of the preprocessing steps. Where emojis are a common communication mechanism in text, this preprocessing step can have a notable impact on the classification results. To address this, this research proposes an Emoji Dictionary (ED), designed to capture the meaning of emojis in specific contexts, and an innovative Sarcasm Detection Approach (SDA), which leverages the interplay between emojis and text to improve sentiment analysis outcomes. This thesis builds upon the hypothesis that emojis contribute substantially to both sentiment analysis and sarcasm detection.

X has been chosen as a source for the collected data. X, the free microblogging application, is one of the most popular social networking applications where people can disseminate their opinions and ideas, especially in hashtags. On X, people share their beliefs, ideas, hobbies, concerns, etc. Posts could include text, videos, images, links, etc. It is more than a tool for tracking changes in people's opinions; it can also measure changes in people's behaviours because it is a reflection of social action on an international level (Mejova et al., 2015). The Statista website reported that the number of active users in the last three months of 2020 reached 187 million. X is a rich example of a source of data for sentiment analysis. Applying sentiment analysis to social network data is a crucial field of study. X has become the most popular social network that is used for this purpose. This research is about understanding sentiment on social media. X was chosen because it was a good exemplar of micro-blogging sites, popular, and convenient.

This thesis aims to explore the different subtleties of the two sentiment analysis approaches (lexicon-based and machine learning) by using three different case studies: the COVID-19 Vaccine, Vegetarianism, and Electric Cars. These topics have been chosen because they are trending, they are in different domains and have different characteristics, and all three datasets contain sentiments of different types with different levels and types of emoji use. The datasets will be collected approximately in the same period because the period of collecting the posts

affects the types and occurrences of emojis that could appear in the posts. The same experimental setup and experiments will be applied across the three datasets. In this study, the impact of emojis will be examined.

1.1 Research Purpose

The overarching goal of this thesis is to improve sentiment classification and sarcasm detection on social media platforms by incorporating emojis into the sentiment analysis pipeline. This study positions emojis as contextual indicators that can enrich sentiment interpretation and proposes innovative methods to detect sarcasm through the interaction between text and emoji use. The research assesses the performance of two distinct sentiment analysis techniques: lexicon-based (VADER) and machine learning-based (BERT) classifiers, with an emphasis on optimizing preprocessing steps, particularly emoji handling, on classification performance.

Rather than focusing on a single domain, this research leverages multiple datasets from different domains, allowing for greater generalization of the findings. By analysing sentiment across three diverse topics, the study ensures that the results are not limited to one specific context, increasing the robustness and applicability of the proposed Sarcasm Detection Approach (SDA) and Emoji Dictionary (ED).

1.2 Rationale

While some of the previous studies have removed emojis in the preprocessing stage, emojis are one aspect of social media data that is rich in emoji use, and their use is increasing day after day. Analysing emojis provides a qualitative layer to sentiment analysis, offering a more nuanced understanding of emotions, particularly in cases of sarcastic expression. This research exploits the existence of emojis in sentiment analysis and detecting sarcasm by creating an Emoji Dictionary (ED) and proposing a Sarcasm Detection Approach (SDA) for both approaches, lexicon-based and machine learning. It is likely that the contextual use of emojis may be cultural or related to particular demographics; the advantage of the Emoji Dictionary (ED) is that it could be adjusted based on that. This would also be helpful in the future if new emojis were released. The research evaluates three distinct datasets, each collected within the same time period and manually labelled by three annotators, and compares the performance of sentiment classifiers using these datasets to ensure the findings are generalizable across various contexts and not specific to only one dataset that has specific characteristics. The three datasets vary in terms of the nature of the topic, domain, ethical considerations, text length, emoji use, and imbalances in the distribution of sentiment classes. In most of the studies, they do not provide technical information about using the classifiers and how their choice would affect the classification

results, like the VADER threshold and BERT training dataset size. In this research, some optimisation issues are examined and discussed.

1.3 Research Questions

The primary aim of this research is to enhance sentiment classification performance and improve sarcasm detection in social media posts by incorporating emojis into the sentiment analysis pipeline. The main research question guiding this study is:

How do various factors, including dataset characteristics, preprocessing methods, and novel methodologies such as Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA), influence the accuracy and effectiveness of sentiment analysis and sarcasm detection in X data?

This research aims to address several key gaps in the current literature. Most studies on sentiment analysis do not adequately explore the effects of preprocessing steps on model performance, particularly when handling unique elements like emojis. Additionally, while sentiment analysis has been widely studied, the specific role of emojis in both sentiment classification and sarcasm detection has received less attention. To address these gaps, this thesis proposes an optimized preprocessing pipeline and introduces two novel methodologies—the Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA). Through these contributions, the research seeks to improve overall performance in both lexicon-based and machine learning-based sentiment analysis models.

To answer the main research question, the study will explore the following sub-questions:

RQ 1: How might different sentiment analysis approaches give different results depending on the nature of the datasets?

RQ 2: What is the significance of preprocessing steps in improving the classification performance of sentiment analysis and sarcasm detection models on X data?

RQ 3: To what extent do the proposed Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) improve sentiment classification performance?

1.4 Research Contribution

This thesis makes several notable contributions to the field of sentiment analysis and sarcasm detection in social media data. The primary contributions are outlined below:

- Incorporation of Emojis in Sentiment Analysis: A key contribution of this thesis is the incorporation of emojis in sentiment analysis. By creating an Emoji Dictionary (ED) and

integrating it into the sentiment analysis pipeline, this research enhances the accuracy and depth of sentiment interpretation, particularly in informal and expressive text found on social media platforms. This advancement addresses a major gap in existing sentiment analysis methodologies, where emojis are either ignored or treated with limited nuance.

- Development of a Novel Sarcasm Detection Approach (SDA): This thesis introduces a novel Sarcasm Detection Approach (SDA) that goes beyond traditional text-based sentiment analysis, which strategically harnesses a blend of lexical features, contextual insights, and the inclusion of emojis to identify sarcastic expressions within social media discourse. Emojis play a pivotal role in encapsulating the nuanced subtleties of sarcasm, acting as supplementary contextual indicators that enrich the model's comprehension of sarcastic intent. Through the incorporation of emojis, the Sarcasm Detection Approach (SDA) achieves greater robustness and sensitivity in sarcasm identification, particularly in cases where textual cues alone may be ambiguous or insufficient.
- Extensive Experimental Evaluations: To validate the effectiveness of the proposed methodologies, the thesis conducts extensive experiments across multiple datasets, including three diverse case studies (COVID-19 Vaccine, Vegetarianism, and Electric Cars). These evaluations compare the performance of the proposed ED and SDA with existing sentiment analysis and sarcasm detection models, using both lexicon-based (VADER) and machine learning-based (BERT) classifiers. The experiments are designed to assess not only the overall performance of the models but also the impact of different preprocessing methods.
- Practical Implications for Real-World Applications: The findings and methodologies presented in this thesis have practical implications for various real-world applications, including social media monitoring and sentiment analysis of customer feedback. By offering ED and SDA for analysing sentiment and detecting sarcasm using emojis, this research contributes to improving decision-making processes and enhancing user experiences on online platforms.
- Creation and Collection of Three Novel Datasets: Another key contribution is the creation, collection, and manual annotation of three distinct datasets—COVID-19 Vaccine, Vegetarianism, and Electric Cars. These datasets were meticulously labelled by three independent annotators, ensuring high-quality, reliable data for experimentation. Each dataset represents a different domain, providing a rich source of diverse data for sentiment analysis. These datasets serve as a valuable resource for

future research in sentiment analysis and sarcasm detection, addressing the relative scarcity of manually labelled, high-quality datasets in this field.

1.5 Thesis Structure

The remainder of the thesis is organised in the following way:

Chapter 2 – Background and Literature Review: Explores existing research on sentiment analysis, sarcasm detection, and the use of emojis in social media. Additionally, it provides an overview of existing emoji dictionaries and sarcasm detection methodologies, identifying key gaps that this thesis aims to address.

Chapter 3 - Research Methodology: Outlines the data collection methods, including a detailed description of the datasets used. It integrates insights from a pilot experiment on initial methods and preprocessing, focusing on testing and refining the approaches. The chapter also describes the methodology for creating the Emoji Dictionary (ED) and implementing the Sarcasm Detection Approach (SDA), laying the foundation for the experimental work.

Chapter 4 – Data Pipeline (Implementation): provides an in-depth overview of the data pipeline, including data acquisition, filtering, and labelling processes. It emphasizes the importance of preprocessing, with a focus on handling emojis, mentions, and URLs. The chapter also touches on optimization issues encountered with sentiment analysis tools like VADER and BERT during the preprocessing phase.

Chapter 5 – Emoji Dictionary (ED): presents the development of the Emoji Dictionary (ED), a specialized lexicon designed to interpret the sentiment of emojis in social media content. A comparative analysis evaluates the performance of the ED against existing emoji dictionaries across three datasets: COVID-19 Vaccine, Vegetarianism, and Electric Cars. Performance metrics such as precision, recall, F1-score, and accuracy are used to demonstrate the advantages of the proposed dictionary in handling emoji-rich content.

Chapter 6 – Sarcasm Detection Approach (SDA): focuses on comparing sarcasm detection methods, with a particular emphasis on the proposed Sarcasm Detection Approach (SDA). The SDA utilizes the interplay between text and emojis to identify sarcasm more effectively. The performance of the SDA is evaluated and compared to existing sarcasm detection models that do not account for emojis, demonstrating its improved accuracy and utility in sentiment classification tasks.

Chapter 7 - Applying ED and SDA Across Multiple Datasets: applies the combined Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) to diverse datasets. The results are

analysed to assess the generalizability and robustness of the proposed methods across datasets with varied linguistic styles and content, further validating their effectiveness in handling the complexities of sentiment analysis and sarcasm detection in social media.

Chapter 8 – Discussion: Discusses the key findings from the experimental work, highlighting the importance of emojis in sentiment analysis and sarcasm detection. It addresses the optimization issues of VADER and BERT and reflects on the role of preprocessing, particularly the impact of the ED and SDA, in enhancing sentiment classification performance.

Chapter 9 – Conclusion and Future Work: Summarizes the contributions of the thesis and provides an overview of the key findings. It outlines potential future research directions.

Chapter 2 Background and Literature Review

2.1 Introduction

Medhat et al. (2014) define sentiment analysis, or opinion mining, as "the computational study of people's opinions, attitudes, and emotions towards an entity." Sentiment analysis is basically a classification procedure that works at different levels (document, sentence, and aspect level) to determines the sentiment (positive, negative, or neutral). H. M. K. Kumar & Harish (2020) describe sentiment analysis as "a process of automatically extracting opinions or emotions from text, especially in user-generated textual content."

The detection of sarcasm in textual communication present a challenge on the domains of sentiment analysis. Recent research has demonstrated that sarcasm frequently involves a contrast between literal expressions and intended meanings, making it difficult for computational models to detect. For instance, Ghosh & Veale (2016) explore the efficacy of neural networks in detecting sarcasm, indicating the importance of finding a computational approaches in tackling this challenge. Similarly, Felbo et al. (2017) demonstrate how the analysis of emoji usage can provide valuable insights into detecting sentiment, emotion, and sarcasm. It suggests that emojis in text can play a crucial role sentiment analysis and sarcasm detection. The work by Riloff et al. (2013) further underscores the complexity of sarcasm detection by examining the contrast between the sentiments and the situations as a feature of sarcastic text. Additionally, Bamman & Smith (2015) focus on the role of contextual information for accurate sarcasm identification on social media platforms. These studies collectively underscore the multifaceted challenges of sarcasm detection in text-based communication, a field where the interplay of context not sentiment just computational but also a deep understanding of human expression.

Emojis are a digital communication tool that bridges the gap between text and emotional expression. They can convey attitudes, tones, and feelings that are difficult to express when face-to-face interaction isn't present. Kralj Novak et al. (2015) highlight the potential of emojis to influence the perceived sentiment of text messages, enriching the text with layers of emotions. Felbo et al. (2017) expand on this idea by using emoji to develop deep learning models that can detect sentiment, emotion, and sarcasm in text. All these illustrate the impact of emojis in digital communication and how emojis can serve as potent indicators of underlying emotional states and intentions.

2.2 Background

2.2.1 Sentiment Analysis: Concepts and Applications

Sentiment analysis is a vital tool for understanding, identifying, and categorising the emotions present in written language. It is a well-established field within natural language processing (NLP). This field is devoted to extracting textual material in order to identify the authors' feelings and sentiments, which range from positive and negative to neutral. Its application extends to various digital channels, including social media exchanges, user reviews, forum posts, and news. As a result, companies, scholars, and legislators can use it to gain insight into public opinion, consumer preferences, and cultural trends. These perspectives are invaluable for understanding human behaviour, market dynamics, and societal movements (Liu & Zhang, 2012; Pang & Lee, 2008). This capability represents a critical first step in fusing computational intelligence with the complexities of human emotional expression. It improves customer relationship strategies and product development while also augmenting the collective understanding of intricate social phenomena through the sentiment aggregate (Cambria et al., 2013).

Sentiment analysis has many uses, but social media monitoring is one of the most important ones. Users now have the freedom to share their ideas in text, pictures, and videos, with text being the most popular format thanks to the widespread availability of social media platforms (Sailunaz & Alhajj, 2019). Researchers use this to analyse society and its position on particular issues while keeping an eye out for trends. One technique is to apply sentiment analysis to social data from a variety of fields. Applications of sentiment analysis can be found in the travel and hospitality sector (Tepavčević et al., 2023), politics (Ceron et al., 2014), and healthcare (Kaur et al., 2024), demonstrating its broad utility in managing public opinion and sentiment within specific sectors.

2.2.2 Challenges in Sarcasm Detection

Sarcasm, with its blend of language and cognitive subtleties stands out as an aspect of human communication. It involves a contrast between the meaning and the intended message of statements. This type of irony is not for humour but also serves as a way to critique and comment on society. Recognizing and interpreting sarcasm requires an understanding of cues and shared context. The core of sarcasm often lies in the difference between the words used and the negative feelings meant by the speaker or writer a theme extensively explored in literature (Gibbs, 2000; Joshi et al., 2017). Indicators like intonation in speech or specific punctuation and capitalization in writing can signal sarcasm though their effectiveness can vary

across environments and cultures (Bryant, 2010; Burgers et al., 2012). Wallace et al. (2014) highlight the role of context in detecting irony and sarcasm, suggesting implications for computational approaches to sarcasm detection. Sarcasm is a topic of continuous interest and research in various domains like linguistics, psychology, and computing because of the interaction between linguistic form, contextual cues, and cognitive processing.

2.2.3 The Emergence of Emojis in Digital Communication

A new generation of emoticons were introduced in Japan called Emojis, and they are another form of emoticons but in 2-D (Kelly & Watts, 2015). Emoji is represented using Unicode characters. Emojis are used to express different kinds of sentiment and play a major role in sentiment classification. Previously, the emoticon (sequence of characters) was used more, but now the emoji has beaten the emoticon, and people on social media tend to use emojis more. About 92% of online users use emojis (Daniel & Camp, 2020).

Emoji include faces with different emotions, flags, animals, plants, careers, food, and much more. With each updated version of the emoji, new emojis are released. There is a lack of interest among the researchers in studying the effect of emoji in NLP (Chen et al., 2018; Pavalanathan & Eisenstein, 2015). Churches et al. (2014) from the school of psychology at Flinders University reported that when people react to emojis as actual faces, it becomes more essential than people expected. In sentiment analysis, mostly emojis are removed (Shiha & Ayvaz, 2017).

Emojis have become essential in conversations adding depth and emotion to text. These vibrant symbols go beyond language differences providing a shared way to convey feelings, responses and attitudes. Emojis are becoming more popular on social media, and they have different uses. They can be used to clarify or enhance the sentiment or to complete the meaning of the text (Peacock & Khan, 2019). Some words can be replaced with emojis; they can convey brief meanings or sentiments (Barbieri et al., 2017). Using emojis can make the expression process much easier and communication faster (Shiha & Ayvaz, 2017). Emojis can be used as codes that refer to something with a shared meaning within a specific society. The text can also be made more lighthearted or funny with emojis (Pavalanathan & Eisenstein, 2015). They can be used for sarcasm or simply to decorate the text (Pohl et al., 2017). The existence of emoticons helps in solving the problem of difficulty conveying their feelings when interacting via text (Walther & D'addario, 2001).

2.3 Sentiment Analysis Approaches and Methodologies

2.3.1 Traditional Sentiment Analysis Techniques

For sentiment analysis purposes, there are two widely used automated approaches: lexicon-based and machine learning.

2.3.1.1 Lexicon-Based Approach

The lexicon-based approach uses a dictionary; it builds its sentiment decision on a created list of words or phrases (called the lexicon) and their labels. This list could be created manually or automatically. Most of these lists are adjectives because they usually carry more sentiment (Taboada et al., 2011). This approach is the most commonly used for sentiment analysis purposes with social media data (Drus & Khalid, 2019). Some researchers prefer this approach more because it's more flexible than the machine learning approach, and the user can manipulate or create a dictionary (Trivedi & Singh, 2021). But it has limitations; the sentiment of the same word could differ from one text to another and from domain to domain (Mejova et al., 2015). A lexicon-based approach works well in sentiment analysis tasks except for text that contains emojis, abbreviations, and informal words that are used too much on X. Because of that, a lexicon-based approach has a low recall of text that contains these. One of the solutions is to add these words to the dictionary, but the problem is that new words always appear. In addition, the polarities of these new words are hard to determine because they are mainly based on context (Zhang et al., 2011). Lexicon-based methods could be built manually or automatically; for that reason, they are unreliable (Taboada et al., 2011).

One of the lexicon and rule-based techniques is the Valence Aware Dictionary for Sentiment Reasoning (VADER). It is a sentiment analysis tool; it contains lexicon features that are labelled according to their sentiment (positive or negative), and beside the polarity, it shows how strong the emotion is (intensity). VADER takes a text and returns metric values that represent the negative, positive, neutral, and compound (that represent the dominant sentiment) scores. VADER is an efficient tool that works with large data sets; it is fast to implement and easy to extend; it performs well in social media; and it supports emojis (Bonta et al., 2019). Hutto & Gilbert (2014) compared VADER, a lexicon-based technique, against seven other lexicon-based techniques: LIWC, GI, ANEW, SWN, SCN, WSD, and Hu-Liu04. They found that VADER has a big difference, which is that it performs better in the social media area. And it has such high accuracy in classification that it even exceeds the result of human classification. VADER is able to quickly classify large amounts of data (Elbagir & Yang, 2019). In another study, M. A. Al-Shabi (2020) also found that VADER has the best performance when compared with the other four

lexicon-based techniques: SentiWordNet, SentiStrength, the Liu and Hu opinion lexicon, and AFINN-111. By using VADER, there is no need to do preprocessing; it can deal with emojis, capitalization, extended punctuation, and stopwords (Malde, 2020). VADER allows working with slang, abbreviations, emoticons, and emojis (강아미, 2021). VADER could misinterpret sarcasm, and spelling and grammar mistakes could cause misinterpretation (DeLancey, 2020). Also, the hashtag symbol doesn't change the sentiment score (e.g., score 'good' = score '#good' = 0.4404). ÇILGIN et al. (2022) in their study, applied lemmatization and removed stopwords before applying VADER.

Linguistic Inquiry and Word Count (LIWC) is a text analysis tool that is used to measure sentiments. Its work is based on counting the frequency of words in a specific category. It includes 74 categories and 4500 words that are categorized into one or more categories. It works based on the word count. Researchers could choose some of the categories that they are interested in in their studies (Hancock et al., 2007). It is used in the social media domain to extract emotions and polarities. It has been used by psychologists, sociologists, and linguists and in senatorial speeches to calculate the sentiment; it has also been used to measure the change in the sentiment of pregnant mothers from their posts (De Choudhury et al., 2013; Hutto & Gilbert, 2014). LIWC does not support slang, emoticons, and abbreviations, so it is less sensitive in social media to sentiment expression. It works based on the count of the word, not the intensity of the word. For example, 'exceptional' and 'ok' are both considered positive words without any respect to their intensity (Hutto & Gilbert, 2014).

General Inquirer (GI) is one of the oldest lexicon-based approaches that is developed manually by using existing dictionaries. It is used by sociologists, political scientists, and psychologists. It includes 183 categories and 11,000 words (1,915 positive and 2,291 negative). It has a shortage of words that are usually used in social media to express sentiment. It doesn't determine the intensity of the sentiment (Taboada et al., 2011).

Affective Norms for English Words (ANEW) provides sentiment ratings for 1,034 words. Each word is assigned a polarity from 1 to 9. 5 is considered neutral, less than 5 is negative, and more than 5 is positive. It is not sensitive to the sentiment lexicon features that are popular in the social domain (Hutto & Gilbert, 2014).

SentiWordNet is derived from WordNet, the lexical database. It contains 147,306 synsets that are labelled with three scores (positive, negative, and neutral) in the range from 0 to 1. For each synset, the total sum of the three scores will equal 1. Most of the synsets do not have negative or positive polarities. It doesn't work well for finding sentiment in microblogs (Hutto & Gilbert, 2014).

In a comparative study of lexicon-based classifiers, it found that VADER achieved higher precision, recall, f1-score, and accuracy when it compared with Text blob and NLTK (Bonta et al., 2019).

In another comparative study, five lexicon classifiers performance were compared: VADER, SentiWordNet, SentiStrength, Liu and Hu lexicon and AFINN-111. The experiment was conducted on two datasets, in both datasets VADER achieved higher accuracy (M. Al-Shabi, 2020).

2.3.1.2 Machine Learning

The machine learning approach is also called a non-lexical approach. It has training and testing data sets. A training dataset that is close to the data that wants to be sentimentally analysed to train the algorithm contains input vectors and their labels. Then, a testing dataset is used to examine the performance of the classifier in predicting the right labels. Naïve Bayes and Support Vector Machine are the most popular machine learning approaches because they are faster than others in their performance (Kirilenko et al., 2018). The input in machine learning techniques could be unigrams, bigrams, or N-grams that describe the number of input words. The machine learning algorithm is classified into one of three categories: supervised: the train dataset is prelabelled, which helps in producing reasonable output; unsupervised: the train dataset is unlabelled, and it conducts clustering. and semi-supervised: the train dataset is both labelled and unlabelled (Ahmad et al., 2017; Gautam & Yadav, 2014; Neethu & Rajasree, 2013).

Naive Bayes (NB) is a popular supervised machine learning algorithm. It classifies the data based on calculating its probability by using Baye's theorem (Ahmad et al., 2017). From the training dataset, it calculates the occurrence of the words, and based on that, it determines the sentiment of the other sentences (Smeureanu & Bucur, 2012). In one of the studies, the accuracy of applying the classification model to 5000 sentences divided into two groups was 79.99% (Singh & Husain, 2014). In another study, they classified 5000 sentences into two groups, and the accuracy was 81.43% (Smeureanu & Bucur, 2012). NB is efficient for large data sets. Its implementation and interpretation are not difficult (Ahmad et al., 2017). It is the most appropriate for text classification (Singh & Husain, 2014). It's fast, and the result that it produces is good (Smeureanu & Bucur, 2012). On its features, it works with independent assumptions. Researchers discover some challenges that need to be solved (Ahmad et al., 2017).

Support Vector Machine (SVM) is a supervised machine learning model that is used for classification and regression. That learns from examples (training) to predict the classification for unclassified data (Noble, 2006). Sentiment analysis is one of the classification problems that SVM could be used to solve. SVM could be used with term-weighting schemes to enhance its

performance. Zainuddin & Selamat (2014) found that SVM has a higher accuracy in sentiment analysis when used with TFIDF. With regard to accuracy and efficiency, SVM is the best (Ahmad et al., 2017). It works based on finding the largest hyperplane margin by using the training dataset that is already classified into separate clusters, then finding the test data on which side it falls in the hyperplane. As long as the margin is large, the chance of misclassification will decrease (Khairnar & Kinikar, 2013).

Maximum Entropy (ME) It is a feature-based model, and it doesn't make independent assumptions for its features (Ahmad et al., 2017). It means it is possible to add features (e.g., phrase, bigram) and that will not cause any feature overlapping, and that's unlike NB (Go et al., 2009). It estimates the probability distribution. It used a training data set to set constraints (that state the data set characteristics), and that's by determining the features first, then measuring their values based on their occurrences in the training dataset. They found that ME has worked better than NB in two out of three data groups (Nigam et al., 1999). The problem of feature overlapping could be handled better by using ME more than NB (Ahmad et al., 2017). If the features are not selected carefully, that could affect the quality of the result (Nigam et al., 1999).

BERT (Bidirectional Encoder Representations from Transformers) is one of the trained transformer models developed in 2018 by Google. BERT trained on 2500 million words in Wikipedia and 800 million words in different books (Agrawal et al., 2021; Kenton & Toutanova, 2019). BERT has had a notable impact on the field of NLP since it was published (Rogers et al., 2021). BERT was trained by using two different tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). With MLM, they masked 15% of the words randomly, generated a training sample, and learned to predict masked words. With NSP, they use pairs of sentences to train the model and learn to predict the second sentence. Through these two tasks that are used in pre-training BERT, researchers were able to get a very effective language model. However, for other tasks requiring the use of BERT, fine-tuning is necessary. This involves adding another layer at the end to train the model for a specific task using a dataset tailored to that task's objectives. Other tasks include sentiment analysis, question answering, text classification, and named entity recognition. At the beginning, BERT is initialized by using pre-trained parameters, and then these parameters are updated and fine-tuned using the labelled dataset. There are two versions of BERT: BERTBASE and BERTLARGE, each with a different number of encoder layers (L), hidden units (H), and number of self-attention heads (A). BERTBASE : L = 12, H = 768, and A = 12; BERTLARGE: L = 24, H = 1024, and A = 16 (Kenton & Toutanova, 2019).

Wang et al. (2012) applied Naïve Bayes and logistic regression by training the model using the same large dataset that has about 2.5 million posts, and the maximum accuracy attained was

65.57%. Neethu & Rajasree (2013) compared the performance of Naive Bayes, SVM, and Maximum Entropy by classifying posts into positive and negative classes. The dataset is balanced and has 1200 posts in total; they used 83.3% for training and the remaining posts for testing the classifiers. They found that all of them were having similar performances. However, Naive Bayes has less recall and accuracy, but it has better precision. The accuracy of Naive Bayes is 89.5%, and the accuracy of SVM and Maximum Entropy is 90%. A. Kumar & Sebastian (2012) reported that SVM outperformed Naive Bayes and MaxEnt models and that using unigrams as its feature was more efficient. Because of its excellent performance, it is widely used. It has many extensions that make it more efficient and flexible (Ahmad et al., 2017). Based on many studies that have compared the performance of the text classification methods, SVM's performance outperforms many of the methods. It is efficient and accurate even with a small training data set; it is even better than NB (Khairnar & Kinikar, 2013).

In a comparative study on various machine learning and deep learning techniques for sentiment analysis, there are SVM, naive bayes, LSTM, and BERT. A publicly available dataset is used; it has more than 1.6 million posts (7,98,988 positive posts and 8,01,011 negative posts.). The split ratio of the training and testing datasets is 80:20. BERT achieved higher precision, recall, f1-score, and accuracy (Dhola & Saradva, 2021).

In another comparative study, seven machine learning models their sentiment analysis performance in classifying the posts as positive, negative, or neutral was compared. The models are: Random Forest, XGBboost, Logistic Regression, Support Vector Machine, SGD Classifier, Decision Tree, and BERT. From the results, it was found that BERT is the best model for sentiment analysis tasks. It achieved higher precision, recall, f1-score, and accuracy (T. S. S. Kumar et al., 2021)

In a review of comparisons, six sentiment classification models on Amazon consumer reviews, three lexicon-based techniques (VADER, Pattern, and SentiWordNet), and three machine learning approaches (SVM, Gradient Boosting, and LR) It found that the performance of all the machine learning models surpassed the performance of lexicon-based techniques. Among the lexicon-based classifiers, VADER achieved the highest classification performance in all metrics (Nguyen et al., 2018).

2.3.2 Sentiment Analysis with Emojis

Despite their widespread use as emotional indicators in digital communication, emojis do not always perfectly capture the tone of the text they are used with. The overall message may be difficult to understand if there is a difference between the emojis used and the text's intended sentiment. Emojis serve a variety of functions, including sarcasm, irony, and other complex

communicative goals in addition to being used for basic sentiment expression, as research has shown.

Emojis, often considered universal in their communication, can carry diverse meanings across different cultures, leading to varied interpretations and usage patterns. For instance, the sleeping face emoji, commonly conveying a sense of sleep or rest, is understood differently in various cultures. While Malays interpret it as a signal to 'do not disturb,' the Chinese perceive it as a cue to 'ignore' the message. Similarly, the loudly crying face emoji evokes contrasting sentiments: in China, it may signify 'disbelief,' 'disappointment,' or 'overwhelming happiness,' while in India, it may express 'help' or 'regret.' Additionally, the face with tears of joy emoji elicits multifaceted responses: Malays view it as indicative of 'cry' or 'why,' Chinese perceive it as 'awkward,' and Indians associate it with 'lame.' These examples illustrate how cultural nuances strongly influence emoji usage and highlight the importance of considering cultural context in emoji interpretation and communication(Amalina & Azam, 2020).

One of the ways to analyse the emojis' sentiment is by labelling them manually, but their use is very contextual; they could be used in ways other than their original use, causing false classification. For example, a crying face might usually be considered a negative sentiment, but there are some uses where it conveys a positive sentiment in the text (e.g., "The dress is so pretty 😭"), so emojis do not always convey the same sentiment as the text (Chen et al., 2018; Felbo et al., 2017). Another way is to replace each emoji with its equivalent meaning; one limitation of this method is that the usage of the emoji could change over time (Felbo et al., 2017).

Kralj Novak et al. (2015) found the existence of emojis play a important role in identifying the sentiment of the posts. There is more agreement between the annotators when they are labelling the posts with emojis.

These studies highlight how emojis play an integral role in digital communication. Emojis are helpful tool for conveying emotion, but they can be difficult to understand in digital messages due to their ambiguous interpretations and complex relationships with text.

2.4 Selection of X as the Source of Data

With an extensive record of online human behaviour and social interaction that is open to the public, X has solidified its place as the "socioscope" for social scientists. It has become the unrivalled platform of choice for obtaining precise, time-stamped records at the individual event level. With billions of digital footprints from social interactions accumulated on a daily basis, X offers an opportunity for collecting observational data. It is massive and microscopic, with the

ability to time-stamped record every micro-interaction. This rich repository improves the ability to notice behavioural shifts, understand the organisation of social networks, and look back and examine what is leading to important events, in addition to offering a thorough chronicle of daily life, social dynamics, and the ebb and flow of relationships. In addition, X data has proven invaluable for researching economic behaviour, including tracking consumer confidence and unemployment as well as social mood, investor sentiment, and market trends. It has also proven useful for keeping an eye on public opinion as it is expressed in political discourse. X will always be a valuable source of information about human behaviour, even in spite of future technological advancements, as it grows and becomes more and more ingrained in daily life. This is because X has the capacity to mediate a wide range of everyday interactions, including social networking, news consumption, and various personal activities (Mejova et al., 2015).

2.5 Sarcasm Detection Techniques

Sarcasm is a subtle kind of verbal irony that requires context awareness and the ability to recognise certain linguistic cues. Saying the opposite of what is meant is a common technique in sarcasm. Many computational methods have been developed recently to identify sarcasm in written texts automatically by using context and linguistic cues.

2.5.1 Linguistic Features and Indicators of Sarcasm

Sarcasm frequently uses linguistic cues like hyperbole, rhetorical questions, or a particular tone that shows insincerity. Sarcasm and other verbal irony require a contrast between the literal and intended meaning, which is frequently expressed through stylistic cues in writing or speech tonality (Gibbs, 2000).

They investigate the conditions necessary for the successful communication and perception of sarcasm. The factors that are essential to understanding sarcasm, such as the surrounding environment or the situational context, along with the larger cultural and situational contexts that affect how sarcastic remarks are interpreted across various social and cultural backgrounds, vocal cues or particular stylistic elements in writing could also be highlighted as crucial indicators that signal the presence of sarcasm. In addition, the speaker's intentionality as well as the listener's perceptual sensitivity may be considered important elements in the intricate interaction of variables that support the identification and understanding of sarcastic irony. This indicates an interplay of many factors that determine the effectiveness of the transmission and reception of sarcastic intent (Campbell & Katz, 2012).

Joshi et al. (2017) review a number of computational methods that use algorithms to detect sarcasm in text by analysing its features. They noted three key elements: the identification of sarcastic patterns, the use of hashtags as sarcastic cues, and the use of contextual data.

2.5.2 Role of Emojis in Sarcasm Detection

Sarcasm is one of the challenges in sentiment analysis. It implies negative sentiment by using non-negative words. Most of the existing approaches use the text to detect if there is sarcasm or not (Joshi et al., 2017). In detecting sarcasm, some approaches look for the user's previous activities to understand the user's characteristics. In most of the studies, they train a model to identify sarcastic posts. Emojis could be used to help in identifying sarcasm. Emojis are frequently used in social media instead of using emotion signals in real-world situations. They could notably improve the detection of sarcasm (Subramanian et al., 2019).

Incorporating emojis into sarcasm detection models has proven beneficial. For instance, Prasad et al. (2017) created a slang and emoji dictionary and used them to help define sarcastic posts. In the preprocessing stage, they replace all the emojis with their labels and all slangs with their meanings using the dictionaries. In their emoji dictionary, they collect only emojis that carry positive or negative sentiments. They tested how using emoji and slang dictionaries would help them identify sarcastic posts. They applied six various algorithms to 2000 manually labelled posts: Random Forest, Gradient Boosting, Decision Tree, Adaptive Boost, Logistic Regression, and Gaussian Naïve Bayes. They applied each algorithm three times by using different splits of training to testing dataset: 60:40, 70:30, and 80:20. In 33.33% of their experiments, the accuracy decreased. The maximum increase in accuracy was 8.22% when they used logistic regression with a 60:40 split ratio.

This study highlights the potential of combining emojis with textual content for more accurate sarcasm detection, showing that while emojis play a crucial role in sentiment expression, they can also signal sarcastic undertones when used in specific contexts.

2.5.3 Related Work

Several studies have focused on advancing sarcasm detection methodologies, each offering distinct computational approaches, dataset collection techniques, and challenges in identifying sarcasm, particularly in short-text social media environments.

Ghosh & Veale (2016) investigated the challenges of sarcasm detection through the lens of neural networks. Their dataset consisted of 39,000 tweets, which were labelled as either

sarcastic or non-sarcastic, with a focus on sarcastic tweets being collected using hashtags like #sarcasm and #yeahright. For testing, they manually annotated 2,000 tweets. The primary method they employed was Support Vector Machines (SVM), with their performance evaluation relying on the F1-score, which assesses the balance between precision and recall. One of the notable contributions of their work was the exploration of feature sets and different machine learning techniques, showcasing that SVM performs well when coupled with neural networks for sarcasm detection. Their approach, while innovative at the time, did not explicitly address the nuances of sarcastic language beyond text cues. While the study offered insights into the use of neural networks and traditional machine learning models, it primarily focused on textual cues, not incorporating other modalities such as emojis or contextual information beyond the text.

P. Kumar & Sarin (2022) presented a more sophisticated model named WELMSD (Word Embedding and Language Model-Based Sarcasm Detection), which integrates word embeddings and deep learning approaches to capture the complexities of sarcasm in social media text. Their approach combined FastText embeddings with BERT, a transformer-based language model, to offer better context-aware classification. One of the strengths of their research was the focus on improving sarcasm detection by using contextual information, a crucial element given that sarcasm often depends on the surrounding linguistic environment. Their evaluation involved using multiple benchmark datasets, and they reported their results using metrics like misclassification rates and AUC (Area Under the Curve). WELMSD outperformed traditional sarcasm detection models by leveraging both shallow word embeddings and deep contextual language models, indicating the importance of combining multiple layers of linguistic information for more accurate sarcasm detection.

Subramanian et al. (2019), in their study "Exploiting Emojis for Sarcasm Detection," aligns closely with the focus of this thesis, as it emphasizes the role of emojis as key indicators of sarcasm in digital communication. The majority of existing sarcasm detection algorithms primarily focus on text, overlooking the emotional and contextual signals carried by emojis. To address this gap, the authors proposed the ESD (Emoji-enhanced Sarcasm Detection) framework, which captures both textual and emoji signals to improve sarcasm detection accuracy.

In the ESD framework, the text encoder converts the textual information into numerical representations using word embeddings, enabling the model to map the relationships between words and understand sentiment more effectively. Similarly, the emoji encoder performs the same function for emojis, transforming them into numerical representations that the model can process. The final component of the framework, the sarcasm prediction model, combines the outputs from the text and emoji encoders. Using a machine learning model trained on labelled

sarcasm data, the system predicts the likelihood of sarcasm by integrating both textual and emoji-based signals.

In conclusion, the Sarcasm Detection Approach (SDA) and Emoji Dictionary (ED) presented in this research offer a comprehensive framework for enhancing sentiment and sarcasm detection in social media communication. By integrating both text and emoji signals, the approach ensures a broader classification spectrum (positive, negative, neutral), enabling more accurate and flexible sentiment analysis across various contexts and domains. SDA's adaptability makes it a valuable tool for detecting nuanced emotional content, particularly in scenarios where traditional text-only approaches may miss or misinterpret underlying sentiments. Furthermore, SDA can be used as an additional layer alongside either lexicon-based methods or machine learning models. This allows it to enhance sentiment analysis by leveraging both linguistic and contextual cues. Through this multi-dimensional strategy, the approach addresses the need for more accurate and versatile tools for sentiment and sarcasm analysis, offering deeper insights into the complex nature of online interactions.

2.6 Emoji Dictionaries in Sentiment Analysis and Sarcasm Detection

Dictionaries are a fundamental tool in sentiment analysis, offering a predefined set of words or phrases mapped to sentiment values. These approaches allow for efficient and straightforward sentiment analysis, as they classify individual words or phrases based on established sentiments. However, as digital communication has evolved—particularly with the rise of emojis in social media—there has been a growing need to incorporate emojis into sentiment analysis frameworks. Emojis are now essential in conveying emotions, tones, and even sarcasm, making their inclusion crucial for a more accurate analysis of sentiment and sarcasm in online environments.

Emoji dictionaries, designed to interpret and quantify the sentiment conveyed by emojis, have become essential in addressing the gaps left by traditional lexicon-based sentiment analysis. Emojis can either enhance or change the meaning of a message, thus playing a pivotal role in both sentiment and sarcasm detection. Without an understanding of emojis, sentiment analysis models may misinterpret the sentiment of posts, especially on social media, where brevity and non-verbal signals like emojis dominate the communication landscape.

Several prominent emoji dictionaries have been developed to bridge this gap. VADER, a popular sentiment analysis tool, extends its lexicon-based approach to emojis by converting them into their corresponding textual descriptions before performing sentiment analysis. This allows VADER to account for the influence of emojis on overall sentiment, treating emojis as words that

contribute to the emotional tone of a message. By converting emojis into textual representations, VADER integrates them into traditional sentiment models (Othman et al., 2022)

Demojize offers a different approach by converting emojis into their CLDR short names, which are standardized textual descriptions of emoji meanings. This method enables conventional sentiment analysis systems to process emojis as text, simplifying their integration into pre-existing sentiment analysis pipelines. Demojize highlights the importance of translating visual symbols into a textual form that can be easily interpreted by lexicon-based tools, making it a versatile option for emoji sentiment analysis (Gupta et al., 2021).

Emojinet represents one of the most comprehensive emoji dictionaries available. It connects Unicode emoji representations to their corresponding English translations, creating an extensive machine-readable inventory of emoji senses. Emojinet's scope allows for a deep understanding of how emojis are used in online communication, offering valuable insights into the sentiment conveyed by different emoji symbols. This extensive database helps to map emojis to their most accurate interpretations, ensuring that their emotional and contextual weight is appropriately captured (Wijeratne et al., 2017).

Each of these emoji dictionaries contributes to the growing need for accurate interpretation of emojis in digital communication. By linking emoji use to their corresponding textual descriptions or meanings, these tools help ensure that the full emotional intent of the message is captured, reflecting the increasing dominance of emojis in online communication.

2.7 Data Preprocessing in Sentiment Analysis and Sarcasm Detection

To do sentiment analysis, one of the important steps after collecting the data is preprocessing, which is underestimated in some of the studies. Preprocessing plays a role in the result of sentiment analysis; it ensures that the data is clean, consistent, and ready for analysis, which can lead to improved accuracy of the classification result (Angiani et al., 2016; Haddi et al., 2013).

Jianqiang (2015) reported that using different methods of preprocessing produces different classification performance. They proposed a preprocessing pipeline that included removing URLs, numbers, repeated letters, negation handling, and expanding acronyms. The study found that expanding acronyms and replacing negation increased the accuracy of the classification, while removing stopwords, URLs, and numbers barely affected the accuracy.

Saif et al. (2014) measured the impacts of removing stopwords, and they found that keeping the stopwords would enhance the performance of the classification.

(Bao et al., 2014) found that the accuracy of the classification is increased by using URL features reservation, negation transformation, and repeated letter normalisation, and reduced by using stemming and lemmatization in the preprocessing phase.

Go et al. (2009) suggested removing the repeated letters; if the letter is repeated more than twice, it will be replaced with just two occurrences.

Angiani et al. (2016) compared the accuracy of using five different preprocessing methods on the same datasets separately, with all five methods, and without any of them. Also, they have done basic operations and cleaning of the data before applying the preprocessing method, like removing URLs, hashtags, and punctuation. They found that higher accuracy is reached when the basic and stemming methods are used.

These studies highlight the importance of selecting appropriate preprocessing methods based on the characteristics of the dataset and the goals of the analysis. By refining preprocessing techniques, researchers can achieve better results in sentiment analysis and sarcasm detection tasks.

2.8 Evaluation Measures

To evaluate the performance of sentiment analysis and sarcasm detection models, several key metrics are commonly used, including accuracy, precision, recall, and F1 score. These metrics provide insights into the effectiveness of the model in correctly classifying sentiment and sarcasm.

- Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations.
- Recall is the ratio of correctly predicted positive observations to all actual positive observations.
- F1 score is the harmonic mean of precision and recall, providing a single measure of a model’s balance between these two metrics. It is particularly useful when both false positives and false negatives are important.

In this study, these metrics will be calculated using Python's sklearn.metrics library. A high F1 score suggests that the model effectively balances precision and recall, resulting in a low rate of false positives and false negatives (Alakus & Turkoglu, 2020).

2.9 Conclusion

This chapter provided an in-depth review of the existing literature on sentiment analysis, sarcasm detection, and the role of emojis in digital communication. The discussion highlighted various sentiment analysis approaches, including lexicon-based and machine learning-based methods, and their strengths and limitations. The complexities of sarcasm detection were also explored, emphasizing the importance of context, linguistic features, and emojis in improving detection accuracy.

The inclusion of emojis in sentiment analysis presents unique challenges and opportunities, particularly in identifying sarcasm, where textual cues alone may not be sufficient. The development of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) introduced in this study aims to address these challenges by providing a structured method for incorporating emojis into sentiment analysis pipelines.

Additionally, the importance of preprocessing in sentiment analysis was underscored, as it notably impacts model performance. Different preprocessing techniques were discussed, offering insights into how they can be tailored to improve sentiment classification accuracy.

Due to its many features and benefits, VADER was selected in this study as a representative of lexicon-based classifiers. In addition to giving polarity scores to individual words, VADER also takes into account the degree of sentiment connected to each word. This makes it possible to represent sentiment in text in a more complex way. The punctuation, emoticons/emojis, informal language, slang, and abbreviations present in social media content are all handled by VADER. VADER does sentiment analysis fast and with computational efficiency. VADER can be easily found in widely used NLP libraries, like Python's NLTK (Natural Language Toolkit).

BERT was selected as the representative machine learning-based model for a number of reasons. BERT considers the entire context of a word within a sentence. BERT uses a bidirectional transformer architecture; it takes into account a word's left and right context when it is being trained. Due to its extensive pre-training on a vast corpus of text data, BERT is able to acquire sophisticated and broadly applicable language representations. BERT can identify a variety of linguistic patterns and relationships. The pre-trained representations of BERT can be adjusted for particular tasks like named entity recognition, sentiment analysis, and question answering. Because BERT's embeddings are contextualised, they are able to capture various interpretations of a word depending on its context. This is especially helpful when trying to distinguish between words that have different meanings. Because BERT and its trained models are open-source, researchers and developers can use and expand upon its capabilities for a variety of applications.

It's important to understand how emojis and preprocessing steps impact sentiment analysis on X for a number of reasons. X is a microblogging site with a unique linguistic style. It is important to know how preprocessing affects sentiment analysis so that the analysis is customised to the characteristics of the X data. Emojis are a big part of how people express sentiment on X. They might alter or enhance the text's expressed sentiment. To fully capture the emotional context of posts, examining how emojis affect sentiment analysis would help. In order to reduce noise, preprocessing steps are crucial. Accurate sentiment analysis depends on understanding which preprocessing steps effectively clean the X data without removing important sentiment indicators. Researchers can improve the accuracy, reliability, and applicability of sentiment analysis models for social media data by looking into how preprocessing steps and emojis affect sentiment analysis in the context of X.

Understanding how different sentiment analysis approaches yield varied results based on the nature of datasets is essential for several reasons. Sentiment analysis is applied to a wide range of datasets from different domains. Investigating how various approaches perform across diverse datasets helps identify the strengths and weaknesses of each method in different contexts. Datasets may vary in terms of language nuances, colloquialisms, and linguistic styles. applying Different sentiment analysis approaches across different datasets help in understanding their performance and selecting the most suitable method for a specific language or domain. The size and composition of datasets (imbalanced classes) can impact the performance of sentiment analysis models. Analysing how different approaches handle varying data sizes and compositions ensures robustness and generalizability. Models that exhibit robustness and generalisation across diverse datasets are more likely to perform well in real-world applications where data characteristics may vary. In summary, investigating the performance of sentiment analysis approaches across diverse datasets enhances the understanding of their applicability, robustness, and limitations, enabling informed decisions in selecting or developing models for real-world scenarios.

The proposed Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) that incorporate the use of emojis can be important for sentiment analysis. Emojis can express feelings and emotions more clearly, when compared to text alone. The Emoji Dictionary (ED) can be used as a guide for matching particular emojis to specific feelings; this would help improve the model's understanding of emotional content. Emoji usage could be contextual; it may be culturally specific or related to particular demographics. This dictionary allows them to customise it to suit their requirements. By taking into account the wider context in which emojis are used, the Emoji Dictionary (ED), when combined with an advanced Sarcasm Detection Approach (SDA), can be helpful in determining the intended sentiment. Sarcasm can occasionally be expressed through emojis. A dedicated Sarcasm Detection Approach (SDA) that considers both textual

content and emojis to detect instances of sarcasm can enhance the model's ability to detect sarcasm.

In sentiment analysis, The choice between using three classes (positive, negative, and neutral) or two classes (positive and negative) depends on the specific goals and requirements of the study and the characteristics of the dataset. Three classes, however, enable a more detailed examination of sentiments. It gives neutral feelings a middle ground and acknowledges that not all expressions can be classified as only positive or negative. By incorporating a neutral class, the model can identify situations where the sentiment isn't explicitly positive or negative. Using three classes is a better reflection of reality. Having a neutral class makes it easier to recognise situations in which users are offering facts or expressing neutral opinions.

Lastly, the chapter acknowledged the importance of manual labelling, where people manually assign labels to posts, is widely seen as an efficient and reliable approach for many reasons. Human annotators have the ability to make nuanced judgements that capture the meaning of posts in context. This becomes crucial, especially when posts involve sarcasm or unclear expressions. Posts frequently include informal language, slang, abbreviations, and cultural references that can be quite diverse. Human annotators are able to overcome these challenges that automated systems may encounter. It leads to the development of a high-quality labelled dataset, which can serve as a basis for training and assessing machine learning models.

Chapter 3 Research Methodology

This chapter presents the methodology used in this study, encompassing the processes for data collection, initial insights from the pilot experiment, and the development of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA). The pilot experiment laid the foundation for the research, informing the choices made in designing and refining the data pipeline and methods.

3.1 Research Design

The overall research design is divided into several phases, as see Figure 3.1, starting with the literature review and pilot experiment, which were crucial for refining the data pipeline and methodology. This led to further data acquisition, data preprocessing, and the development of ED and SDA. The key phases of the research design are:

Phase 1: Literature Review, Pilot Experiment, and Data Acquisition

- Conduct literature review to identify gaps in sentiment analysis and sarcasm detection.
- Pilot experiment to explore preliminary data processing methods and refine the research design.
- Collect and filter datasets (COVID-19 Vaccine, Vegetarianism, Electric Cars).

Phase 2: Data Preprocessing Using VADER and BERT

- Apply preprocessing techniques (remove quotes, mentions, URLs, hashtags, emojis).
- Optimize performance for both VADER and BERT by addressing dataset splits, VADER thresholds, and handling large-scale posts.

Phase 3: Development of the Emoji Dictionary (ED) and Comparative Evaluation

- Construct the ED based on manual labelling and emoji surveys.
- Compare ED with other emoji dictionaries to evaluate its effectiveness in sentiment analysis across multiple datasets.

Phase 4: Application and Evaluation of the Sarcasm Detection Approach (SDA)

- Apply the SDA by detecting conflicts between text and emoji sentiments.
- Compare SDA with other sarcasm detection methods to evaluate its effect on sarcasm detection accuracy.

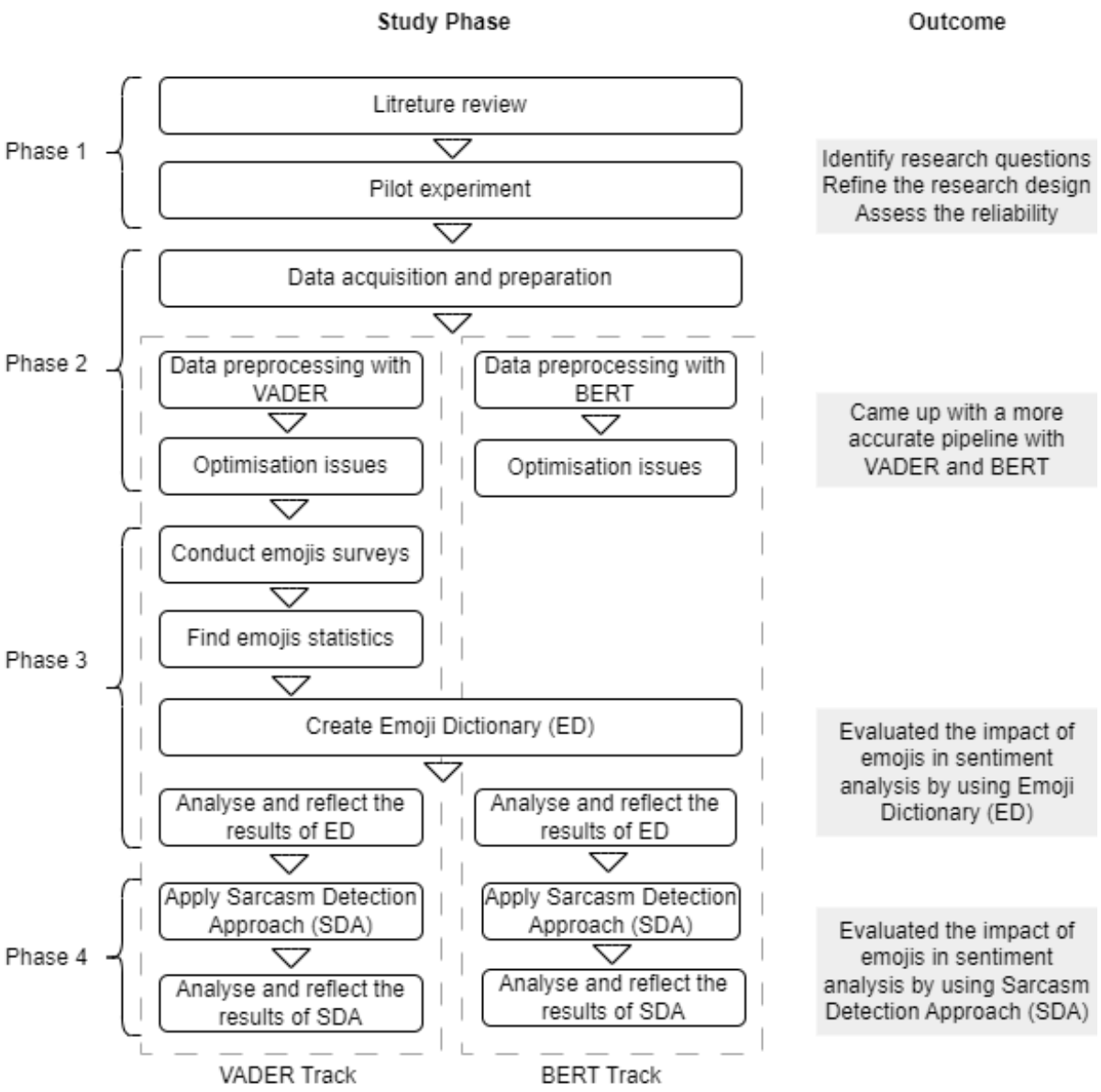

Figure 3.1 Research journey: phases and outcomes overview

3.2 Data Collection

Data collection was performed using the X API for Academic Researchers, focusing on posts related to three main topics: COVID-19 Vaccine, Vegetarianism, and Electric Cars. The posts were filtered based on the following criteria:

- Only English-language posts were retained.
- Reposts and duplicates were removed to ensure the diversity and integrity of the dataset.

Ethical approval was obtained from the Ethics Committee at the University of Southampton (research ethics number: 64049). The cleaned datasets were saved in structured CSV files for further processing.

3.3 Pilot Experiment Insights

The pilot experiment was instrumental in shaping the approach for the final research methodology. It provided initial insights into the impact of various preprocessing techniques and the role of emojis in sentiment analysis. The small dataset used in the pilot study (focused on posts about the COVID-19 vaccine) enabled a detailed investigation of how preprocessing steps influenced classification accuracy for both VADER and BERT.

Key insights from the pilot experiment included:

- Effect of Preprocessing: It was found that removing quotes, hashtags, and updating emojis resulted in an increase in overall classification accuracy for VADER, especially in detecting sarcasm through emoji usage. However, removing mentions and URLs had a negligible impact on accuracy, which suggested that these elements were not crucial for sentiment classification in the given dataset. Similarly, BERT's performance improved after applying preprocessing steps like removing quotes, hashtags, mentions, and URLs.
- Emojis in Sentiment Analysis: The handling of emojis by sentiment analysis tools like VADER revealed critical shortcomings during the pilot experiment. VADER often assigned neutral or unreliable sentiment scores to emojis, which either undermined or conflicted with the overall sentiment of the post (e.g., neutral scores for clearly positive or negative emojis). This limitation highlighted the need for a more robust approach to process emojis directly, which inspired the creation of the Emoji Dictionary (ED). During the analysis of misclassified posts, a recurring pattern emerged in posts containing emojis; in many cases, the sentiment conveyed by the emojis conflicted with the sentiment expressed in the text. This conflict often suggested sarcasm, particularly where the text might express positivity while the emojis conveyed negativity or vice versa. Recognizing this interplay, I realized that this conflict could be a strong indicator of sarcasm. This observation led to the development of the Sarcasm Detection Approach (SDA), specifically designed to detect sarcasm by analysing the sentiment conflict between text and emojis. The SDA allows for a more nuanced approach, ensuring that both text and emojis contribute meaningfully to the sentiment analysis process, particularly in detecting sarcastic tones.
- Optimization Issues: The pilot experiment also shed light on key optimization challenges, such as determining the optimal VADER threshold. Testing various

thresholds on neutral posts revealed that a threshold of 0.05 offered the best trade-off between precision and recall across all sentiment classes. For BERT, the experiment explored different training dataset sizes, showing that the model could still achieve high accuracy (84%) with only 20% of the dataset used for training, compared to 87% with 80%.

These early insights were crucial in guiding the development of the final methodology, including the design of the Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA). The experiment highlighted the importance of considering the characteristics of the dataset when choosing preprocessing steps and analysing sentiment, leading to a more refined strategy for the subsequent phases of the research.

3.4 Dataset Characteristics

Once data was collected, the characteristics of the three datasets were analysed. Each dataset was inspected for attributes such as the presence of emojis, hashtags, mentions, URLs, and the length of posts. This analysis provided a foundation for understanding how different features, especially emojis, contributed to the choosing of the preprocessing pipeline. For instance, the COVID-19 Vaccine dataset contained a high percentage of URLs (79.65%) and hashtags (29.65%), which were removed or retained based on their influence on sentiment.

3.5 Data Preprocessing

Before applying sentiment analysis techniques, several preprocessing steps were conducted:

- Removal of URLs, hashtags, mentions, stopwords, and emojis: To assess the impact of these elements on sentiment classification.
- Lemmatization: To reduce words to their base form.
- Emoji Handling: Emojis were processed by either retaining, removing, or replacing them with their sentiment scores (in ED).

The pilot experiment showed that removing URLs and mentions had a negligible effect on model performance. However, handling emojis—particularly with the introduction of ED—resulted in more accurate sentiment classification.

3.6 Sentiment Analysis

Two sentiment analysis approaches were implemented in this research:

- VADER: As a lexicon-based model, VADER was tested with the processed data to evaluate its ability to classify sentiment in social media posts. The threshold for classification was optimized to improve precision and recall.
- BERT: A machine learning model that provided context-aware sentiment classification. BERT's performance was optimized by adjusting the size of training datasets and experimenting with different dataset splits to achieve balanced and accurate classification results. Additionally, the validity of using different topics to train the classifier was explored, highlighting the flexibility of BERT in adapting to diverse content domains.

The incorporation of emojis and their dedicated processing through the Emoji Dictionary (ED) greatly enhanced both models' performance, particularly in the accurate identification of sentiment and sarcasm. The ED enabled more precise handling of emoji sentiment, while the Sarcasm Detection Approach (SDA) leveraged emoji-text sentiment conflicts to improve sarcasm detection, offering a more nuanced analysis of social media content.

3.7 Development of Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA)

3.7.1 Emoji Dictionary (ED)

The Emoji Dictionary (ED) was developed to enhance sentiment classification by addressing how emojis are handled in sentiment analysis. Instead of assigning sentiment scores to emojis, the ED assigns specific actions to each emoji: whether to retain the emoji, replace it with a synonym, or remove it entirely from the analysis. Using a combination of reviewers' opinions and manual labelling of 9,000 posts across three sentiment categories—NEG, NEU, and POS—the final action for each emoji was determined, ensuring the final decisions for each emoji were informed by both data and human judgment, resulting in a more precise approach to sentiment classification.

3.7.2 Sarcasm Detection Approach (SDA)

The Sarcasm Detection Approach (SDA) was introduced as a novel way to identify sarcasm using emojis. The approach divided each post into text and emoji components, evaluating the sentiment of both parts. If there was a conflict in sentiment (e.g., positive text but negative emojis), the SDA flagged the post as sarcastic.

3.8 Validation

3.8.1 Performance Metrics and Validation Process

To validate the effectiveness of the ED and SDA, both were compared with existing approaches. The validation was done through confusion matrices, and performance metrics such as accuracy, precision, recall, and F1-score were calculated using sklearn.metrics.

The performance of the Emoji Dictionary (ED) was compared against other prominent emoji dictionaries. Similarly, the Sarcasm Detection Approach (SDA) was evaluated against existing sarcasm detection techniques.

In another part of the validation process, we applied the classifiers (VADER and BERT) across the three datasets both with and without the integration of ED and SDA. The goal was to evaluate the impact of integrating ED and SDA on classification performance. By comparing the results before and after incorporating ED and SDA, we quantified any observed changes in accuracy, precision, and F1-scores in sentiment and sarcasm detection tasks. This highlighted the enhanced accuracy, precision, and F1-scores when these models were included in the pipeline.

3.8.2 Z-Score-Based Table Colouring

Visualizing differences in performance or values across datasets is crucial for identifying trends, outliers, and areas that require further attention. To achieve this, a Z-score-based color-coding scheme was implemented to enhance the clarity and interpretability of data tables.

By utilizing Z-scores, which indicate how far a particular value deviates from the mean of its row, this approach normalizes the data and highlights notable variations in a visually intuitive manner. A Z-score represents the number of standard deviations a data point is from the mean of its dataset. This standardization makes it possible to compare values across different rows and tables, regardless of the scale or distribution of the original data.

Z-scores naturally highlight outliers, values that are significantly higher or lower than the average, making it easier to spot exceptional cases. The Z-score-based color-coding scheme is designed with the following thresholds:

- Green for Z-Scores ≥ 1: A Z-score of 1 or higher suggests that the value is significantly higher than the average, making it a positive outlier.
- Orange for Z-Scores ≤ -1: A Z-score of -1 or lower suggests that the value is significantly lower than the average, marking it as a negative outlier.

- Lighter Green for Positive Z-Scores Between 0 and 1: Values with Z-scores between 0 and 1 are above the mean but within one standard deviation. These are moderately positive but not extreme.
- Lighter Orange for Negative Z-Scores Between 0 and -1: Values with Z-scores between 0 and -1 are below the mean but within one standard deviation. These are moderately negative but not extreme.
- No Colour for Z-Scores Close to 0 (Between -0.1 and 0.1): Values with Z-scores close to 0 are near the mean, indicating that they are typical or average within the dataset.

3.9 Summary

This chapter has provided a detailed overview of the methodology employed in this research, including key insights from the pilot experiment, data preprocessing techniques, and the development of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA). The validation process, through a comparison with existing methods, demonstrated the effectiveness of the ED and SDA in enhancing sentiment analysis and sarcasm detection. These methodologies were validated against established benchmarks, highlighting their contribution to improving the accuracy and reliability of sentiment classification.

Chapter 4 Data Pipeline

4.1 Introduction

The process of sentiment analysis is heavily dependent on the underlying data pipeline, which ensures that the data is prepared in a way that enables reliable and effective classification. This chapter outlines the steps involved in the data pipeline used for this study. Specifically, it focuses on optimizing preprocessing for sarcasm detection across different classifiers and datasets, utilizing VADER and BERT for sentiment analysis. These classifiers will be applied to three distinct datasets, assessing how emojis are integrated into the classification process and evaluating the impact of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) on sentiment analysis performance.

4.2 Data Acquisition

Posts from the three datasets—the COVID-19 Vaccine, Electric Cars, and Vegetarianism —were collected by using a Python query to access the X for Academic Researchers API in the period from January 1, 2022, to January 31, 2022. Table 4.1 shows the search words that were used to collect each dataset and the number of posts that are collected for each dataset. CSV files are used to store all of the findings. The number of posts is not determined in the query.

Table 4.1 Search words and post collection details for the COVID-19 Vaccine, Electric Cars, and Vegetarianism datasets, including the number of posts collected

Dataset	**Search words**	**Number of posts collected**
COVID-19 Vaccine	Covid19 vaccine, COVID-19 Vaccine, coronavirus vaccine, corona vaccine, Koronavirus vaccine, Corona virus vaccine, Sars-cov-2 vaccine, chinese virus vaccine, Covid vaccine.	440,276 posts
Electric Cars	Electric Cars, electric car, electric vehicle, electric vehicles, electric automobile, electric automobiles, electronic vehicles, robotic vehicle, automated vehicle, automated vehicles, battery vehicle, battery vehicles, ecars, ecar, e-car, electric motor, electric motors.	104,449 posts

Vegetarianism	Vegetarianism, vegan, vegetarian	248,782 posts

4.3 Data Preparation

This subsection is concerned with how data was prepared at a high level. To enhance the research's robustness and reliability, this study was conducted on three different datasets. The dataset's characteristics will be identified to understand how contextual variations can influence research outcomes. All the experiments, including the following process of data acquisition and preparation, are conducted three times, once for each dataset. Figure 4.1 shows the pipeline of these stages that are taken in this chapter, and each section presents the results of each process.

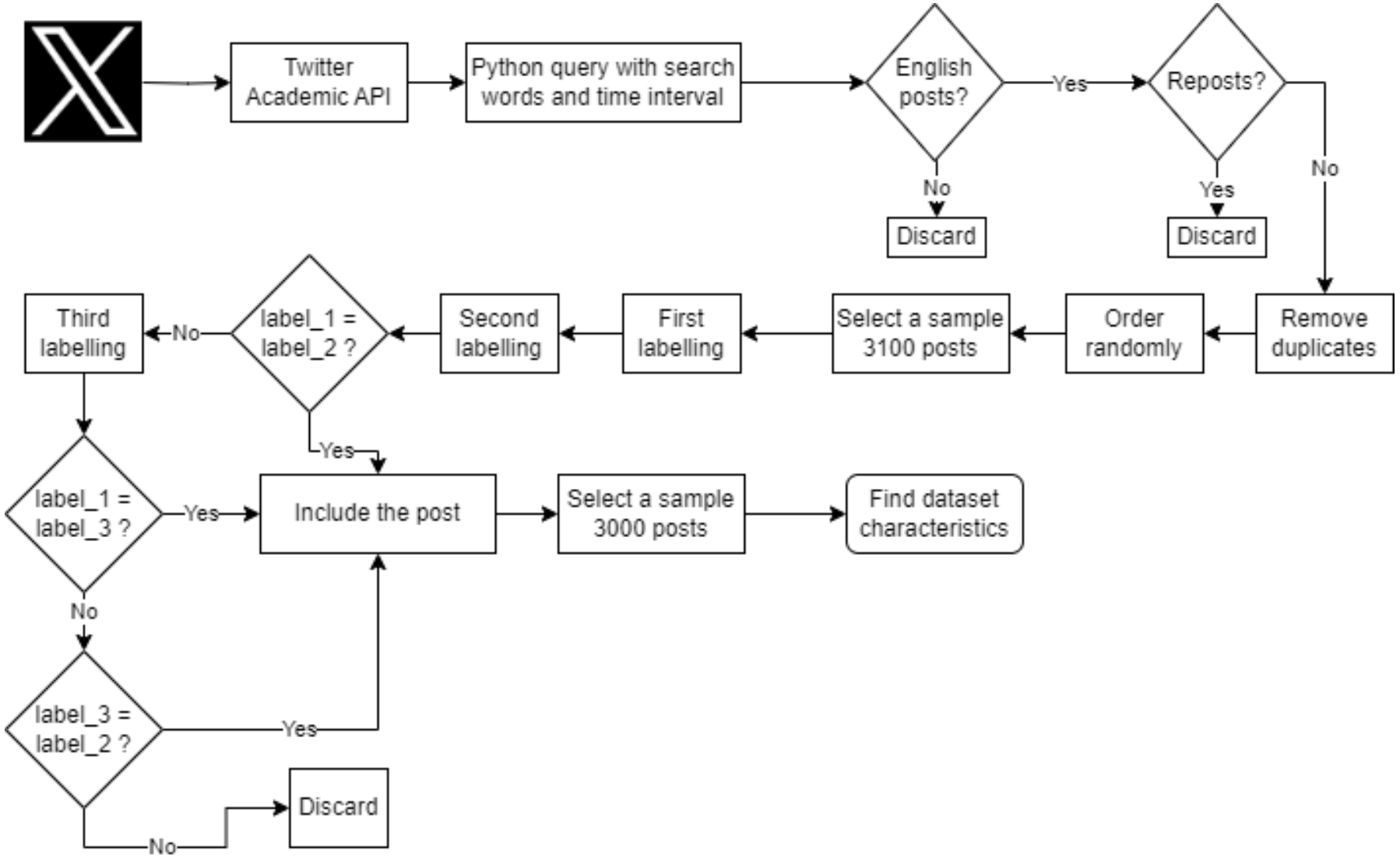

Figure 4.1 Data acquisition and preparation workflow chart

4.3.1 Filtering the Data

To refine the dataset, all collected posts are filtered through a multi-step process. Non-English posts are identified and excluded based on language metadata analysis. Reposts are determined by employing the Regular Expression module in Python, which allows for precise pattern matching to identify repeated content. Duplicate posts are detected and efficiently removed using the Pandas library's drop_duplicates() function, ensuring the dataset is comprised only of unique, original English-language posts for subsequent analysis. Number of the remaining posts is presented in Table 4.2. Then, the posts ordered randomly.

Table 4.2 Summary of dataset composition before and after filtering - detailing initial collection and final post-filtering counts

Dataset	**Number of Posts Collected**	**Number of Posts After Filtering**
COVID-19 Vaccine	440,276 posts	253,337 posts
Electric Cars	104,449 posts	64,638 posts
Vegetarianism	248,782 posts	173,128 posts

4.3.2 Manual Labelling

To establish a reliable foundation for analysis, each dataset was designed to include 3,000 posts. An additional 100 posts per dataset were gathered to account for potential inconsistencies in human interpretation during the labelling process. The posts were reviewed by three annotators, each tasked with assigning one of three sentiment labels: positive (+), negative (-), or neutral (/). Posts where all three annotators disagreed on the label were removed from the dataset to reduce ambiguity and potential bias. This step was crucial to ensure that only posts with a clear majority agreement on sentiment classification remained.

The goal of this multi-annotator approach was to enhance the reliability of the dataset by mitigating personal biases and allowing diverse perspectives to converge on a common label. This strategy proved particularly useful for addressing nuanced or ambiguous sentiments that could be interpreted differently depending on individual perceptions. Posts were included if at least two annotators agreed on the sentiment label. After this refinement process, the final post

counts stood at 3,019, 3,025, and 3,002 for the COVID-19 Vaccine, Electric Cars, and Vegetarianism datasets, respectively.

The slight differences in the number of posts across datasets highlight the variability in human interpretation, underscoring the need for a consensus-based approach to labelling. By incorporating multiple viewpoints, the dataset benefits from a higher level of confidence in the assigned labels, ensuring that the analysis is built on a solid foundation of accurate sentiment classification.

Figure 4.2 illustrates the sentiment distribution across the three datasets, categorized into positive, negative, and neutral classes. The distribution percentages shown in the figure provide insight into the sentiment composition after labelling. Across all datasets, 57.99% of posts were classified as neutral (NEU), 27.02% as negative (NEG), and 14.99% as positive (POS). This distribution reflects the natural variability in sentiment expression across different topics.

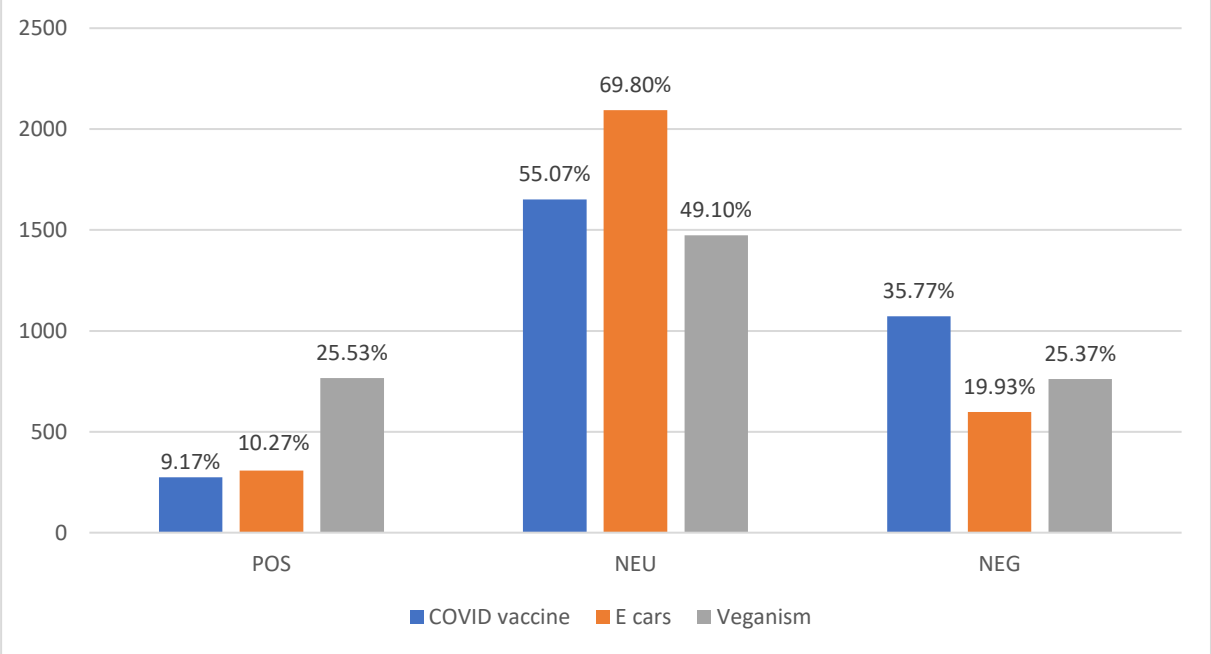

Figure 4.2 Sentiment distribution across datasets

4.3.3 Dataset Characteristics

Following the labelling process, an in-depth analysis of the characteristics of each dataset was conducted. Datasets vary in several ways, including size, temporal span, language, text length, and content domain. These factors considerably influence the performance of sentiment analysis methods, making it essential to thoroughly examine the datasets to understand the potential affordances, limitations, and areas requiring further refinement.

The three chosen datasets—COVID-19 Vaccine, Electric Cars, and Vegetarianism—represent distinct topics, each with unique characteristics. These differences include not only the subject matter but also ethical considerations, text structure, emoji usage, and imbalances in sentiment distribution. Such variability allows for a comprehensive evaluation of sentiment analysis techniques across diverse contexts.

The COVID-19 Vaccine topic belongs to the fields of public health, epidemiology, and healthcare. It relates to public health and how people react to COVID-19 vaccinations. People's sentiments towards vaccines can vary widely and include concerns about the safety, effectiveness, and accessibility of vaccines, as well as public health actions associated with the epidemic. This topic is widely important because of the pandemic's extensive effects and the necessity of vaccination campaigns everywhere, which include talks about the development of vaccines, the use of technology, and vaccine distribution. It has some ethical concerns, like vaccine equity and using poor and vulnerable people for vaccine trials.

The discussion surrounding Electric Cars topic includes aspects such as eco-friendliness, energy sources, and the automotive industry. It involves vehicles that are powered by electricity. This subject delves into the influence of Electric Cars on the environment, technical developments, and consumer acceptance. Viewpoints on this topic may vary, including the advantages and disadvantages of this technology, the environmental effects, the viability of Electric Cars, and charging infrastructure. This topic is important globally as many countries focus on reducing carbon emissions and switching to environmentally friendly transportation.

The topic of Vegetarianism falls under the categories of animal rights, ethical decisions, and nutrition. This topic is related to diet and lifestyle, which exclude animal-derived products. This involves conversations about ethical issues, animal welfare and rights, plant-based diets, environmental sustainability, and other ethical issues. Sentiments on this topic may include how Vegetarianism affects both the environment and personal health, as well as individual beliefs and moral dilemmas. This subject influences aspects such as the food industry, agriculture, and consumer decisions. Conversations may vary depending on the culture, food, and lifestyle among regions.

Beside these general characteristics, further characteristics of the collected datasets are detailed in Table 4.3. These characteristics were identified through an analysis of 3,000 posts from each dataset.

Table 4.3 Datasets' characteristics

Characteristic	**COVID-19 Vaccine**	**Electric Cars**	**Vegetarianism**
Percentage of NEG posts	35.77%	19.93%	25.37%
Percentage of NEU posts	55.07%	69.8%	49.1%
Percentage of POS posts	9.17%	10.27%	25.53%
Percentage of posts with quotes	7%	5.13%	3.13%
Percentage of posts with hashtags	29.53%	36.07%	34.77%
Percentage of posts with mentions	15.87%	20.63%	14.2%
Percentage of posts with URLs	78.23%	84.1%	64.63%
Percentage of posts with Emojis	11.47%	11.47%	25.13%
Percentage of posts with has intense words	0.87%	0.67%	1.9%
max number of words in a post	60	57	59
min number of words in a post	1	1	2
mean number of words in a post	28.4	25.71	22.52
max number of hashtags in a post	23	26	26
min number of hashtags in a post	0	0	0
mean number of hashtags in a post	0.88	1.13	1.58
Total number of words in the dataset	53,126	50,489	43,276
Number of unique words in the dataset	11,245	11,513	11,715
Percentage of posts has slangs	94.5%	93.13%	89.7%

4.4 Data Preprocessing and Evaluation

This section explores the technical and analytical methods used to refine, clean, and transform the raw data into a format suitable for further analysis. Each of the preprocessing steps is examined by itself with the two classifiers VADER and BERT on the three datasets to determine if they are worthy of keeping or removing. At the end, a comparative analysis will be performed by comparing the results obtained with and without preprocessing after combining all the

preprocessing steps that enhance the performance. This will show the advantages that our preprocessing approach provides for the influence of sentiment analysis.

Each of the following paragraphs explains one of the preprocessing steps.

Quotations, or text enclosed in quote marks, frequently express the feelings of others rather than the author themselves. Consequently, their presence could affect the evaluation of sentiment in the end. Therefore, removing quotes from the analysis is expected to improve sentiment classification performance by guaranteeing that the evaluation more accurately reflects the author's personal sentiment.

Since mentions and URLs do not usually have sentiment values, removing them from the text won't likely have a direct effect on the results of sentiment classification. Nevertheless, removing these components can simplify the analysis procedure, possibly improving overall classification performance by lowering noise and focusing the assessment on text that is more likely to express strong sentiment.

Users often use hashtags to increase the reach of their posts and make it easier for people to find the posts that are related to a specific topic. While hashtags are generally useful, they can also convey emotions independently. Some users employ hashtags on social media as a medium to express their emotions. In order to better understand how hashtags contribute to sentiment within posts, this experiment will examine and evaluate their influence on sentiment analysis.

Emojis are powerful sentiment conveyers that are frequently eliminated during preprocessing in different studies. The purpose of this experiment is to specifically examine how emojis affect sentiment analysis and determine whether keeping them in the dataset or removing them will enhance classification performance. The analysis seeks to uncover how useful emojis are for precisely identifying and categorising expressed sentiments.

Stopwords, which include common words like "a," "the," and "in," usually do not change the main idea of a sentence. Three datasets will be used in this experiment to test the removal of stopwords in order to assess whether stopwords are necessary when using sentiment analysis. Finding out if the removal of these common and apparently unimportant words improves or decreases classification accuracy.

Some analyses use lemmatization, a preprocessing technique that reduces words to their base or root form. In order to determine whether lemmatization improves sentiment analysis in general, this section examines how effective it is to apply it to posts. The goal is to understand the influence of lemmatization, on the accuracy of the classification approach.

This study provides an overview of preprocessing techniques for sentiment analysis by analysing the effects of different preprocessing steps. Each step is assessed based on how well it can improve sentiment classification performance. This study attempts to determine the best preprocessing techniques by carefully evaluating the elimination of non-sentiment elements, incorporating sentiment-conveying emojis, and the simplicity offered by removing stopwords and lemmatizing words. The main objective is to identify a set of preprocessing steps that enhance the efficiency and precision of sentiment analysis.

4.4.1 Preprocessing with VADER

During the phase of processing when VADER is used for sentiment analysis, each dataset is assessed to determine the effective preprocessing steps. Initially, the original VADER is used on unmodified posts to establish a baseline for sentiment classification accuracy. Each preprocessing method is evaluated individually. Decisions are made accordingly. Sklearns metric reports are utilised to compare the outcomes of these evaluations.

The purpose of this comparison is to find out whether there are any noticeable advantages to each preprocessing step in terms of improving the performance of sentiment classification using VADER. Subsequently, all successful preprocessing steps are combined and evaluated to identify the impactful changes that could potentially enhance sentiment analysis performance, as depicted in Figure 4.3.

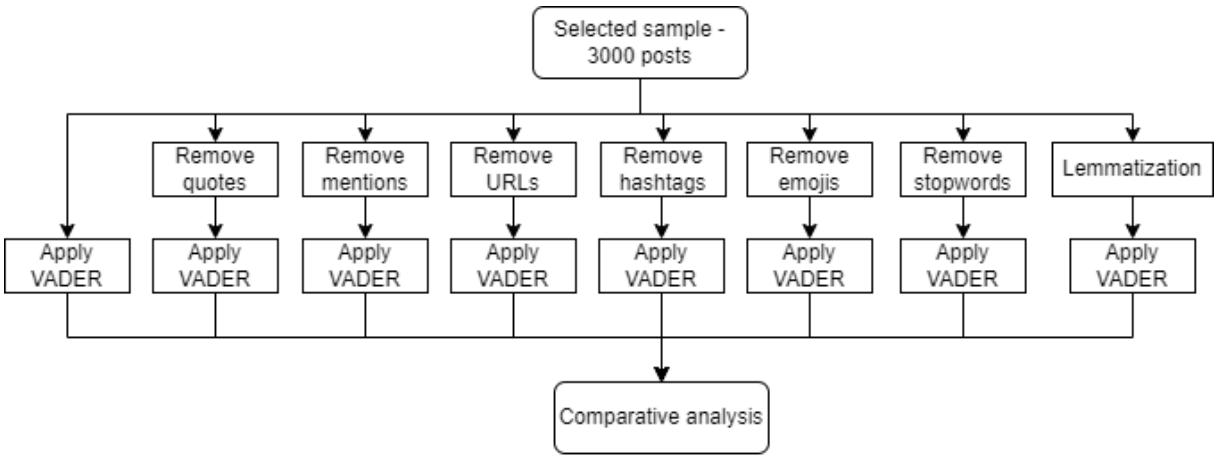

Figure 4.3 VADER preprocessing workflow chart

Table 4.4, Table 4.5, and Table 4.6 offer a comparison across three datasets, showcasing the impact of various preprocessing techniques on the performance metrics when applying VADER for sentiment analysis. These techniques include the removal of quotes, mentions, hashtags, emojis, stopwords, and the application of lemmatization to the posts. To facilitate a clear comparison, each preprocessing outcome is benchmarked against the results obtained from analysing the original posts with VADER, utilising a Green-White-Orange colour scale to visually

represent improvements, neutrality, or declines in performance metrics within the same category.

Notably, across all datasets, the removal of quotes (No_Quotes column) consistently leads to an enhancement in performance metrics, highlighting the distracting nature of quotes in sentiment analysis.

The exclusion of emojis (referenced as the No_Emojis column) clarifies there is a decrease in performance, especially evident in Dataset 3, where the number of posts containing emojis is more than double the proportion of posts with emojis compared to Datasets 1 and 2. This highlights the important function emojis serve in expressing emotions.

The removal of hashtags has a marginal impact, with slight variations in performance observed in Datasets 2 and 3, which contain a higher frequency of hashtags.

Interestingly, both lemmatization and the removal of stopwords generally result in lowered classification performance across the datasets.

Further analysis reveals a direct correlation between the dataset characteristics discussed in Section 6.3.3 and the observed preprocessing effects. Specifically, removing quotes shows the positive impact on Dataset 1, where 7% of posts contain quotes. Dataset 2 (5.13%) and Dataset 3 (3.13%) are the datasets that come after this. These findings indicate that the presence of quotes directly influences how effective this preprocessing step is. Similarly, the minor adjustments noticed when removing hashtags are more impactful in Datasets 2 and 3, which exhibit higher levels of hashtag usage.

The negative effects of emoji removal on performance are especially marked in Dataset 3, where emojis are more prevalent, highlighting their importance in sentiment expression. These results show how different preprocessing steps interact with the unique features of each dataset. This highlights the need for customised preprocessing methods in sentiment analysis to improve model performance.

Table 4.4 VADER performance on original posts and after applying each of the preprocessing steps in the COVID-19 Vaccine dataset

Dataset 1: COVID-19 Vaccine		Orig.	No_Quotes	No_Mentions	No_URLs	No_Hashtags	No_Emojis	No_Stopwords	Lemmatization
NEG: 1073 posts	Prec.	0.54	0.55	0.55	0.54	0.55	0.55	0.55	0.53
	Rec.	0.56	0.55	0.56	0.56	0.55	0.56	0.54	0.56
	F1	0.55	0.55	0.55	0.55	0.55	0.55	0.54	0.55
NEU: 1652 posts	Prec.	0.79	0.79	0.79	0.80	0.79	0.78	0.79	0.81
	Rec.	0.33	0.35	0.33	0.33	0.34	0.33	0.34	0.31
	F1	0.47	0.49	0.47	0.47	0.47	0.47	0.47	0.45
POS: 275 posts	Prec.	0.18	0.19	0.18	0.18	0.18	0.18	0.18	0.18
	Rec.	0.80	0.81	0.80	0.80	0.79	0.79	0.80	0.81
	F1	0.29	0.30	0.29	0.29	0.29	0.29	0.29	0.30
Accuracy		0.46	0.47	0.46	0.46	0.46	0.46	0.45	0.45

Table 4.5 VADER performance on original posts and after applying each of the preprocessing steps in the Electric Cars dataset

Dataset 2: Electric Cars		Orig.	No_Quotes	No_Mentions	No_URLs	No_Hashtags	No_Emojis	No_Stopwords	Lemmatization
NEG: 598 posts	Prec.	0.51	0.52	0.52	0.51	0.51	0.51	0.52	0.51
	Rec.	0.47	0.46	0.47	0.47	0.47	0.46	0.45	0.47
	F1	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.49
NEU: 2094 posts	Prec.	0.85	0.86	0.85	0.85	0.86	0.84	0.86	0.85
	Rec.	0.43	0.46	0.44	0.43	0.44	0.44	0.44	0.42
	F1	0.58	0.60	0.58	0.58	0.58	0.58	0.58	0.56
POS: 308 posts	Prec.	0.19	0.19	0.19	0.19	0.19	0.19	0.18	0.18
	Rec.	0.84	0.84	0.84	0.84	0.84	0.83	0.85	0.83
	F1	0.31	0.31	0.31	0.31	0.31	0.30	0.30	0.29
Accuracy		0.48	0.50	0.48	0.48	0.49	0.48	0.48	0.47

Table 4.6 VADER performance on original posts and after applying each of the preprocessing steps in the Vegetarianism dataset

Dataset 3: Vegetarianism		Orig.	No_Quotes	No_Mentions	No_URLs	No_Hashtags	No_Emojis	No_Stopwords	Lemmatization
NEG: 761 posts	Prec.	0.66	0.66	0.66	0.66	0.66	0.64	0.65	0.65
	Rec.	0.47	0.47	0.47	0.47	0.47	0.44	0.43	0.47
	F1	0.55	0.55	0.55	0.55	0.55	0.52	0.52	0.55
NEU: 1473 posts	Prec.	0.76	0.76	0.76	0.76	0.76	0.71	0.76	0.77
	Rec.	0.41	0.42	0.41	0.41	0.43	0.42	0.41	0.40
	F1	0.53	0.54	0.53	0.53	0.55	0.53	0.54	0.52
POS: 766 posts	Prec.	0.40	0.40	0.40	0.40	0.41	0.39	0.39	0.40
	Rec.	0.87	0.87	0.87	0.87	0.87	0.81	0.87	0.87
	F1	0.55	0.55	0.55	0.55	0.55	0.53	0.54	0.55
Accuracy		0.54	0.55	0.54	0.54	0.55	0.53	0.54	0.54

To enhance the accuracy of the VADERs analysis, steps were taken like removing quotes, mentions, URLs, and hashtags while keeping emojis to capture the emotional context of posts. This approach is quantitatively assessed by combining all the preprocessing steps in Table 4.7. They are visually represented in Figure 4.4, showing how they impact performance metrics across datasets. Although the overall change in performance appears marginal at first glance, a deeper analysis reveals notable enhancement in the recall and F1-score metrics for the neutral (NEU) class across all datasets. This suggests that the preprocessing steps improve the model's ability to correctly identify and classify neutral sentiments, reducing misclassifications. The improvement might be caused by the removal of distracting information (like URLs and mentions) that would have confused the sentiment analysis process and made it a bit harder for VADER to identify the sentiment.

This result highlights the importance of customised preprocessing in sentiment analysis by showing how adjustments to the preprocessing of data can enhance model performance in detecting sentiment classes. The impacts of these modifications are considerable for researchers who seek to improve sentiment analysis methods, emphasising the fine line that must be drawn between data preprocessing and the keeping of emotional information in order to achieve accurate sentiment classification.

Table 4.7 VADER performance before and after suggested preprocessing steps in all datasets

		Dataset 1		**Dataset 2**		**Dataset 3**	
		Before	**After**	**Before**	**After**	**Before**	**After**
NEG	**Prec.**	0.54	0.55	0.51	0.52	0.66	0.66
	Rec.	0.56	0.55	0.47	0.46	0.47	0.47
	F1	0.55	0.55	0.49	0.49	0.55	0.55
NEU	**Prec.**	0.79	0.79	0.85	0.86	0.76	0.76
	Rec.	0.33	0.36	0.43	0.47	0.41	0.44
	F1	0.47	0.49	0.58	0.60	0.53	0.55
POS	**Prec.**	0.18	0.19	0.19	0.19	0.40	0.41
	Rec.	0.80	0.80	0.84	0.83	0.87	0.86
	F1	0.29	0.30	0.31	0.31	0.55	0.55
Accuracy		0.46	0.47	0.48	0.50	0.54	0.55

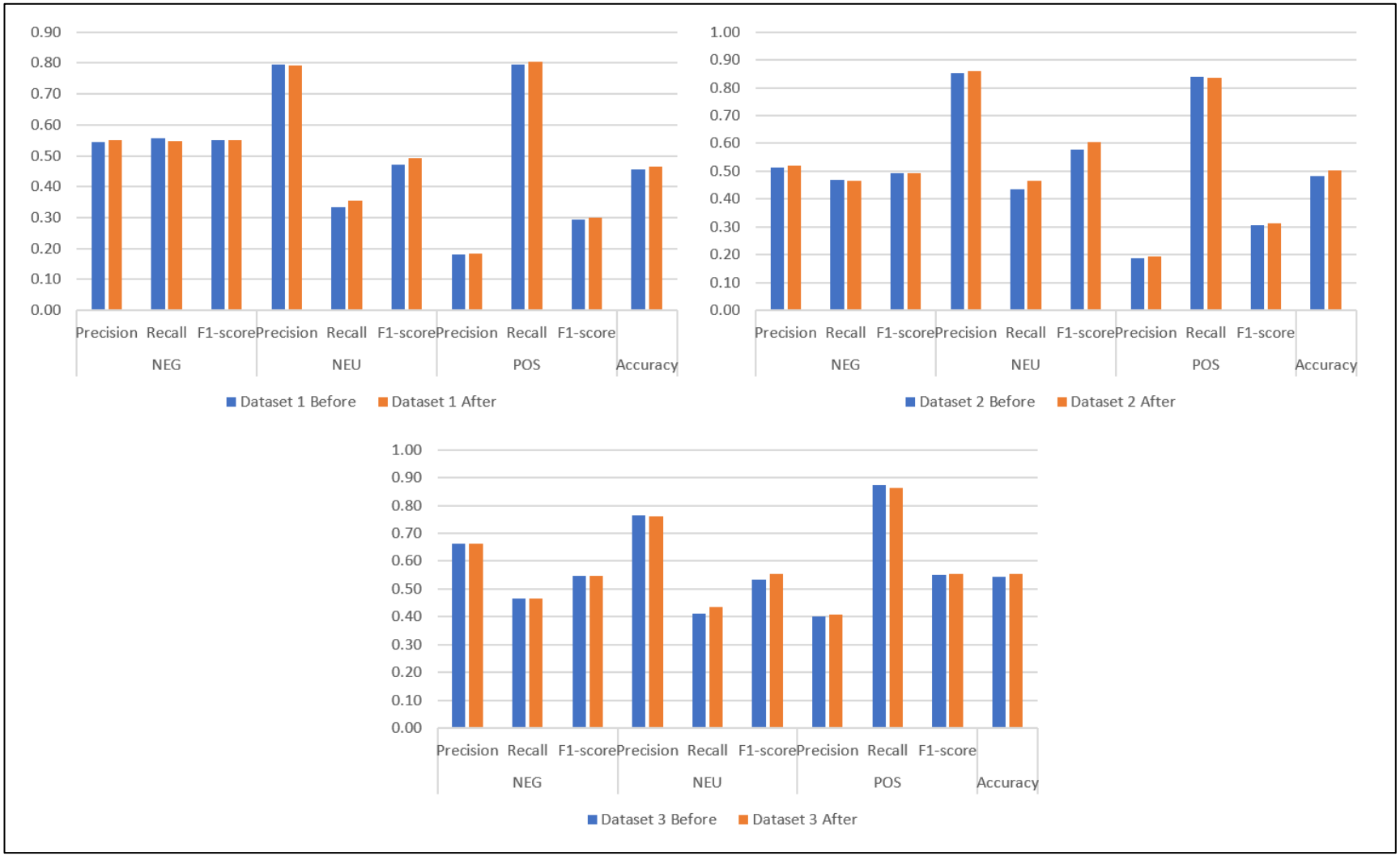

Figure 4.4 Three bar charts represent the changes in VADER evaluation metrics of the three datasets before and after the suggested preprocessing steps

4.4.2 Preprocessing with BERT

In the preprocessing stage for sentiment analysis using BERT, every dataset is carefully evaluated to determine the best preprocessing techniques. BERT is first applied to the unprocessed, raw posts in order to create a baseline for sentiment classification precision. The effects of all preprocessing steps are then assessed separately. The effectiveness of these preprocessing steps is then evaluated using Sklearn's metric reports to identify any impacts these actions provide in enhancing BERT's sentiment classification. Following this, the preprocessing methods that contribute to improved classification accuracy are combined and re-evaluated. This iterative approach ensures that preprocessing is optimized for BERT, enhancing its ability to accurately classify sentiment. Figure 4.5 provides an overview of the preprocessing workflow chart.

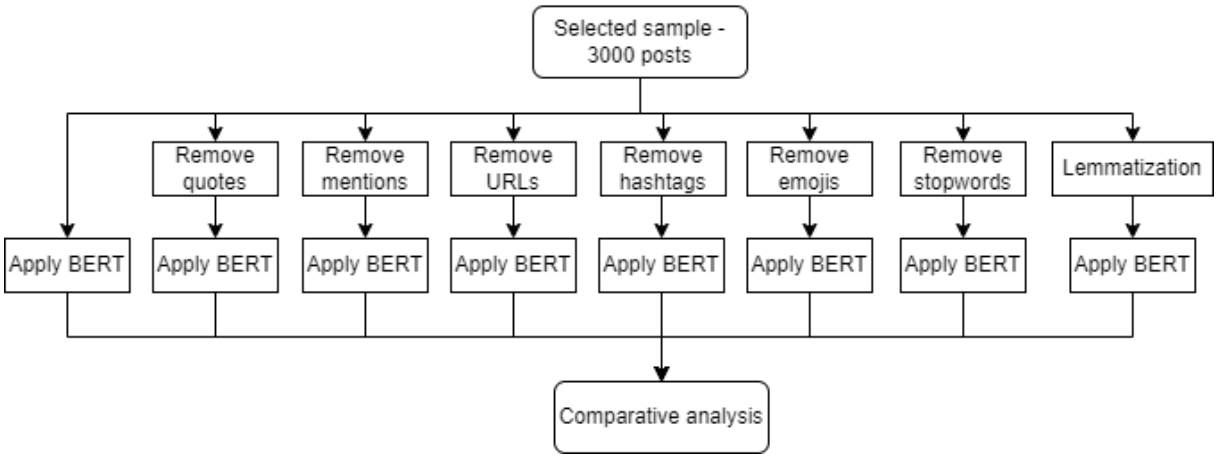

Figure 4.5 BERT preprocessing workflow chart

Table 4.8, Table 4.9, and Table 4.10 illustrate the preprocessing analysis and show how it affects the classification results for all datasets. This means that text preparation needs to be done effectively before using BERT for classification.

Notably, removing quotes generally results in a small but consistent increase in classification accuracy of between 1% and 2% for all datasets. This suggests that quotes could be eliminated in the preprocessing steps because they might reduce the classifier's focus by introducing unrelated text.

Similarly, performance is slightly enhanced by removing mentions and URLs. Their presence confuses the classifier through integrating elements into the text that may not have sentiment value, in addition to increasing computational load and running times. Therefore, in order to simplify the classifier's focus on sentiment-relevant content, mentions and URLs will be removed.

The impact of hashtags gives a more complicated picture. In the first two datasets, their removal improves the classification performance; however, in the third dataset, the opposite effect is seen. The percentages of posts containing hashtags in all datasets are 29.53%, 36.07%, and 34.77%, respectively. The variation in performance change could be explained by the different functions hashtags perform based on context; they might offer sentiment indicators in certain scenarios but introduce noise in others. The choice to exclude hashtags aims to reduce confusion in classification results due to their presence across datasets and the overall positive impact of their removal on two out of three datasets.

Emojis are particularly important for expressing emotion; their absence markedly reduces performance, especially in the negative (NEG) class. This demonstrates the importance of emojis in enhancing textual sentiment expression and justifies their use in all posts in order to maintain these important sentiment indicators.

The analysis further explores the effects of removing stopwords and applying lemmatization. Our results show a decline in classification performance for the first and last datasets, with only the second dataset demonstrating improvement. This variation highlights the contextual sensitivity of sentiment analysis, implying that these kinds of preprocessing steps could potentially oversimplify the textual data, making it more difficult for BERT to detect sentiment accurately.

Overall, this thorough analysis of preprocessing effects informs our strategy for refining the textual data for BERT classification by finding a balance between the need to reduce noise and computational complexity and the preservation of important sentiment indicators. These observations highlight the importance of a personalised preprocessing approach that is sensitive to the different characteristics and challenges that every dataset presents.

Table 4.8 BERT performance on original posts and after applying each of the preprocessing steps in the COVID-19 Vaccine dataset

Dataset 1: COVID-19 Vaccine		Orig.	No_Quotes	No_Mentions	No_URLs	No_Hashtags	No_Emojis	No_Stopwords	Lemmatization
NEG: 108 posts	Prec.	0.85	0.85	0.84	0.83	0.83	0.81	0.84	0.8
	Rec.	0.81	0.8	0.81	0.83	0.84	0.81	0.8	0.85
	F1	0.83	0.82	0.83	0.83	0.84	0.81	0.82	0.83
NEU: 165 posts	Prec.	0.87	0.88	0.89	0.89	0.89	0.89	0.86	0.9
	Rec.	0.91	0.93	0.9	0.9	0.91	0.89	0.92	0.86
	F1	0.89	0.91	0.89	0.89	0.9	0.89	0.89	0.88
POS: 27 posts	Prec.	0.68	0.71	0.64	0.73	0.83	0.65	0.68	0.7
	Rec.	0.63	0.63	0.67	0.7	0.7	0.63	0.56	0.7
	F1	0.65	0.67	0.65	0.72	0.76	0.64	0.61	0.7
Accuracy		0.85	0.86	0.85	0.86	0.87	0.84	0.84	0.84
Time		43min 43s	42min 3s	41min 24s	37min 27s	41min 49s	41min 48s	44min 50s	42min 57s

Table 4.9 BERT performance on original posts and after applying each of the preprocessing steps in the Electric Cars dataset

Dataset 2: Electric Cars		Orig.	No_Quotes	No_Mentions	No_URLs	No_Hashtags	No_Emojis	No_Stopwords	Lemmatization
NEG: 60 posts	Prec.	0.79	0.79	0.79	0.81	0.79	0.76	0.83	0.8
	Rec.	0.77	0.73	0.77	0.78	0.82	0.68	0.73	0.8
	F1	0.78	0.76	0.78	0.8	0.8	0.72	0.78	0.8
NEU: 210 posts	Prec.	0.92	0.91	0.91	0.91	0.93	0.89	0.91	0.91
	Rec.	0.9	0.91	0.92	0.91	0.89	0.92	0.94	0.92
	F1	0.91	0.91	0.92	0.91	0.91	0.9	0.93	0.92
POS: 30 posts	Prec.	0.58	0.62	0.65	0.56	0.61	0.61	0.76	0.69
	Rec.	0.7	0.7	0.67	0.6	0.77	0.57	0.73	0.6
	F1	0.64	0.66	0.66	0.58	0.68	0.59	0.75	0.64
Accuracy		0.85	0.86	0.86	0.86	0.86	0.84	0.88	0.87
Time		40min 48s	40min 47s	40min 42s	38min 2s	39min 26s	40min 48s	40min 51s	34min 40s

Table 4.10 BERT performance on original posts and after applying each of the preprocessing steps in the Vegetarianism dataset

Dataset 3: Vegetarianism		**Orig.**	**No_Quotes**	**No_Mentions**	**No_URLs**	**No_Hashtags**	**No_Emojis**	**No_Stopwords**	**Lemmatization**
NEG: 76 posts	**Prec.**	0.75	0.73	0.77	0.75	0.74	0.74	0.7	0.67
	Rec.	0.64	0.78	0.63	0.74	0.59	0.64	0.67	0.62
	F1	0.7	0.75	0.7	0.74	0.66	0.69	0.68	0.64
NEU: 148 posts	**Prec.**	0.79	0.83	0.8	0.81	0.75	0.79	0.82	0.77
	Rec.	0.79	0.77	0.82	0.78	0.79	0.84	0.81	0.76
	F1	0.79	0.8	0.81	0.8	0.77	0.82	0.82	0.77
POS: 76 posts	**Prec.**	0.67	0.71	0.66	0.68	0.64	0.7	0.6	0.63
	Rec.	0.76	0.76	0.74	0.74	0.71	0.7	0.64	0.7
	F1	0.71	0.73	0.7	0.71	0.68	0.7	0.62	0.66
Accuracy		0.75	0.77	0.75	0.76	0.72	0.76	0.73	0.71
Time		41min 10s	36min 33s	41min 22s	36min 38s	41min 18s	40min 38s	41min 24s	34min 33s

As part of optimising text for sentiment analysis using BERT, key preprocessing steps were undertaken, including the removal of quotes, mentions, URLs, and hashtags, while keeping emojis to maintain emotional context. This approach is quantitatively evaluated by combining all the preprocessing steps. Table 4.11 and visually in Figure 4.6, which show the performance changes across each dataset with and without these preprocessing steps. Notably, the results showed an increase in precision, recall, and overall accuracy for all classes in all datasets, as well as reduced model running times. Specifically, the precision of the positive (POS) class showed a marked increase. Moreover, there was an increase in the recall of the negative (NEG) class, which was particularly evident in Dataset 1, due to the substantial proportion of the NEG class (35.8%) within this dataset.

Table 4.11 BERT performance before and after suggested preprocessing steps in all datasets

		Dataset 1		**Dataset 2**		**Dataset 3**	
		Before	**After**	**Before**	**After**	**Before**	**After**
NEG	**Prec.**	0.85	0.86	0.79	0.81	0.75	0.74
	Rec.	0.81	0.86	0.77	0.77	0.64	0.71
	F1	0.83	0.86	0.78	0.79	0.70	0.72
NEU	**Prec.**	0.87	0.90	0.92	0.91	0.79	0.80
	Rec.	0.91	0.89	0.90	0.92	0.79	0.79
	F1	0.89	0.90	0.91	0.92	0.79	0.80
POS	**Prec.**	0.68	0.76	0.58	0.70	0.67	0.70
	Rec.	0.63	0.81	0.70	0.70	0.76	0.75
	F1	0.65	0.79	0.64	0.70	0.71	0.73
Accuracy		0.85	0.87	0.85	0.87	0.75	0.76
Time		43min 43s	36min 20s	40min 48s	38min 4s	41min 10s	32min 25s

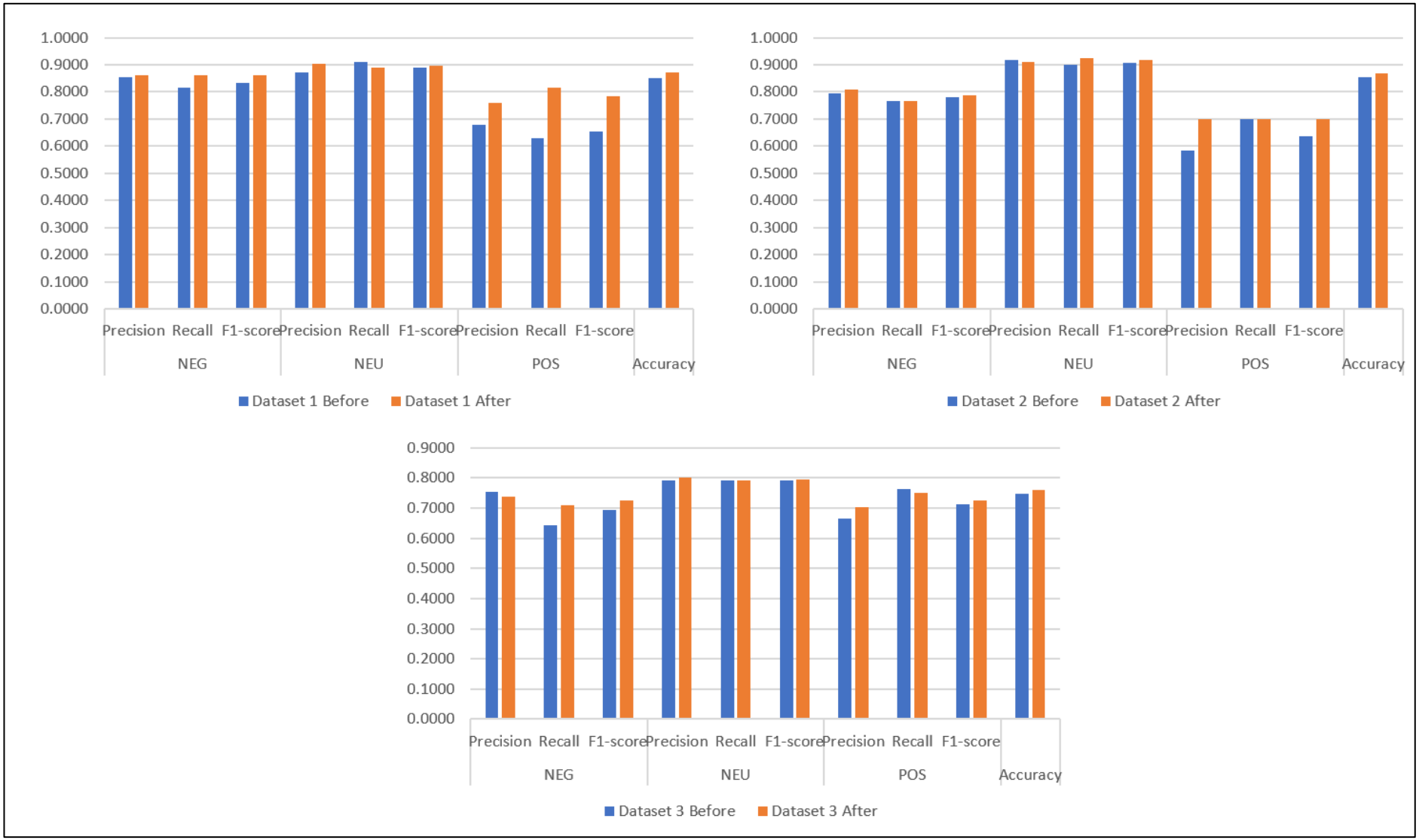

Figure 4.6 Three bar charts represent the changes in BERT evaluation metrics of the three datasets before and after the suggested preprocessing steps

4.4.3 Summary

There are similarities and variances in how VADER and BERT perform before and after the recommended preprocessing pipeline. Analysing the performance metrics across classes in the datasets reveals a nuanced impact of preprocessing on each sentiment analysis tool.

Post-preprocessing, both VADER and BERT exhibit an increase in accuracy across all datasets. This underscores the significance of preprocessing in refining input data for sentiment classification.

After preprocessing, both tools display an increase in the precision of the POS class. This implies that removing noise like hashtags, quotes, mentions, and URLs from posts can help in identifying sentiments.

The impact of preprocessing on the NEG and NEU classes shows a distinction between VADER and BERT. While VADER's performance remains relatively stable or shows minimal change, BERT experiences benefits, particularly in precision and recall for the NEG class as well as precision for the NEU class. This suggests that BERT performs better when working with cleaned data.

While both tools exhibit increased precision in the POS class, the recall metric for VADER remains unchanged after preprocessing, whereas BERT shows a notable increase. This difference demonstrates the BERT's ability to understand context better after data cleaning, making it more sensitive to nuanced expressions of positivity.

The post-preprocessing running times for BERT are reduced, indicating increased efficiency in both performance and processing time. This change is not observed with VADER since it is a rule-based model, and the complexity of the dataset usually has less of an impact on its processing time.

It is important to note that preprocessing consistently brings advantages when applied to different sentiment analysis tools, but to varying levels and in different ways. Although VADER exhibits moderate increases, mainly in accuracy and small precision adjustments in the POS class, BERT shows more substantial increases in precision, recall, and F1 scores in all classes, along with greater efficiency. These differences highlight the significance of selecting appropriate sentiment analysis tools and preprocessing techniques based on the objectives and characteristics of the dataset.

The different results between preprocessing with VADER and BERT illustrate the main distinctions between rule-based and machine-learning-based techniques. These insights

illustrate how important preprocessing is to improving sentiment analysis and help researchers modify their approaches for maximum efficacy and precision.

To sum up, this chapter has outlined the methods and approaches used to clean up and prepare the data for the subsequent stages of analysis. The removal of quotes, mentions, URLs, and hashtags is part of the preprocessing pipeline that is advised for both the VADER and BERT sentiment analysis tools. This preprocessing pipeline improves the data by removing words from the text that could otherwise confuse the classifiers' interpretations. Even though they actually exist in the text, the meaningful sentiment analysis is not enhanced by these unnecessary parts. Moreover, this type of preprocessing reduces memory usage and speeds up training times, which not only optimises the data by reducing its feature space but also boosts computational efficiency, a crucial factor when working with big datasets.

4.5 Optimization Issues for VADER and BERT

This section explores several key challenges related to enhancing the sentiment analysis process to achieve the main goal of increasing the accuracy and reliability of sentiment classification. The chapter discusses these challenges and their effects through three distinct case studies. The main focuses include adjusting the VADER threshold for more accurate sentiment detection, optimising BERT dataset splits for balanced training and validation, the effect of using varying topics in training the classifier, and how effective it is to use a dataset with a mix of topics. These methodological decisions aim to enhance the model's capability to accurately classify sentiments and optimise the performance of sentiment analysis models.

4.5.1 VADER Thresholds

Research that uses VADER for sentiment classification frequently mentions adopting a 0.05 threshold to distinguish between the negative (NEG), neutral (NEU), and positive (POS) sentiment classes, even though many of these studies do not explicitly mention the threshold settings. For our analysis in this study, there are two experiments to find the optimal threshold: finding the most frequent VADER score of all NEU posts and evaluating different thresholds.

First, the most frequent VADER score for all NEU posts was found in all datasets. As seen in Figure 4.7, 0 is the most frequent score for NEU posts in all datasets by 31.78%, 42.93%, and 40.67%, respectively. The second and what follows all have a very small percentage, and they are not close to zero.

Figure 4.7 Three bar charts show the most frequent VADER scores in all the posts that are labelled as in the three datasets

As a result of these observations, it was found that the most frequent VADER score for NEU posts is 0 in all three datasets. For a clearer understanding of which is the better threshold, a variety of threshold values will be tested on the three datasets in order to determine whether or not the 0.05 threshold really maximises the accuracy of sentiment classification.

For the COVID-19 Vaccine dataset, as seen in Table 4.12, adjustments to the VADER threshold from 0 to 1 demonstrate incremental enhancements in precision across the negative (NEG) sentiment class. However, this precision gain is coupled with a gradual decline in recall, indicating a trade-off between accurately identifying negative sentiments and the model's ability to capture all relevant instances. Neutral (NEU) and positive (POS) classes exhibit similar trends, where a higher threshold improves precision at the cost of recall, especially notable in the POS class with its modest increase in precision against a stable recall rate.

Similar findings are seen in the Electric Cars dataset, as seen in Table 4.13, where the NEG class shows a slight increase in precision at higher thresholds at the expense of recall. With thresholds at 0.08 and 0.09, the NEU class's performance peaks. This suggests that a higher threshold improves the model's focus on neutral sentiments. This highlights the balance required to maximise sentiment classification. The POS class again gains in precision without an obvious effect on recall.

In the Vegetarianism dataset, as seen in Table 4.14 the pattern differs slightly; the NEG class observes an ideal combination of recall and precision at a threshold of 0.08. There is an obvious advantage to decreasing the classifier's sensitivity to neutral sentiments, as the NEU class's performance improves gradually with a higher threshold. Raising the thresholds can more accurately identify positive sentiments, as the POS class shows, with improved precision and stable recall.

Table 4.12 VADER performance comparison with different thresholds in the COVID-19 Vaccine dataset

Dataset 1: COVID-19 Vaccine		**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**1**
NEG: 1073 posts	**Prec.**	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
	Rec.	0.57	0.57	0.57	0.57	0.57	0.56	0.56	0.56	0.55	0.55	0.55
	F1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
NEU: 1652 posts	**Prec.**	0.81	0.81	0.81	0.80	0.80	0.80	0.79	0.79	0.79	0.79	0.79
	Rec.	0.34	0.34	0.34	0.35	0.36	0.36	0.37	0.37	0.38	0.39	0.39
	F1	0.48	0.48	0.48	0.49	0.49	0.49	0.50	0.50	0.52	0.52	0.52
POS: 275 posts	**Prec.**	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.20	0.20	0.20
	Rec.	0.82	0.82	0.82	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
	F1	0.30	0.30	0.30	0.31	0.31	0.31	0.31	0.31	0.32	0.32	0.32
Accuracy		0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.48	0.48	0.49	0.49

Table 4.13 VADER performance comparison with different thresholds in the Electric Cars dataset

Dataset 2: E cars		**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**1**
NEG: 598 posts	**Prec.**	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.54	0.54	0.54
	Rec.	0.49	0.49	0.49	0.48	0.48	0.48	0.47	0.47	0.47	0.47	0.47
	F1	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50
NEU: 2094 posts	**Prec.**	0.87	0.87	0.87	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85
	Rec.	0.46	0.46	0.46	0.46	0.46	0.46	0.47	0.47	0.49	0.49	0.49
	F1	0.60	0.60	0.60	0.60	0.60	0.60	0.61	0.61	0.62	0.62	0.62
POS: 308 posts	**Prec.**	0.19	0.19	0.19	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
	Rec.	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83
	F1	0.31	0.31	0.31	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
Accuracy		0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.52

Table 4.14 VADER performance comparison with different thresholds in the Vegetarianism dataset

Dataset 3: Vegetarianism		**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**1**
NEG: 761 posts	**Prec.**	0.68	0.66	0.69	0.69	0.69	0.69	0.69	0.69	0.70	0.70	0.70
	Rec.	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.48	0.48	0.47
	F1	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.56
NEU: 1473 posts	**Prec.**	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77	0.76	0.76	0.76
	Rec.	0.43	0.43	0.44	0.44	0.44	0.44	0.44	0.44	0.45	0.45	0.45
	F1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.57	0.57	0.57
POS: 766 posts	**Prec.**	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
	Rec.	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	F1	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Accuracy		0.56	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57

Overall, the results imply that raising the VADER threshold generally increases precision for all sentiment classes; recall may suffer as a result, especially for the NEG class. The best threshold seems to differ between datasets, suggesting that the right balance between precision and recall is influenced by the context and type of data. These results highlight the importance of evaluating the threshold when using VADER for sentiment analysis, since it greatly impacts its ability to precisely and effectively categorise sentiments.

According to the results obtained from the tables of the three datasets (COVID-19 Vaccine, Electric Cars, and Vegetarianism), For both the NEG and POS classes, higher thresholds enhance precision while diminishing recall. This is because the model requires more explicit sentiment indicators to classify expressions as either negative or positive, thereby enhancing the precision of these classifications. However, this may cause some expressions to be missing that convey softer emotions with milder sentiment indicators, thereby diminishing the recall of these classes. Conversely, lower thresholds have the opposite effect; the model becomes more inclusive, including posts with less noticeable sentiment indicators but potentially misclassifying non-negative or non-positive expressions as NEG or POS. This improves recall and worsens the precision of these classes.

For the NEU class, the conditions are different; higher thresholds result in greater recall but less precision. Higher thresholds cause the model to be more careful when classifying expressions as positive or negative, which makes it easier to identify truly neutral expressions and improves the recall of the NEU class. However, this decrease in precision in the NEU class is because higher thresholds result in more expressions being considered NEU. On the other hand, with lower thresholds, the model becomes more aggressive in labelling posts as POS or NEG. This leads to increased precision by lowering the misclassification of non-neutral posts, but it may also miss some really neutral sentiments, which have an impact on recall.

In summary, the 0.05 threshold appears to be a reasonable option in multiple aspects. It generally improves the overall accuracy of sentiment classification with VADER without more compromise in recall observed at higher thresholds. While the 0.05 threshold may not be the absolute best in every single metric or dataset, it represents a reliable and balanced choice for optimising VADER's sentiment classification. It manages to maintain a good balance between identifying specific sentiments accurately (precision) and capturing the majority of relevant sentiment instances (recall). In addition, the choice of the optimal threshold also depends on the particular needs of an analyst. For example, a different threshold that maximises recall for the negative class might be better if the goal is to fully identify all negative posts. This experiment highlights the importance of matching the threshold setting to the particular requirements of the sentiment analysis.

4.5.2 BERT Dataset Size and Splits

BERT is a very useful tool with very accurate results. But this higher accuracy has a cost, in terms of the computational resources required for training. The purpose of this experiment is to investigate the relationship between the size of the dataset and the dataset splits (the separation of the data into parts for training, validation, and testing) effect on sentiment classification. These splits' size and balance are crucial factors that affect sentiment analysis's effectiveness. This investigation is essential for expanding our knowledge of BERT's use in sentiment analysis, especially in situations where resource limitations are a factor.

80% is usually used in most of the studies to train the classifier. Prasad et al. (2017) used different training and testing ratios in applying six different techniques to the sarcastic classification task. They applied Random Forest, Gradient Boosting, Decision Tree, Adaptive Boost, Logistic Regression, and Gaussian Naïve Bayes to 2000 manually labelled posts. They used the following splits: 60:40, 70:30, and 80:20 (the first number presenting the training percentage and the second presenting the testing percentage). They applied each tool twice (with and without emojis and slang words). The accuracies are not changing that much; the average of the maximum changes in the accuracies is 2.8%. In addition, the 80:20 split always gave the highest accuracy; two times 60:40 gave the highest accuracy, three times 70:30, and seven times 80:20.

This section examines the use of different dataset splits: 80:10:10, 70:10:20, and 60:10:30, across all datasets having different dataset sizes 3000, 2000, and 1000 posts. As seen in Table 4.15, Table 4.16, and Table 4.17, across all three datasets, in terms of dataset size, accuracy tends to be higher with larger dataset sizes; as the dataset is bigger, the performance is better. In general, the performance of the 3000-posts dataset is better than that of the 2000-posts dataset, and the 1000-posts dataset achieves the worst performance. In terms of data split, with a 3000-post dataset, the best split is 80:10:10. However, with the 2000-post dataset, 60:10:30 is giving the best performance. With a 1000-post dataset, the metric values of 70:10:20 are better for the first and last datasets, and 80:10:10 is better split for the second dataset. Moreover, the experiment reveals that the size of each class (POS, NEG, or NEU) is another factor influencing performance, which is the number of posts in each class. Extremely imbalanced datasets would suffer more when having smaller training datasets, because this would not offer enough information for the model to make a good generalisation.

Table 4.15 Compare BERT performance with different training dataset sizes and different dataset splits in COVID-19 Vaccine dataset

Dataset size		**3000**			**2000**			**1000**		
split ratio		**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**
NEG	**Prec.**	0.88	0.86	0.85	0.79	0.78	0.79	0.74	0.84	0.74
	Rec.	0.83	0.86	0.80	0.85	0.80	0.85	0.83	0.83	0.81
	F1	0.86	0.86	0.83	0.82	0.79	0.82	0.77	0.84	0.78
NEU	**Prec.**	0.88	0.87	0.86	0.88	0.85	0.87	0.88	0.86	0.89
	Rec.	0.92	0.88	0.89	0.83	0.86	0.86	0.78	0.85	0.81
	F1	0.90	0.88	0.87	0.85	0.86	0.86	0.83	0.86	0.85
POS	**Prec.**	0.68	0.67	0.69	0.60	0.73	0.67	0.67	0.52	0.44
	Rec.	0.63	0.62	0.66	0.67	0.61	0.55	0.89	0.58	0.54
	F1	0.65	0.64	0.68	0.63	0.67	0.60	0.76	0.55	0.48
Accuracy		0.86	0.85	0.84	0.82	0.82	0.83	0.80	0.82	0.78

Table 4.16 Compare BERT performance with different training dataset sizes and different dataset splits in Electric Cars dataset

Dataset size		**3000**			**2000**			**1000**		
split ratio		**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**
NEG	**Prec.**	0.83	0.81	0.74	0.81	0.74	0.78	0.74	0.76	0.83
	Rec.	0.72	0.66	0.75	0.65	0.78	0.73	0.70	0.65	0.63
	F1	0.77	0.73	0.74	0.72	0.76	0.75	0.72	0.70	0.72
NEU	**Prec.**	0.92	0.89	0.89	0.88	0.90	0.90	0.89	0.86	0.88
	Rec.	0.92	0.93	0.92	0.93	0.91	0.92	0.90	0.87	0.92
	F1	0.92	0.91	0.90	0.90	0.90	0.91	0.89	0.86	0.90
POS	**Prec.**	0.50	0.57	0.49	0.55	0.55	0.59	0.60	0.36	0.50
	Rec.	0.63	0.61	0.34	0.55	0.44	0.59	0.60	0.43	0.55
	F1	0.56	0.59	0.40	0.55	0.49	0.59	0.60	0.39	0.52
Accuracy		0.85	0.85	0.83	0.84	0.84	0.85	0.83	0.78	0.83

Table 4.17 Compare BERT performance with different training dataset sizes and different dataset splits in Vegetarianism dataset

Dataset size		**3000**			**2000**			**1000**		
split ratio		**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**	**80:10:10**	**70:10:20**	**60:10:30**
NEG	**Prec.**	0.74	0.76	0.71	0.74	0.77	0.81	0.63	0.65	0.63
	Rec.	0.76	0.78	0.72	0.73	0.71	0.72	0.65	0.73	0.69
	F1	0.75	0.77	0.71	0.73	0.74	0.77	0.64	0.69	0.66
NEU	**Prec.**	0.84	0.81	0.79	0.77	0.74	0.79	0.71	0.76	0.75
	Rec.	0.82	0.82	0.78	0.80	0.80	0.81	0.76	0.69	0.75
	F1	0.83	0.82	0.78	0.78	0.77	0.80	0.73	0.72	0.75
POS	**Prec.**	0.73	0.74	0.68	0.67	0.66	0.70	0.67	0.66	0.62
	Rec.	0.74	0.71	0.67	0.65	0.60	0.74	0.56	0.69	0.57
	F1	0.73	0.72	0.68	0.66	0.63	0.72	0.61	0.67	0.59
Accuracy		0.79	0.78	0.74	0.74	0.73	0.77	0.68	0.70	0.69

In summary, as the training dataset size increases, the performance of the classification increases, and vice versa. There is no best data split that always gives the best classification results based on the size of the dataset. Larger datasets require more computational resources; thus, careful consideration is necessary when selecting the optimal split. The choice of the best dataset split depends on the dataset's characteristics and specific requirements. This highlights the importance of considering data splits to optimise BERT's performance in sentiment analysis. The findings suggest that while BERT is capable of achieving high accuracy in sentiment classification, the configuration of training, validation, and testing data plays a critical role in maximising its effectiveness.

4.5.3 Use Different Topic Dataset to Train the Classifier

In order to ease the use of BERT, there is an idea to use different topic datasets to train a classifier. In this section, using different topics to train the BERT classifier is examined. Each dataset is divided into two parts: the training dataset (2000 posts) and the testing dataset (1000 posts). Each of the three training datasets is used to train the classifier and is tested by all three other testing datasets. The training dataset is divided into two parts: 90% (1800) to train the classifier and 10% (200 posts) to validate the classifier and to fine-tune the model's parameters. Stratified is to ensure each group (POS, NEU, and NEG) presents with the same percentages to have the best representation for the dataset.

As seen in Table 4.18, Table 4.19 and Table 4.20, it is observed that using the same dataset topic for classifier training generally results in better performance, and using other topics to train the classifier gives reasonable performance. However, in some classes, there is an issue when there are insufficient posts in a particular class within the training dataset. For example, as seen previously in Figure 4.8, dataset 1 has 9.2% POS posts. Consequently, this can impact the classifiers performance when using dataset 1 to train dataset 3, which has a percentage of POS posts of 25.5%. Furthermore, dataset 3 has the lowest percentage of NEU posts (49.1%) compared to all other classes, which means that when dataset 3 is used to train the classifier for datasets 1 or 2, it performs the worst.

Table 4.18 Comparing BERT performance using Dataset 1 as a testing dataset

Training dataset		Dataset 1	Dataset 2	Dataset 3
NEG:357 posts	Prec.	0.85	0.81	0.69
	Rec.	0.76	0.71	0.68
	F1	0.81	0.76	0.69
NEU: 551 posts	Prec.	0.84	0.79	0.77
	Rec.	0.89	0.90	0.80
	F1	0.87	0.84	0.78
POS: 92 posts	Prec.	0.57	0.68	0.64
	Rec.	0.60	0.46	0.54
	F1	0.59	0.55	0.59
Accuracy		0.82	0.79	0.73

Table 4.19 Comparing BERT performance using Dataset 2 as a testing dataset

Training dataset		Dataset 1	Dataset 2	Dataset 3
NEG: 199 posts	Prec.	0.73	0.75	0.81
	Rec.	0.70	0.71	0.55
	F1	0.72	0.73	0.66
NEU: 698 posts	Prec.	0.91	0.90	0.85
	Rec.	0.90	0.92	0.94
	F1	0.91	0.91	0.89
POS: 103 posts	Prec.	0.58	0.58	0.55
	Rec.	0.66	0.55	0.52
	F1	0.62	0.57	0.54
Accuracy		0.84	0.84	0.82

Table 4.20 Comparing BERT performance using Dataset 3 as a testing dataset

Training dataset		Dataset 1	Dataset 2	Dataset 3
NEG: 254 posts	Prec.	0.52	0.58	0.73
	Rec.	0.89	0.83	0.72
	F1	0.65	0.68	0.72
NEU: 491 posts	Prec.	0.85	0.81	0.78
	Rec.	0.57	0.65	0.78
	F1	0.68	0.72	0.78
POS: 255 posts	Prec.	0.66	0.68	0.66
	Rec.	0.60	0.64	0.68
	F1	0.63	0.66	0.67
Accuracy		0.66	0.69	0.74

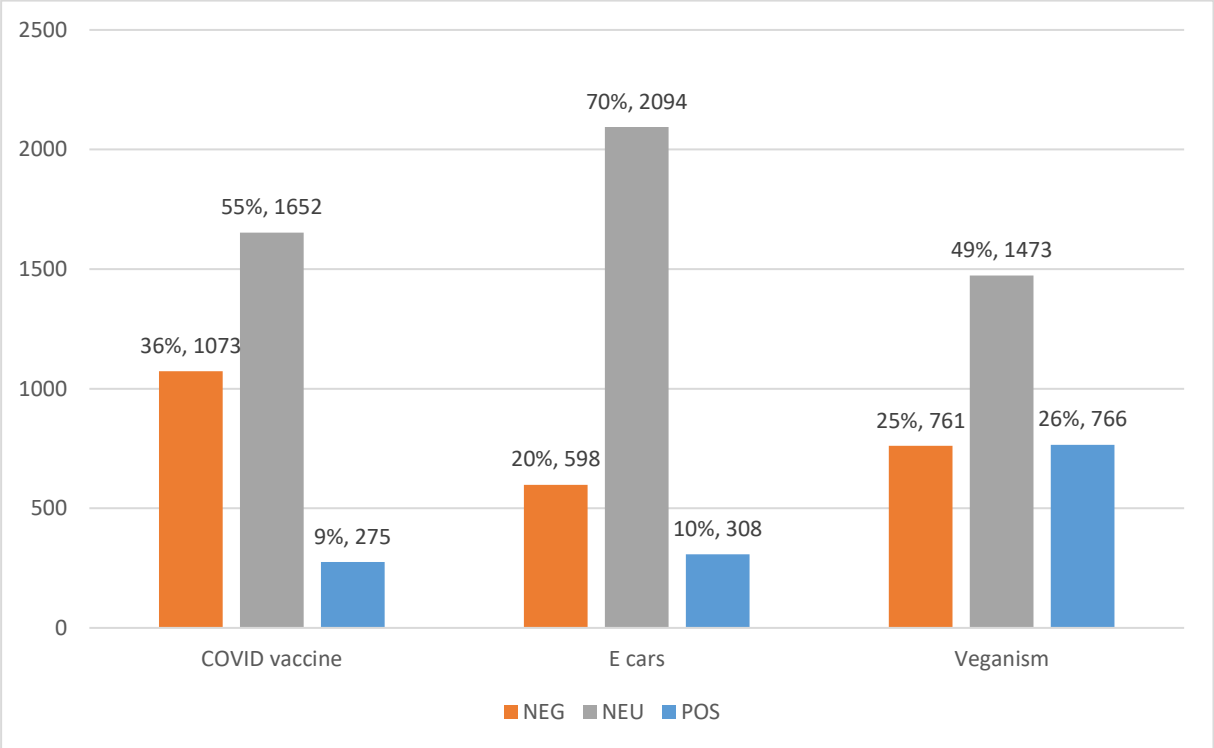

Figure 4.8 Bar chart represents the distribution of the post's sentiments (NEG, NEU, POS) of the three datasets

These findings highlight how the proportion of each class within a dataset influences the classifier's performance. It is important to have a balanced class distribution, guiding the optimisation of classifiers to enhance the performance of all the classes. Furthermore, it's

acceptable to use an imbalanced dataset if the focus is on classifying specific sentiments, such as NEG posts, where training the classifier with a dataset mostly composed of the NEG class can yield better specificity and performance for the targeted sentiment.

4.5.4 Use Mix-Topic Dataset

This section explores whether training a classifier on datasets with the same topic content is necessary to achieve the best sentiment analysis performance. Traditionally, it has been believed that classifier performance is highly impacted by how much the content of the dataset is concentrated on particular topics. This study seeks to examine this idea by combining various datasets into a single mixed-topic collection and investigating how topic variety affects the classifier's accuracy in detecting sentiment.

Each of the three datasets is divided into two parts: 2000 posts and 1000 posts. In creating the training dataset, 2000 posts from each dataset were combined into a singular large dataset that has 6000 posts, and then 2000 posts were randomly chosen. There are three testing datasets, one for each dataset, and each one has 1000 posts. The stratification method in selecting posts was used all the time to ensure that the sample accurately reflected the overall distribution of classes within the dataset. The experiment has two parts; it examines the performance of each testing dataset twice, by using for training the classifier a mixed-topic dataset and a dataset of the same topic.

In the first half of the experiment, the BERT classifier was trained using the mixed-topic dataset that has 2000 posts, and 10% of this dataset is used for validating the classifier while training. Second, the BERT classifier was trained and tested three times using the same topic in the training and testing datasets. The performance of these two parts is illustrated in Table 4.21 and Table 4.22.

Table 4.23 shows the distribution of the three sentiment classes of all the datasets that were used for training and testing in the second half of the experiment.

In this analysis, BERT performance is compared between the mixed-topic and same-topic datasets. However, it was observed that a class's performance correlates with its size in the training dataset; a greater number of posts within a specific class generally leads to enhanced performance. The mixed-topic training dataset has a higher proportion of NEU posts (58%) compared to NEG (27%) and POS (15%). This skew towards NEU posts in the training dataset contributes to the classifier's stronger performance in identifying NEU sentiments. As observed in Table 4.21, the performance of the NEU class across all three testing datasets. Exactly the opposite is observed with the performance of the NEU class across all three testing datasets,

which is the lowest. It is exactly the same when observing the BERT performance when the same topic is used for training the classifier, as seen in Table 4.22. BERT's performance demonstrates adaptability to both mixed and same-topic datasets, with slight variations in performance across sentiment classes. This is due to the composition of the training and testing datasets. The NEU class consistently shows strong performance due to its predominant representation in all the training datasets. This suggests the importance of dataset preparation to enhance classifier performance for specific sentiment classes.

Overall, BERT demonstrates robustness across both mixed-topic and same-topic scenarios; the classifier tends to perform marginally better when the training and testing data are on the same topic. This provides a more coherent learning environment for the model. The results underscore the importance of considering class distribution when training and testing sentiment analysis models.

Table 4.21 Compare BERT performance using a mix-topic dataset that has 2000 posts for training and three different topic datasets for testing; each testing dataset has 1000 posts

		Dataset 1	**Dataset 2**	**Dataset 3**
NEG	**Prec.**	0.78	0.73	0.66
	Rec.	0.80	0.70	0.72
	F1	0.79	0.72	0.69
NEU	**Prec.**	0.86	0.89	0.80
	Rec.	0.84	0.91	0.72
	F1	0.85	0.90	0.76
POS	**Prec.**	0.64	0.61	0.63
	Rec.	0.66	0.56	0.69
	F1	0.65	0.59	0.66
Accuracy		0.81	0.83	0.72

Table 4.22 Compare BERT performance using the same topic for training and testing datasets; the training dataset has 2000 posts, and the testing dataset has 1000 posts

		Dataset 1	Dataset 2	Dataset 3
NEG	**Prec.**	0.79	0.78	0.74
	Rec.	0.81	0.74	0.69
	F1	0.80	0.76	0.71
NEU	**Prec.**	0.84	0.90	0.78
	Rec.	0.87	0.91	0.79
	F1	0.86	0.91	0.79
POS	**Prec.**	0.64	0.61	0.67
	Rec.	0.47	0.63	0.70
	F1	0.54	0.62	0.69
accuracy		0.81	0.85	0.74

Table 4.23 The sentiment classes distribution of the training and each of the testing datasets

	NEG	NEU	POS
mixed_topic training dataset	540	1160	300
Dataset 1_testing	357	551	92
Dataset 2_testing	199	698	103
Dataset 3_testing	254	491	255
Dataset 1_training	716	1101	183
Dataset 2_training	399	1396	205
Dataset 3_training	507	982	511

4.6 Summary

In essence, the success of sentiment analysis research is mostly dependent on having a well-organised data pipeline. By ensuring that the analytical process is strong, dependable, and able to extract valuable insights from textual data, it eventually advances sentiment analysis techniques.

This chapter outlined the comprehensive data pipeline used in this research, detailing each stage from data acquisition through preprocessing. The datasets—COVID-19 Vaccine, Electric

Cars, and Vegetarianism—were collected, filtered for relevance, manually labelled, and refined to ensure the integrity of sentiment analysis results. Preprocessing steps, such as the removal of quotes, mentions, URLs, and hashtags, were evaluated across two classifiers, VADER and BERT, to assess their impact on classification performance.

This chapter marks the preparation of data for the next stages, which will focus on applying and evaluating the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA). However, the recommended preprocessing pipeline will not be applied when examining ED and SDA to ensure that nothing interferes with the results. These upcoming chapters will compare ED and SDA with other state-of-the-art methods, highlighting their contributions to emoji sentiment classification and sarcasm detection.

Chapter 5 Emoji Dictionary (ED)

Acknowledgment of Published Work

Parts of the work presented in this chapter, along with Chapter 6, have been published in the IEEE conference as part of the paper titled "Embracing Emojis in Sarcasm Detection to Enhance Sentiment". This paper, presented at The International Conference on Computer and Applications IEEE/ICCA'23 (Fifth Edition), focuses on the creation of the Emoji Dictionary (ED) and its role in sentiment analysis and sarcasm detection.

5.1 Introduction

With the rise of social media, there has been a shift in how people express emotions, thoughts, and sentiments. Emojis, initially introduced as simple pictographs, have evolved into a rich medium for communicating feelings and emotions in online conversations. Their widespread use on platforms like X, Instagram, and others presents both a challenge and an opportunity for sentiment analysis. Emojis carry context-dependent meanings that can substantially influence the interpretation of textual content. However, most Natural Language Processing (NLP) tools tend to overlook or remove emojis during preprocessing, treating them as noise rather than as valuable sentiment-bearing symbols (Barbieri et al., 2017; Chen et al., 2018).

This chapter presents the development and implementation of a specialized Emoji Dictionary (ED) to address this gap, allowing sentiment classifiers to effectively interpret and process emojis. The ED integrates emojis as key sentiment markers to improve sentiment analysis performance. This chapter details the creation of the ED, ED is compared with other existing dictionaries such as VADER, Demojize, and Emojinet using datasets across three different topics: COVID-19 Vaccine, Electric Cars, and Vegetarianism. These datasets allow for a comprehensive comparison of how well each dictionary handles the nuances of emoji usage in varied contexts.

The ED serves as a complementary layer that can be added to the sentiment analysis pipeline, functioning similarly to other dictionaries in the preprocessing stage but tailored to emoji interpretation. Through case studies and comparative experiments, this chapter demonstrates how ED enhances sentiment classification by offering a more accurate representation of sentiment-rich emojis within online conversations.

The ED aims to bridge the gap between the textual and symbolic (emoji) expressions of sentiment, offering a more comprehensive tool for modern sentiment analysis. By accounting for the diverse roles that emojis play in digital communication, the ED provides a more nuanced understanding of sentiment, particularly in scenarios where emojis carry substantial emotional or contextual weight.

5.2 Methodology

5.2.1 Data Collection and Preparation

The starting point for creating the ED was the collection of social media posts from three distinct datasets: COVID-19 Vaccine, Electric Cars, and Vegetarianism. For the creation of ED, only posts that contained emojis were retained to ensure that the classifier performance could be evaluated specifically in the context of emoji-rich content. Table 5.1 illustrates the number of posts retained after filtering out posts without emojis. As emojis constitute a significant portion of online conversations in these datasets, they present an excellent opportunity to understand their role in sentiment analysis.

Table 5.1 Dataset's sizes

	Number of posts	**Percentage of posts has emojis**	**Number of posts with emojis**
COVID-19 Vaccine	253,337	11.06%	280,26
Electric Cars	64,638	11.17%	722,0
Vegetarianism	173,128	26.23%	454,14

5.2.2 Frequent Emoji Selection

From each dataset, the top 250 most frequent emojis were identified. No more emojis have been chosen to include because the frequency of appearance of the last emojis in the list is too small in comparison with the most frequent emoji, as shown in Figure 5.1, Figure 5.2, and Figure 5.3, which represent the most frequent emojis in all the datasets. For example, the recurrence

of the 1st emoji '🚀' is 4782, and the recurrence of the 200th emoji '💭' is 28 in the COVID-19 Vaccine dataset.

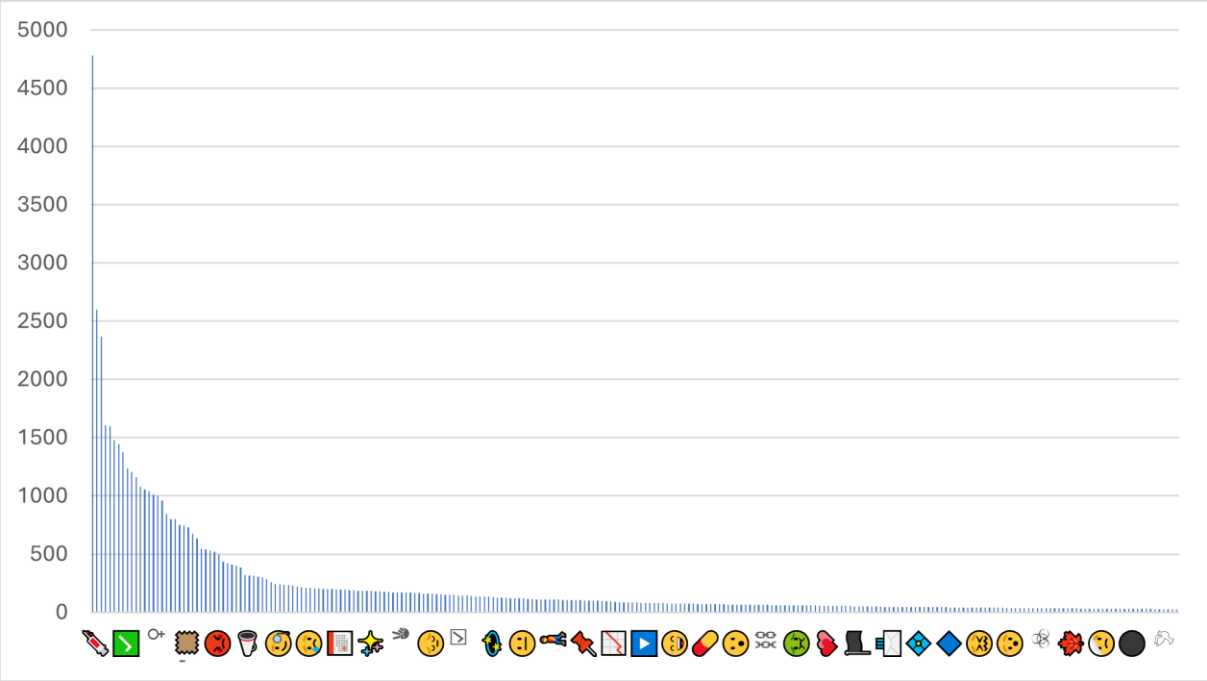

Figure 5.1 Bar chart showing the differences in the occurrence of some of top 250 frequent emojis in the COVID-19 Vaccine dataset

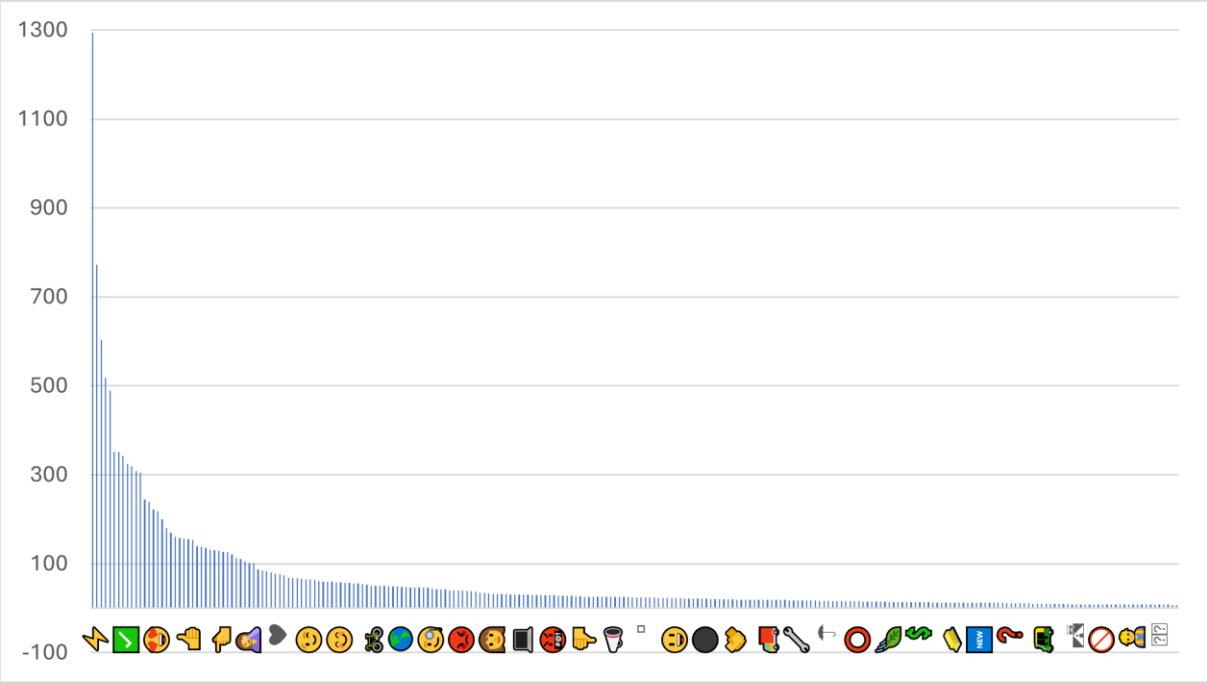

Figure 5.2 Bar chart showing the differences in the occurrence of some of top 250 frequent emojis in the Electric Cars dataset

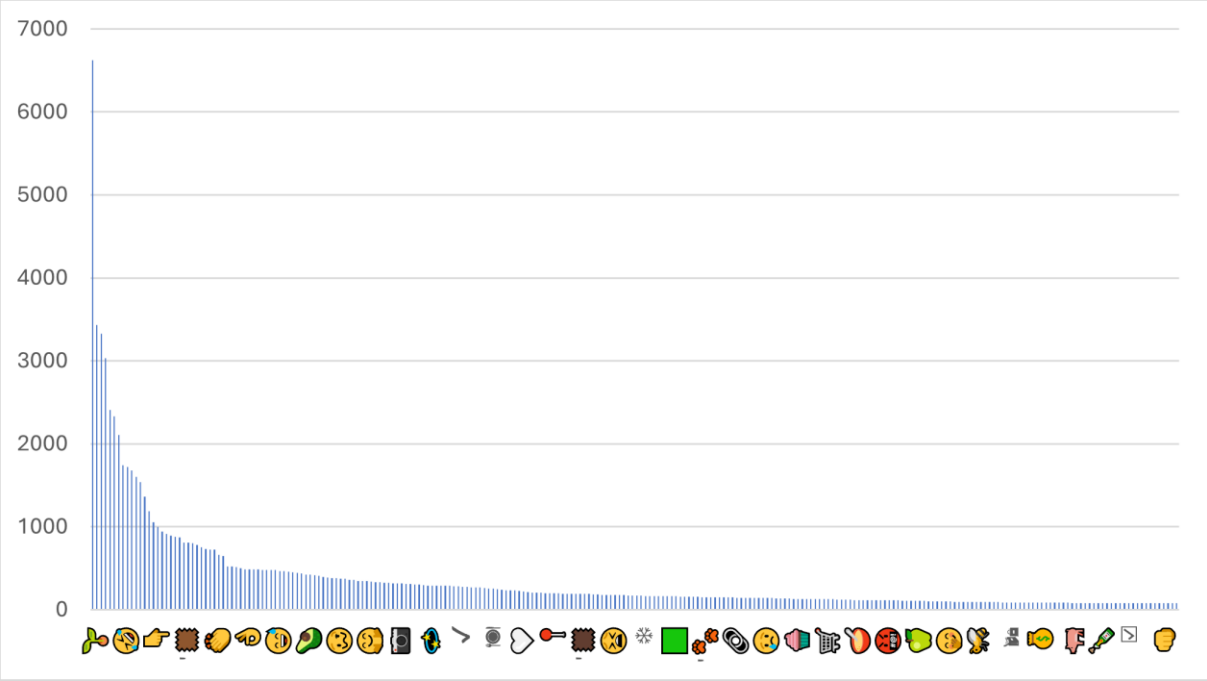

Figure 5.3 Bar chart showing the differences in the occurrence of some of top 250 frequent emojis in the Vegetarianism dataset

5.2.3 Creation of the Emoji Dictionary (ED)

Figure 5.4 outlines the workflow used to develop the ED. The process involved a combination of statistical analysis, sentiment scoring, and manual review by multiple annotators.

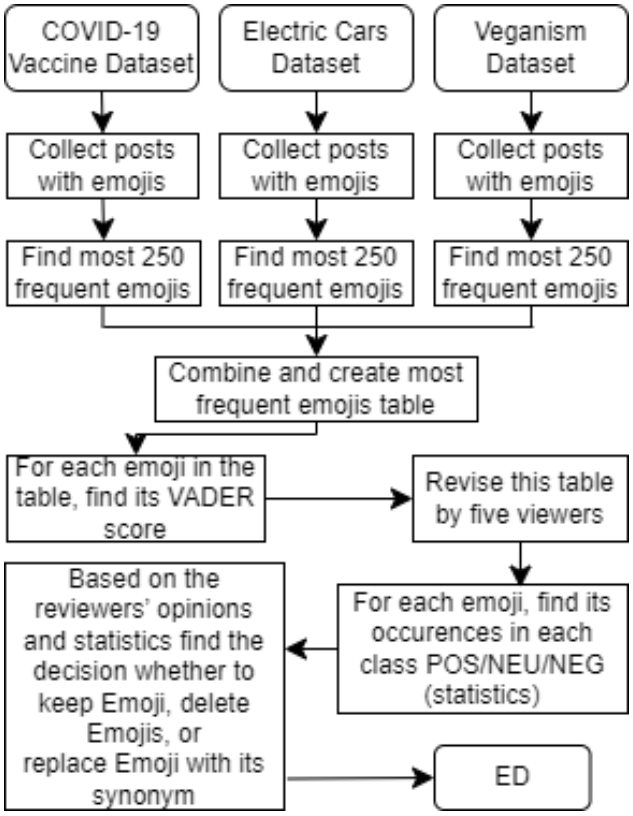

Figure 5.4 The research design used to create Emoji Dictionary (ED)

1. Emoji Frequency Analysis: The 250 most frequent emojis were first identified in each dataset and then combined into a unified list without repetition. This resulted in a list of 410 unique emojis.
2. Initial Sentiment Scoring with VADER: Each emoji in this list was assigned a sentiment score using VADER's sentiment analysis tool. VADER returns a score between -1 and +1, where -1 represents negative sentiment, +1 represents positive sentiment, and 0 represents neutral sentiment.
3. Manual Review by Annotators: Five human reviewers evaluated the VADER-assigned scores for each emoji. They were asked to either confirm the VADER score or propose an alternative score that better reflected their understanding of the emoji's meaning. This manual review process was essential for adjusting sentiment scores to align with human interpretations, refining the accuracy of the ED. The survey, as shown in Table 5.2,

provided valuable insights into the sentiment of each emoji, allowing for a refined understanding based on reviewers' input.

Table 5.2 Head of emojis table

emoji	description	VADER score [-1, +1]	Your opinion
😂	face with tears of joy	0.4404	
🤣	rolling on the floor laughing	0.4939	
👏	clapping hands	0	
😭	loudly crying face	-0.4767	
…	…	…	…

4. Statistical Analysis of Emoji Occurrences in Sentiment-Labelled Posts: Each emoji's frequency across positive (POS), neutral (NEU), and negative (NEG) posts was analysed using 9000 manually labelled posts, evenly distributed across the three datasets—3000 each from the COVID-19 Vaccine, Electric Cars, and Vegetarianism datasets. These posts were labelled by three annotators based on their sentiment. This analysis identified how often each emoji appeared within posts of different sentiment classes and informed further adjustments to the sentiment scores.
5. Final Decision Process: Using a combination of the reviewers' feedback and statistical data on emoji occurrences within POS/NEU/NEG posts, final decisions were made for each emoji. The possible actions were to keep the emoji with its existing sentiment score, remove the emoji entirely, or replace it with a more contextually accurate synonym. This systematic approach ensured that ambiguous or context-dependent emojis were refined to improve their contribution to sentiment analysis. A table was then created, categorizing each emoji based on the action taken—keep, remove, or replace—ensuring the final dictionary was fine-tuned for optimal performance.

5.2.4 Final Emoji Dictionary (ED)

as seen in Table 5.3, based on the combination of statistical analysis and reviewer feedback, each emoji was assigned one of three actions to ensure better sentiment classification:

- Keep: The emoji retains its original meaning.

- Replace with Synonym: The emoji is replaced with a more contextually appropriate synonym. For example, the "👏" emoji (clapping hands) was replaced with "proud."
- Remove: The emoji is removed if it doesn't convey any clear sentiment, as with "⏰" (alarm clock).

Table 5.3 Examples of emojis and the actions taken based on the analysis

VADER score	**Opinions and Statistics**	**Action**	**Example**
POS	NEU	Remove	▶, VADER score=0.34, remove
POS	NEG	Replace with synonym	😔, VADER score=0. 0772, Replace it with 'sad'
NEG	NEU	Remove	⏰, VADER score=-0.34, remove
NEU	POS	Replace with synonym	👏, VADER score= 0, Replace it with 'proud
NEU	NEG	Replace with synonym	🤬, VADER score= 0, Replace it with 'outburst of anger'
POS	POS	Keep	❤, VADER score= 06369, Keep it
NEG	NEG	Keep	😭, VADER score= -0.4767, Keep it

A small subset of emojis and their corresponding decisions is presented in Table 5.4. The complete list of 110 emojis with actions can be found in Appendix A. These emojis were carefully categorized to either keep, replace, or remove them based on both human interpretation and statistical analysis of their usage. This approach ensures a more accurate sentiment classification that reflects real-world usage and emotional context.

Table 5.4 Emojis with survey results, statistics, and decisions

emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
👏	clapping hands	0	0.3	0.3	1	0.3	1	17	2	1	Replace it with ‘proud’
😭	loudly crying face	-0.4767					-1	8	0	47	Keep
😳	flushed face	0	-0.1			-0.2	-0.4	1	0	15	Replace it with ‘embarrassment’
▶️	play button	0.34	0	0		0	0	1	4	1	Remove
😐	face with raised eyebrow	0	-0.1		-0.4	-0.2		0	0	11	Replace it with ‘suspicion’
⏰	alarm clock	-0.34	0	0		0	0	1	4	0	Remove
😒	unamused face	0	-0.2	-0.2	-1		-0.6	0	0	8	Replace it with ‘skepticism’
💩	pile of poo	0	-0.3	-0.6	-1	-0.3	-0.4	0	0	5	Replace it with ‘aggression’
💔	broken heart	0.2732	-0.6	-0.1	-1	-0.2	-0.8	0	0	6	Replace it with ‘grief’
✍️	writing hand	0.4939	0	0.1	0	0	0	1	4	1	Remove

5.3 Comparative Experiment: Evaluating the Performance of Sentiment Analysis Dictionaries

This section evaluates the Emoji Dictionary (ED) against three other emoji sentiment analysis tools—VADER, Demojize, and Emojinet—across seven datasets, which are divided into three topics (COVID-19 Vaccine, Electric Cars, and Vegetarianism). These datasets were selected to provide a range of contexts, with both general and emoji-specific subsets. The goal was to determine how well each dictionary captures emoji-based sentiment, using VADER and BERT for analysis. The metrics used to compare their performance include precision, recall, F1-score, and accuracy across different sentiment classes (NEG, NEU, POS). The dictionaries used in this comparison are as follows:

- VADER: VADER is a widely used rule-based tool for sentiment analysis, especially for social media text. It was used as a baseline in this comparison. VADER converts emojis into their textual descriptions before performing sentiment analysis. Despite its effectiveness in general sentiment analysis, it does not specifically focus on the nuanced role emojis play in sentiment expression.
- Demojize: Demojize is a Python package that converts emojis into their corresponding Unicode CLDR (Common Locale Data Repository) short names, essentially turning each emoji into a text label.
- Emojinet: Emojinet is a large machine-readable inventory of emoji senses that maps Unicode emoji representations to their English translations, helping provide context for emoji meanings. It aims to cover the broader spectrum of meanings associated with emojis.
- ED: The proposed Emoji Dictionary, designed as a specialized lexicon for emojis based on human-reviewed and statistically analysed sentiment associations.

The seven datasets used in this study include both general posts (which may or may not contain emojis) and emoji-only datasets. Here is a breakdown:

- Dataset 1 (COVID-19 Vaccine - General): 3,000 posts about COVID-19 vaccines, with 11.47% containing emojis.
- Dataset 2 (Electric Cars - General): 3,000 posts about electric cars, with 11.47% containing emojis.

- Dataset 3 (Vegetarianism - General): 3,000 tweets on vegetarianism, with a higher proportion (25.13%) of posts containing emojis.
- Dataset 4 (COVID-19 Vaccine - Emoji Only): 1,000 posts about COVID-19 vaccines, each containing at least one emoji.
- Dataset 5 (Electric Cars - Emoji Only): 1,000 posts on electric cars, all containing emojis.
- Dataset 6 (Vegetarianism - Emoji Only): 1,000 posts about vegetarianism, with all posts containing emojis.
- Dataset 7 (Combined): 12,000 posts from the combined general and emoji-only datasets, with 33.6% containing emojis.

Results and Analysis

5.3.1.1 VADER Sentiment Analysis Results

Each dictionary was evaluated by converting emojis to their text equivalents and applying VADER to analyse sentiment. Results were recorded across different datasets (see Tables for Dataset 1 to Dataset 7) to assess dictionary performance in the NEG, NEU, and POS sentiment classes.

The results from general datasets are presented in Table 5.5, Table 5.6, and Table 5.7, corresponding to Dataset 1 (COVID-19 Vaccine - General), Dataset 2 (Electric Cars - General), and Dataset 3 (Vegetarianism - General). Across these datasets, Emoji Dictionary (ED) demonstrates marginal improvements in the NEG and POS classes. For example, ED slightly improves F1-scores and precision compared to other dictionaries, indicating its capability to enhance sentiment interpretation even when emojis are used sparingly in general datasets. Other dictionaries, such as Demojize and Emojinet, also show comparable performance to VADER but fail to match ED's balance of precision and recall. The NEU class remains consistent across dictionaries with little to no variation, reflecting the inherent challenges of identifying neutrality in general textual datasets.

Table 5.5 Performance comparison of VADER with different dictionaries on COVID-19 Vaccine dataset (general)

Dictionary		**VADER**	**Demojize**	**Emojinet**	**ED**
NEG: 1073 posts	**Prec.**	0.55	0.55	0.55	0.55
	Rec.	0.56	0.56	0.56	0.57
	F1	0.55	0.55	0.55	0.56
NEU: 1652 posts	**Prec.**	0.79	0.78	0.79	0.80
	Rec.	0.33	0.33	0.33	0.33
	F1	0.47	0.47	0.47	0.47
POS: 275 posts	**Prec.**	0.18	0.18	0.18	0.18
	Rec.	0.80	0.79	0.80	0.81
	F1	0.29	0.29	0.29	0.30
Accuracy		0.46	0.46	0.46	0.46

Table 5.6 Performance comparison of VADER with different dictionaries on Electric Cars dataset (general)

Dictionary		**VADER**	**Demojize**	**Emojinet**	**ED**
NEG: 598 posts	**Prec.**	0.51	0.51	0.51	0.52
	Rec.	0.47	0.46	0.47	0.48
	F1	0.49	0.49	0.49	0.50
NEU: 2094 posts	**Prec.**	0.85	0.84	0.85	0.86
	Rec.	0.43	0.44	0.43	0.43
	F1	0.58	0.57	0.58	0.58
POS: 308 posts	**Prec.**	0.19	0.19	0.19	0.19
	Rec.	0.84	0.83	0.84	0.85
	F1	0.31	0.30	0.30	0.31
Accuracy		0.48	0.48	0.48	0.49

Table 5.7 Performance comparison of VADER with different dictionaries on Vegetarianism dataset (general)

Dictionary		VADER	Demojize	Emojinet	ED
NEG: 761 posts	**Prec.**	0.66	0.64	0.67	0.69
	Rec.	0.47	0.44	0.47	0.49
	F1	0.55	0.52	0.55	0.57
NEU: 1473 posts	**Prec.**	0.76	0.71	0.77	0.78
	Rec.	0.41	0.42	0.41	0.41
	F1	0.53	0.53	0.53	0.54
POS: 766 posts	**Prec.**	0.40	0.39	0.40	0.41
	Rec.	0.88	0.82	0.89	0.90
	F1	0.55	0.53	0.56	0.56
Accuracy		0.54	0.53	0.55	0.56

The results from emoji-rich datasets are detailed in Table 5.8, Table 5.9, and Table 5.10, corresponding to Dataset 4 (COVID-19 Vaccine - Emoji Only), Dataset 5 (Electric Cars - Emoji Only), and Dataset 6 (Vegetarianism - Emoji Only). ED consistently outperforms other dictionaries, particularly in the NEU and POS classes. ED shows notable improvements in F1-scores and overall classification accuracy. For instance, in Table 5.8, ED enhances recall and F1-scores for both NEG and POS classes, showcasing its robustness in handling emoji-rich contexts. NEU class enhancements are particularly noticeable in these datasets. ED achieves considerably higher precision and recall than Demojize and Emojinet, indicating its ability to handle ambiguous or nuanced sentiments often conveyed by emojis. Emojinet and Demojize, while improving performance over standard VADER, lack the consistency and balance observed with ED.

Table 5.8 Performance comparison of VADER with different dictionaries on COVID-19 Vaccine dataset (emoji only)

Dictionary		VADER	Demojize	Emojinet	ED
NEG: 646 posts	Prec.	0.76	0.77	0.75	0.89
	Rec.	0.53	0.53	0.53	0.77
	F1	0.62	0.63	0.62	0.83
NEU: 186 posts	Prec.	0.10	0.29	0.11	0.71
	Rec.	0.05	0.27	0.06	0.26
	F1	0.07	0.28	0.08	0.38
POS: 168 posts	Prec.	0.26	0.27	0.26	0.40
	Rec.	0.68	0.62	0.68	0.90
	F1	0.38	0.38	0.37	0.56
Accuracy		0.47	0.50	0.47	0.70

Table 5.9 Performance comparison of VADER with different dictionaries on Electric Cars dataset (emoji only)

Dictionary		VADER	Demojize	Emojinet	ED
NEG: 335 posts	Prec.	0.47	0.48	0.47	0.81
	Rec.	0.40	0.42	0.40	0.72
	F1	0.43	0.45	0.43	0.76
NEU: 323 posts	Prec.	0.15	0.26	0.15	0.74
	Rec.	0.06	0.15	0.06	0.20
	F1	0.08	0.19	0.08	0.31
POS: 342 posts	Prec.	0.43	0.46	0.43	0.52
	Rec.	0.76	0.70	0.76	0.94
	F1	0.55	0.56	0.55	0.67
Accuracy		0.41	0.43	0.41	0.63

Table 5.10 Performance comparison of VADER with different dictionaries on Vegetarianism dataset (emoji only)

Dictionary		**VADER**	**Demojize**	**Emojinet**	**ED**
NEG: 268 posts	**Prec.**	0.56	0.53	0.61	0.85
	Rec.	0.43	0.42	0.42	0.72
	F1	0.49	0.47	0.50	0.78
NEU: 72 posts	**Prec.**	0.00	0.03	0.01	0.31
	Rec.	0.00	0.10	0.01	0.19
	F1	0.00	0.05	0.01	0.24
POS: 660 posts	**Prec.**	0.74	0.75	0.77	0.85
	Rec.	0.73	0.65	0.84	0.93
	F1	0.74	0.70	0.81	0.89
Accuracy		0.59	0.55	0.67	0.82

The results from the combined dataset, presented in Table 5.11, aggregate findings from both general and emoji-rich datasets. This dataset provides a broader evaluation of dictionary performance across diverse contexts. ED maintains its superior performance, achieving the highest F1-scores and accuracy across all sentiment classes. Its ability to generalize well across varied topics and levels of emoji usage is evident from its consistent results. Other dictionaries show marginal enhancements in specific metrics but fail to achieve the same balance as ED.

Table 5.11 Performance comparison of VADER with different dictionaries on combined dataset (all topics)

Dictionary		VADER	Demojize	Emojinet	ED
NEG: 3518 posts	Prec.	0.56	0.56	0.56	0.64
	Rec.	0.50	0.50	0.50	0.60
	F1	0.53	0.53	0.53	0.62
NEU: 6421 posts	Prec.	0.75	0.72	0.76	0.83
	Rec.	0.37	0.39	0.37	0.39
	F1	0.50	0.50	0.50	0.53
POS: 2061 posts	Prec.	0.28	0.28	0.29	0.32
	Rec.	0.77	0.73	0.81	0.89
	F1	0.41	0.40	0.43	0.47
Accuracy		0.48	0.48	0.48	0.54

ED's specialization in handling emojis provides a notable advantage, particularly in emoji-heavy datasets, where its nuanced approach enhances both precision and recall. The NEU class, while consistently challenging across all datasets, sees pronounced enhancements with ED in emoji-rich contexts, emphasizing the importance of integrating emoji semantics in sentiment analysis. The results highlight the value of combining emoji-specific dictionaries with sentiment analysis tools like VADER to refine performance in both general and emoji-heavy datasets.

5.3.1.2 BERT Sentiment Analysis Results

BERT was employed with each dictionary under three data splits (50:10:40, 60:10:30, 70:10:20). Each split represents different proportions of training, validation, and testing sets, ensuring the robustness of the findings. Additionally, the runtime and maximum input sequence length for each dictionary were tracked, as shown in the tables provided.

The results for general datasets are presented in Table 5.12, Table 5.13, and Table 5.14, corresponding to Dataset 1 (COVID-19 Vaccine - General), Dataset 2 (Electric Cars - General), and Dataset 3 (Vegetarianism - General). Across these datasets, ED consistently demonstrates strong performance in the NEG and POS classes, often achieving the highest or near-highest precision and F1-scores. Its nuanced handling of emoji enhances its ability to capture

sentiment cues even when emojis are used sparingly. Demojize performs competitively, particularly in precision for the NEG class, but its performance is less consistent compared to ED. Emojinet occasionally achieves comparable results to ED but generally lags in recall and F1-scores for the NEG and POS classes. The NEU class exhibits the most stable performance across all dictionaries, with minor variations in metrics, reflecting the inherent challenges of neutral sentiment detection.

Table 5.12 Performance comparison of BERT with different dictionaries on COVID-19 Vaccine dataset (general) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.81	0.8	0.82	0.83	0.8	0.81	0.82	0.8	0.86	0.83	0.83	0.84
	Rec.	0.8	0.84	0.82	0.83	0.83	0.79	0.77	0.84	0.8	0.77	0.81	0.8
	F1	0.81	0.82	0.82	0.83	0.82	0.8	0.8	0.82	0.83	0.8	0.82	0.82
NEU	**Prec.**	0.85	0.85	0.86	0.86	0.87	0.85	0.84	0.87	0.87	0.85	0.87	0.85
	Rec.	0.88	0.85	0.87	0.88	0.86	0.89	0.89	0.84	0.89	0.89	0.88	0.89
	F1	0.86	0.85	0.87	0.87	0.87	0.87	0.87	0.86	0.88	0.87	0.88	0.87
POS	**Prec.**	0.6	0.6	0.59	0.64	0.64	0.61	0.63	0.58	0.58	0.6	0.63	0.68
	Rec.	0.47	0.47	0.55	0.53	0.57	0.52	0.56	0.57	0.64	0.6	0.62	0.65
	F1	0.53	0.53	0.57	0.58	0.6	0.57	0.59	0.58	0.61	0.6	0.62	0.67
Accuracy		0.81	0.81	0.82	0.83	0.83	0.82	0.82	0.82	0.84	0.82	0.83	0.83

Table 5.13 Performance comparison of BERT with different dictionaries on Electric Cars dataset (general) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.76	0.77	0.77	0.78	0.79	0.78	0.82	0.82	0.8	0.81	0.83	0.81
	Rec.	0.71	0.77	0.79	0.75	0.76	0.83	0.7	0.73	0.71	0.71	0.73	0.74
	F1	0.74	0.77	0.78	0.76	0.77	0.8	0.76	0.77	0.75	0.76	0.78	0.77
NEU	**Prec.**	0.89	0.9	0.9	0.91	0.9	0.91	0.9	0.9	0.91	0.89	0.92	0.92
	Rec.	0.92	0.9	0.9	0.89	0.93	0.92	0.91	0.93	0.93	0.94	0.93	0.93
	F1	0.91	0.9	0.9	0.9	0.92	0.91	0.91	0.91	0.92	0.92	0.92	0.92
POS	**Prec.**	0.59	0.61	0.6	0.56	0.7	0.71	0.55	0.65	0.61	0.67	0.64	0.63
	Rec.	0.55	0.59	0.54	0.67	0.6	0.54	0.64	0.65	0.65	0.58	0.71	0.69
	F1	0.57	0.6	0.57	0.61	0.64	0.62	0.59	0.65	0.62	0.62	0.67	0.66
Accuracy		0.84	0.84	0.84	0.84	0.86	0.87	0.84	0.86	0.86	0.86	0.87	0.86

Table 5.14 Performance comparison of BERT with different dictionaries on Vegetarianism dataset (general) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.7	0.73	0.73	0.69	0.68	0.71	0.72	0.68	0.72	0.73	0.71	0.73
	Rec.	0.71	0.78	0.75	0.74	0.68	0.7	0.71	0.66	0.75	0.77	0.78	0.74
	F1	0.71	0.75	0.74	0.72	0.68	0.7	0.72	0.67	0.74	0.75	0.75	0.74
NEU	**Prec.**	0.77	0.84	0.83	0.8	0.77	0.79	0.8	0.77	0.78	0.84	0.83	0.79
	Rec.	0.76	0.81	0.79	0.76	0.79	0.79	0.81	0.77	0.82	0.79	0.78	0.79
	F1	0.77	0.83	0.81	0.78	0.78	0.79	0.81	0.77	0.8	0.81	0.8	0.79
POS	**Prec.**	0.65	0.75	0.71	0.68	0.66	0.72	0.76	0.71	0.74	0.76	0.73	0.71
	Rec.	0.65	0.75	0.75	0.69	0.64	0.74	0.76	0.71	0.64	0.81	0.75	0.69
	F1	0.65	0.75	0.73	0.69	0.65	0.73	0.76	0.71	0.69	0.78	0.74	0.7
Accuracy		0.72	0.79	0.77	0.74	0.72	0.75	0.77	0.73	0.76	0.79	0.77	0.75

The results from emoji-only datasets are presented in Table 5.15, Table 5.16, and Table 5.17, corresponding to Dataset 4 (COVID-19 Vaccine - Emoji Only), Dataset 5 (Electric Cars - Emoji Only), and Dataset 6 (Vegetarianism - Emoji Only), respectively. In these datasets, ED consistently outperforms other dictionaries in all sentiment classes, achieving the highest precision, recall, and F1-scores. This reflects its effectiveness in handling emoji-rich content and extracting sentiment nuances. Demojize also performs strongly, particularly in precision for the NEG and POS classes, but its recall tends to be slightly lower than ED. Emojinet shows moderate enhancements over the Original BERT but often falls short of ED and Demojize, especially in recall for the NEG and POS classes. The NEU class remains the weakest across all approaches, highlighting the difficulty of detecting neutrality in emoji-heavy contexts.

Table 5.15 Performance comparison of BERT with different dictionaries on COVID-19 Vaccine dataset (emoji only) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.84	0.94	0.92	0.95	0.84	0.94	0.92	0.94	0.84	0.98	0.95	0.96
	Rec.	0.88	0.96	0.96	0.93	0.9	0.94	0.94	0.95	0.86	0.93	0.95	0.91
	F1	0.86	0.95	0.94	0.94	0.87	0.94	0.93	0.95	0.85	0.95	0.95	0.93
NEU	**Prec.**	0.67	0.81	0.78	0.75	0.65	0.78	0.78	0.74	0.76	0.8	0.82	0.74
	Rec.	0.62	0.74	0.73	0.77	0.62	0.77	0.77	0.82	0.42	0.87	0.84	0.76
	F1	0.64	0.77	0.76	0.76	0.64	0.77	0.77	0.78	0.54	0.84	0.83	0.75
POS	**Prec.**	0.6	0.75	0.78	0.81	0.55	0.76	0.8	0.84	0.34	0.72	0.72	0.72
	Rec.	0.51	0.76	0.7	0.82	0.44	0.76	0.72	0.72	0.48	0.79	0.7	0.85
	F1	0.55	0.76	0.74	0.81	0.49	0.76	0.76	0.77	0.4	0.75	0.71	0.78
Accuracy		0.77	0.89	0.87	0.89	0.77	0.88	0.87	0.89	0.71	0.9	0.89	0.87

Table 5.16 Performance comparison of BERT with different dictionaries on Electric Cars dataset (emoji only) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.73	0.91	0.89	0.92	0.8	0.95	0.9	0.88	0.74	0.94	0.91	0.88
	Rec.	0.78	0.96	0.93	0.93	0.77	0.9	0.85	0.92	0.68	0.92	0.91	0.88
	F1	0.76	0.93	0.91	0.92	0.79	0.92	0.88	0.9	0.71	0.93	0.91	0.88
NEU	**Prec.**	0.68	0.83	0.84	0.79	0.65	0.81	0.82	0.81	0.71	0.86	0.82	0.83
	Rec.	0.5	0.82	0.81	0.84	0.75	0.81	0.69	0.84	0.62	0.83	0.85	0.88
	F1	0.57	0.82	0.82	0.81	0.7	0.81	0.75	0.82	0.66	0.84	0.83	0.85
POS	**Prec.**	0.51	0.86	0.79	0.87	0.67	0.81	0.68	0.87	0.59	0.85	0.82	0.91
	Rec.	0.6	0.82	0.78	0.81	0.59	0.84	0.82	0.81	0.71	0.88	0.8	0.86
	F1	0.55	0.84	0.78	0.84	0.63	0.83	0.74	0.84	0.64	0.87	0.81	0.88
Accuracy		0.63	0.87	0.84	0.86	0.7	0.85	0.79	0.85	0.67	0.88	0.85	0.87

Table 5.17 Performance comparison of BERT with different dictionaries on Vegetarianism dataset (emoji only) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.8	0.92	0.87	0.91	0.79	0.95	0.93	0.91	0.78	0.83	0.87	0.84
	Rec.	0.63	0.91	0.82	0.9	0.66	0.9	0.84	0.93	0.6	0.85	0.75	0.87
	F1	0.7	0.91	0.85	0.91	0.72	0.92	0.88	0.92	0.68	0.84	0.81	0.85
NEU	**Prec.**	0.62	0.67	0.32	0.44	0.29	0.73	0.5	0.75	0.29	1	0.46	0.71
	Rec.	0.17	0.34	0.41	0.24	0.18	0.36	0.41	0.27	0.13	0.27	0.4	0.33
	F1	0.27	0.45	0.36	0.31	0.22	0.48	0.45	0.4	0.18	0.42	0.43	0.45
POS	**Prec.**	0.81	0.92	0.92	0.91	0.83	0.92	0.9	0.91	0.8	0.92	0.89	0.92
	Rec.	0.94	0.97	0.91	0.96	0.91	0.98	0.96	0.97	0.92	0.98	0.95	0.96
	F1	0.87	0.94	0.91	0.93	0.87	0.95	0.93	0.94	0.86	0.95	0.92	0.94
Accuracy		0.8	0.91	0.85	0.89	0.79	0.92	0.89	0.91	0.78	0.9	0.86	0.89

The results from the Combined Dataset (Table 5.18) integrate both general and emoji-only datasets, offering a comprehensive evaluation, Demojize consistently delivers strong performance across all sentiment classes, often achieving higher precision, recall, and F1-scores for the NEG and POS classes compared to ED. This suggests that Demojize's handling of emoji-to-text translations is highly effective for sentiment detection in diverse datasets. ED also performs well, with metrics that are closely aligned with Demojize, though it occasionally falls slightly behind in recall for the NEG and POS classes. Its balanced precision and F1-scores indicate robust performance but not necessarily an advantage over Demojize. Emojinet demonstrates moderate performance, generally lagging behind both Demojize and ED. While it achieves reasonable precision, its recall is often lower, particularly for the NEG class, leading to less balanced F1-scores. The NEU class metrics remain stable across all dictionaries, with minimal variation. This reflects the shared difficulty of accurately detecting neutrality in content influenced by sarcasm and emojis.

Table 5.18 Performance comparison of BERT with different dictionaries on Combined dataset (all topics) across three data splits

		50:10:40				60:10:30				70:10:20			
		Original	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**	**Original**	**Demojize**	**Emojinet**	**ED**
NEG	**Prec.**	0.84	0.9	0.86	0.87	0.87	0.89	0.88	0.89	0.87	0.9	0.9	0.9
	Rec.	0.81	0.88	0.87	0.86	0.84	0.91	0.89	0.89	0.86	0.91	0.9	0.88
	F1	0.83	0.89	0.87	0.86	0.86	0.9	0.89	0.89	0.86	0.9	0.9	0.89
NEU	**Prec.**	0.9	0.92	0.91	0.9	0.9	0.93	0.92	0.92	0.91	0.93	0.91	0.91
	Rec.	0.91	0.93	0.91	0.91	0.92	0.92	0.92	0.92	0.93	0.93	0.93	0.92
	F1	0.9	0.92	0.91	0.91	0.91	0.93	0.92	0.92	0.92	0.93	0.92	0.92
POS	**Prec.**	0.74	0.84	0.8	0.82	0.77	0.85	0.83	0.82	0.78	0.87	0.86	0.83
	Rec.	0.74	0.83	0.78	0.81	0.75	0.83	0.82	0.81	0.77	0.85	0.8	0.83
	F1	0.74	0.84	0.79	0.82	0.76	0.84	0.82	0.81	0.77	0.86	0.83	0.83
Accuracy		0.85	0.9	0.88	0.88	0.87	0.9	0.89	0.9	0.88	0.91	0.9	0.89

ED and Demojize demonstrate closely aligned performance across most datasets, with neither consistently outperforming the other. ED generally provides balanced precision, recall, and F1-scores across all sentiment classes, while Demojize occasionally excels in certain metrics, particularly precision for the NEG class. The differences between ED and Demojize are generally marginal, with both dictionaries markedly outperforming Emojinet and the Original BERT in emoji-rich datasets. This reflects their superior handling of emoji translations for sentiment analysis. Emojinet, while improving over the Original BERT, struggles to match the performance of ED and Demojize, particularly in recall for the NEG and POS classes. This suggests that its emoji semantics are less effective in detecting nuanced sentiments. The NEU class consistently shows lower performance across all dictionaries and datasets, highlighting the inherent difficulty of detecting neutrality, especially in sarcastic or emoji-heavy contexts.

Table 5.19 compares the average running times and maximum sequence lengths (Max Tokens) for BERT using different dictionaries (Original, Demojize, Emojinet, ED) on dataset 7 (the combined dataset across all topics). This comparison highlights notable differences in computational efficiency among the dictionaries.

The Original and ED dictionaries demonstrate the fastest running times across all splits, with a maximum sequence length of 164 tokens, the shortest among the dictionaries. This shorter sequence length likely contributes to their quicker processing times by reducing memory requirements and computational load.

Dictionaries like Demojize, which expand emojis into detailed textual descriptions, result in significantly longer processing times due to their increased sequence length (Max Tokens = 359). Emojinet also exhibits longer processing times than Original and ED, with a maximum sequence length of 233 tokens.

These results underscore a trade-off between the level of emoji interpretability and computational efficiency. For tasks where processing speed is crucial, Original and ED are preferable due to their faster processing times and reduced sequence lengths. ED offers similar efficiency to Original but potentially enhances sentiment analysis with more refined emoji sentiment handling.

In summary, Table 5.19 illustrates that choosing the right dictionary depends on the specific demands of the sentiment analysis task, balancing interpretative depth and computational feasibility.

Table 5.19 Average running time and max sequence length (Max_Len) for BERT with different dictionaries on the Combined dataset.

	Average running time	**Max Tokens**
original	17m 12s	164
Demojize	20h 11m 11s	359
Emojinet	6h 34m 39s	233
ED	17m	164

5.4 Summary

This chapter introduced the Emoji Dictionary (ED) as a specialized lexicon designed to enhance sentiment classification in emoji-rich content. This chapter compared ED with other dictionaries, including VADER, Demojize, and Emojinet, across multiple datasets.

In the VADER-based analysis, ED demonstrated moderate changes in general datasets and stronger performance in emoji-only datasets. Its specialized handling of emojis provided a balanced sentiment interpretation, particularly when emojis were prominently featured. Demojize also performed strongly, often achieving metrics close to or on par with ED, especially in emoji-rich datasets. Emojinet showed slight advancements but generally lagged behind ED and Demojize in precision, recall, and F1-scores for most sentiment classes.

In the BERT-based sentiment analysis, no single dictionary consistently outperformed others across all metrics for each sentiment class in all datasets and data splits. ED and Demojize frequently achieved comparable performance, particularly in emoji-only datasets, where their nuanced handling of emojis proved effective. Running time analysis revealed that ED and the Original BERT had shorter processing times due to their lower maximum sequence lengths, while Demojize incurred higher computational overhead. Both ED and Demojize are recommended for datasets with a high frequency of emojis due to their competitive accuracy in emoji interpretation. However, for applications prioritizing faster processing times, ED or the Original BERT may be preferred.

The results highlight that ED and Demojize offer balanced and effective approaches to sentiment classification, particularly in emoji-rich contexts. The choice between dictionaries should consider the specific dataset characteristics (e.g., emoji frequency) and available

computational resources, as the performance differences between ED and Demojize are generally marginal and context-dependent.

Chapter 6 Sarcasm Detection Approach (SDA)

Acknowledgment of Published Work

Parts of the work presented in this chapter have been published in the IEEE conference as part of the paper titled "Embracing Emojis in Sarcasm Detection to Enhance Sentiment". This paper, presented at The International Conference on Computer and Applications IEEE/ICCA'23 (Fifth Edition), emphasizes the integration of emojis in sarcasm detection to enhance sentiment classification.

6.1 Introduction

Sarcasm detection plays a pivotal role in sentiment analysis, particularly when analysing the nuanced and often complex communication patterns found in social media posts. Sarcasm, which uses irony to mock or convey a different meaning, can make it hard to understand the true emotional tone, creating challenges for sentiment analysis. For instance, a sarcastic comment may appear positive on the surface but conveys a negative sentiment, leading to misclassification in traditional sentiment analysis systems.

In digital communication, emojis further complicate sentiment interpretation. While initially introduced as a way to enhance textual communication with visual cues, emojis have evolved into powerful tools for expressing emotions and sarcasm. Their contextual nature means the same emoji can convey drastically different meanings depending on the accompanying text. This presents both challenges and opportunities for sentiment analysis systems, particularly when sarcasm is heavily reliant on emojis.

This study introduces a Sarcasm Detection Approach (SDA) that identifies sarcasm by detecting conflicts between the sentiment of text and emojis. When such a conflict is found, the sentiment is reclassified as negative. To enhance its accuracy, the approach incorporates the Emoji Dictionary (ED), which is specifically designed for precise sentiment interpretation of emojis.

A comparative analysis was conducted to evaluate the performance of the SDA+ED approach against two other sarcasm detection methods that are more generalized and do not focus on emojis:

- Logistic Regression: Based on the paper "Logistic Regression Method for Sarcasm Detection of Text Data," this approach uses logistic regression trained on a labelled dataset (sarcastic and non-sarcastic) to classify sarcasm.
- WELMSD: As described in the paper "WELMSD – Word Embedding and Language Model-based Sarcasm Detection," this method combines FastText embeddings with BERT as a context-aware language model. It leverages automatic feature engineering and a robust evaluation framework to improve sarcasm detection.

These approaches are evaluated using two sentiment classification methods—VADER (a lexicon-based tool) and BERT (a machine learning model). Each method operates on datasets divided into three sentiment classes: Positive (POS), Neutral (NEU), and Negative (NEG). The sarcasm detection process incorporates additional steps to relabel sarcastic posts as negative, ensuring a more accurate representation of sentiment.

6.2 Methodology

In the collected datasets, the most frequent emojis and the types of posts in which they are found have been observed. As seen in Figure 6.1, among these frequent emojis, some express positive sentiments, but they also appear in negative posts. This curiosity prompted a closer examination of these posts to explore the context of emoji use and identify ways to exploit their existence in sentiment analysis. For example, the 😂 face with tears of joy emoji, generally associated with happiness, appears in negative posts more than in positive ones by a factor of 4.6. This highlights the complexity of emoji usage in digital communication.

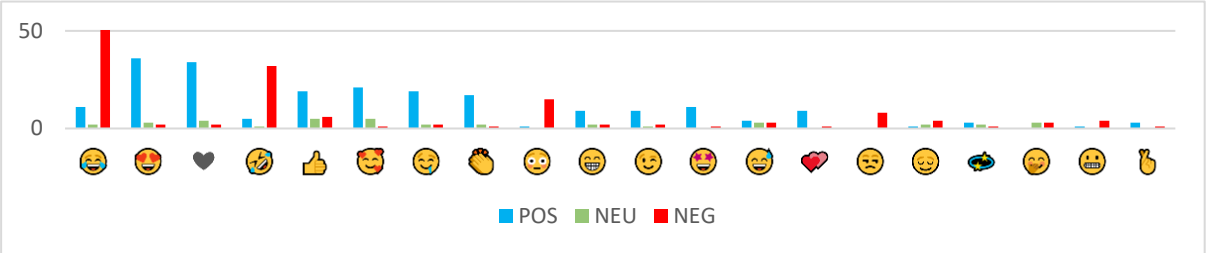

Figure 6.1 Bar chart showing a comparison of the occurrences of some of the most frequent emojis in the collected datasets

This section outlines the proposed Sarcasm Detection Approach (SDA) and the integration of the Emoji Dictionary (ED) to enhance sentiment analysis and sarcasm detection. The

methodology also describes how SDA and ED were evaluated in comparison to two alternative sarcasm detection approaches: Logistic Regression and WELMSD.

6.2.1 Sarcasm Detection Approach (SDA)

The Sarcasm Detection Approach (SDA) operates by analysing conflicts between the sentiment of text and emojis within social media posts. It consists of the following steps:

1. Input Separation: Each post is split into two components:
    - Only Text: The textual content after removing all emojis.
    - Only Emojis: A string comprising only the emojis extracted from the post.
2. Sentiment Classification: The sentiment of both components is determined using either:
    - A lexicon-based approach (e.g., VADER with ED) or
    - A machine learning model (e.g., BERT).
3. Conflict Detection: If the sentiment of the text contradicts the sentiment of the emojis, sarcasm is detected. Specifically:
    - If Text Sentiment = Positive and Emoji Sentiment = Negative
    - Or if Text Sentiment = Negative and Emoji Sentiment = Positive
4. Reclassification: Posts identified as sarcastic are reclassified as Negative to better reflect the underlying sentiment.

6.2.1.1 Sarcasm Detection Approach (SDA) with VADER

Combining the sentiment analysis capabilities of VADER with the Sarcasm Detection Approach (SDA), it is possible to determine the emotional tone of text and emojis together. The main goal of this integration is to find conflicts between the text's sentiment and the emojis' sentiment. The approach identifies sarcasm by dividing each post into text and emoji components and calculating the VADER score for each. If the sentiment in the text is positive but the sentiment in the emojis is negative, or vice versa, the final sentiment classification is changed to negative. False negatives would be decreased because only situations in which the text and emojis strongly contrast in the emotions they convey are classified as sarcastic.

6.2.1.2 Sarcasm Detection Approach (SDA) with BERT

Incorporating the Sarcasm Detection Approach (SDA) with BERT involves a sophisticated analysis of the sentiment conflict between text and emojis within a post. The method identifies sarcasm when a notable sentiment difference is noticed by evaluating the sentiment of text and

emojis independently. If sarcasm is found, the final sentiment is labelled as negative. This study outlines four methods for applying SDA with BERT to identify the most accurate approach. After presenting these four methodologies, the analysis will focus on selecting and applying the single most effective method for integrating SDA with BERT, based on comparative analysis. In the results section, the results of the implementation process of the four methods will be detailed and examined. Here is a detailed explanation for each method, with a design to clarify the difference between them.

In Method 1, as seen in Figure 6.2, first of all, the classifier is trained on the original posts. Then, the trained model was used to predict the sentiment of the posts. After that, to detect the sarcasm, the classifier predicts the sentiment of the only text and of the only emoji of each post; if there is a conflict in the sentiment, the sarcasm is detected, so the final prediction is changed to NEG.

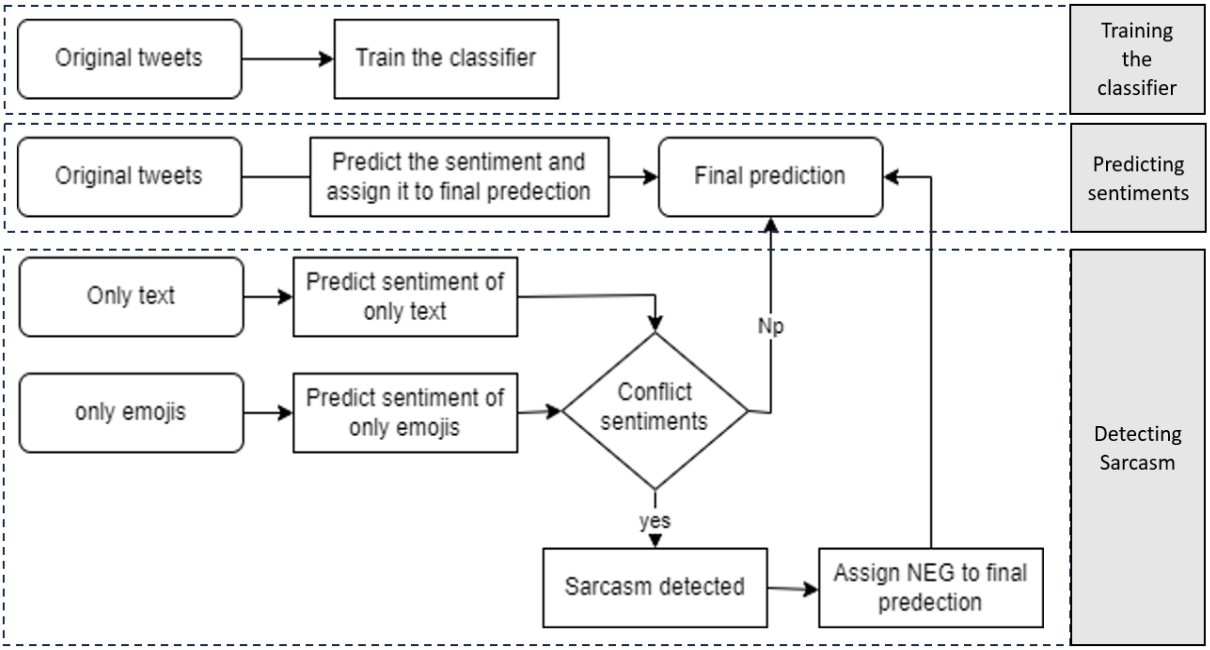

Figure 6.2 Flow chart of detecting sarcasm with BERT - Method1

In Method 2, as seen in Figure 6.3, first of all, the classifier is trained on the original posts. Then, the trained model was used to predict the sentiment of the posts. After that, to detect the sarcasm, the classifier predicts the sentiment of the only text and of the only emoji after replacing all the emojis with their CLDR short names by using the Demojize function in Python. If there is a conflict in the sentiment, the sarcasm is detected, so the final prediction is changed to NEG.

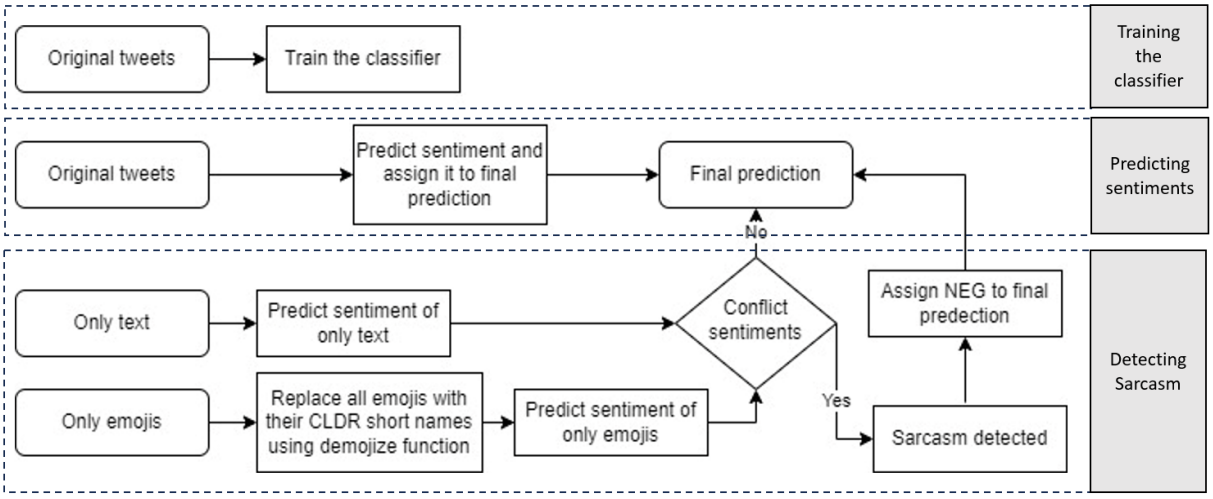

Figure 6.3 Flow chart of detecting sarcasm with BERT - Method2

In Method 3, as seen in Figure 6.4, first of all, the classifier is trained on the original posts after replacing all the emojis with their CLDR short names by using the Demojize function in Python. Then, the trained model was used to predict the sentiment of the posts after using the Demojize function. After that, to detect the sarcasm, the classifier predicts the sentiment of the only text and of the only emoji. After using the Demojize function, if there is a conflict in the sentiment, the sarcasm is detected, so the final prediction is changed to NEG.

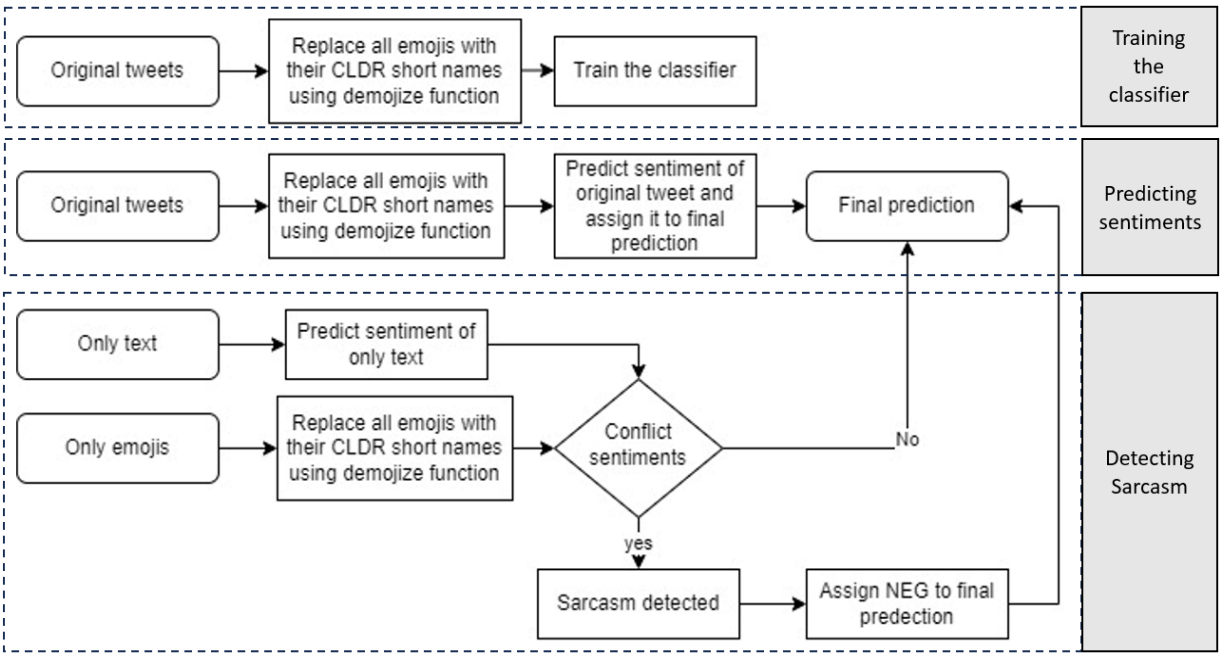

Figure 6.4 Flow chart of detecting sarcasm with BERT - Method3

In Method 4, as seen in Figure 6.5, this method is exactly the same as Method 3, but instead of using the Demojize function to replace emojis, the Emoji Dictionary (ED) that have been created before is used.

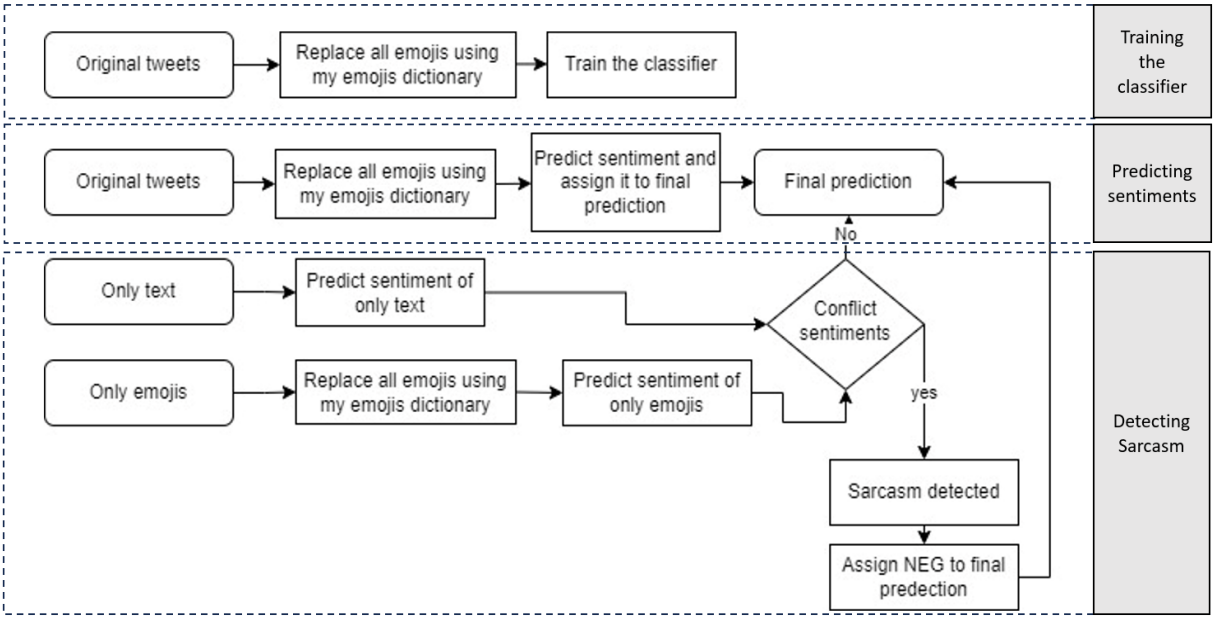

Figure 6.5 Flow chart of detecting sarcasm with BERT - Method4

6.2.2 Emoji Dictionary (ED) Integration

The Emoji Dictionary (ED) was initially developed to enhance sentiment analysis by providing precise sentiment interpretations for emojis. Unlike generic approaches that often treat emojis as neutral or context-independent symbols, the ED works based on rigorous analysis, making it an indispensable tool for sentiment-rich data.

When integrated with the Sarcasm Detection Approach (SDA), the ED further amplifies the effectiveness of sarcasm detection. By improving the accuracy of emoji sentiment interpretation, the ED directly supports SDA's ability to identify conflicts between text and emoji sentiments. This synergy ensures:

- Improved Conflict Detection: Enhanced sentiment scores for emojis allow SDA to more reliably identify instances where text and emoji sentiments clash, a key indicator of sarcasm.
- Accurate Reclassification: Posts with detected sarcasm are more precisely reclassified as negative, reducing errors in sentiment classification.

6.2.3 Comparative Approaches

To evaluate the effectiveness of the proposed Sarcasm Detection Approach (SDA) enhanced with the Emoji Dictionary (ED), two widely recognized sarcasm detection methods were selected for comparison. Below is a detailed description of these comparative approaches:

1. Logistic Regression

This approach, based on the methodology outlined in "Logistic Regression Method for Sarcasm Detection of Text Data" (Bipin Gupta, 2019), uses a logistic regression model trained on labelled datasets with sarcasm (sar) and non-sarcasm (non-sar) annotations. The method incorporates features such as n-grams, part-of-speech tagging, and specific textual patterns to detect sarcasm. The simplicity and interpretability of logistic regression make it a reliable baseline method for sarcasm detection.

2. WELMSD (Word Embedding and Language Model-Based Sarcasm Detection)

As described in "WELMSD – Word Embedding and Language Model-based Sarcasm Detection" (P. Kumar & Sarin, 2022), this method leverages advanced NLP techniques to improve sarcasm detection. It combines FastText embeddings for word-level representation with BERT to capture contextual nuances in text. The architecture integrates embedding layers and language models, enabling robust feature extraction and sarcasm detection. Attention mechanisms are employed to focus on sarcasm-relevant parts of the text, enhancing its ability to identify complex sarcastic patterns. The method is trained on datasets annotated with sarcasm (sar) and non-sarcasm (non-sar) labels.

The performance of SDA+ED is evaluated against these comparative approaches using standard sentiment analysis classifiers like VADER and BERT, with a focus on metrics such as precision, recall, F1-score, and accuracy. This ensures a comprehensive assessment of SDA+ED's ability to detect sarcasm, particularly in emoji-rich content.

6.3 Implementation and Results

6.3.1 SDA with VADER and BERT

All the posts that have one of the emojis that are in Appendix A were collected; the percentages of these posts from the posts that have emojis are 43.5%, 32.96%, and 62.5%, respectively. From each dataset, 1000 posts were collected randomly. These posts were labelled by three annotators. Table 6.1 presents the distribution of each class in the three datasets.

Table 6.1 The classes distributions (NEG, NEU, POS) in the three datasets

	COVID-19 Vaccine	**Electric Cars**	**Vegetarianism**
NEG	64.6%	33.5%	26.8%
NEU	18.6%	32.3%	7.2%
POS	16.8%	34.2%	66%

6.3.1.1 Sentiment Analysis and Detecting sarcasm with VADER

This section evaluates the classification performance of utilizing an Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA) with VADER. Each dataset is divided into three sets of information: original_post, only_text, and only_emojis. The VADER was applied three times to compare the performances:

- Original: apply VADER to the original_post.
- ED: translate all emojis using the Emoji Dictionary (ED) before applying VADER.
- ED & SDA: translate all emojis using the Emoji Dictionary (ED) before applying VADER. Then, apply DSA to only_emojis and only_text columns.

As seen in Table 6.2, Table 6.3, and Table 6.4, in general, the performance of the NEU class is the lowest. When the Emoji Dictionary (ED) is used, performance is enhanced in all metrics. The biggest improvement is in the precision of the NEU class; it was enhanced by 46.36% on average; this reflects the increases in the True NEU posts. The average accuracy has improved by 21.93%. When the Sarcasm Detection Approach (SDA) is added to the Emoji Dictionary (ED), the biggest change is in the NEG class; its precision is decreased by 5.09% on average, and the recall is increased by 14.35% on average. This means that the classifier detects more actual NEG posts at the same time that the False NEG is increased. VADER is converting each emoji to text before analysing the sentiment; from the results, it is evident that this converting process is

not accurate. This issue was addressed by using the Emoji Dictionary (ED) to work on emojis before running VADER. As seen in Figure 6.6, that represents how much each evaluation metric is improved in all datasets when ED & SDA are used in comparison to using VADER on original posts. The most significant increase in the F-1 score is observed in the NEG class, this reflects the success of the Sarcasm Detection Approach (SDA) and Emoji Dictionary (ED) in detecting NEG posts.

Table 6.2 Performance comparison of VADER, VADER with ED, VADER with ED & SDA in COVID-19 Vaccine dataset

		Orig.	**ED**	**ED & SDA**
NEG: 646 posts	**Prec.**	0.76	0.89	0.87
	Rec.	0.53	0.76	0.91
	F1	0.62	0.82	0.89
NEU: 186 posts	**Prec.**	0.13	0.69	0.75
	Rec.	0.08	0.27	0.27
	F1	0.1	0.39	0.4
POS: 168 posts	**Prec.**	0.26	0.4	0.5
	Rec.	0.69	0.9	0.76
	F1	0.38	0.56	0.6
Accuracy		0.47	0.7	0.77

Table 6.3 Performance comparison of VADER, VADER with ED, VADER with ED & SDA in Electric Cars dataset

		Orig.	ED	ED & SDA
NEG: 335 posts	**Prec.**	0.49	0.81	0.76
	Rec.	0.4	0.72	0.86
	F1	0.44	0.76	0.81
NEU: 323 posts	**Prec.**	0.2	0.74	0.78
	Rec.	0.08	0.2	0.2
	F1	0.11	0.31	0.31
POS: 342 posts	**Prec.**	0.43	0.52	0.56
	Rec.	0.76	0.94	0.88
	F1	0.55	0.67	0.68
Accuracy		0.42	0.63	0.65

Table 6.4 Performance comparison of VADER, VADER with ED, VADER with ED & SDA in Vegetarianism dataset

		Orig.	ED	ED & SDA
NEG: 268 posts	**Prec.**	0.58	0.85	0.77
	Rec.	0.43	0.72	0.87
	F1	0.49	0.78	0.83
NEU: 72 posts	**Prec.**	0.02	0.3	0.36
	Rec.	0.04	0.19	0.19
	F1	0.03	0.24	0.25
POS: 660 posts	**Prec.**	0.75	0.85	0.9
	Rec.	0.73	0.93	0.88
	F1	0.74	0.89	0.89
Accuracy		0.6	0.82	0.84

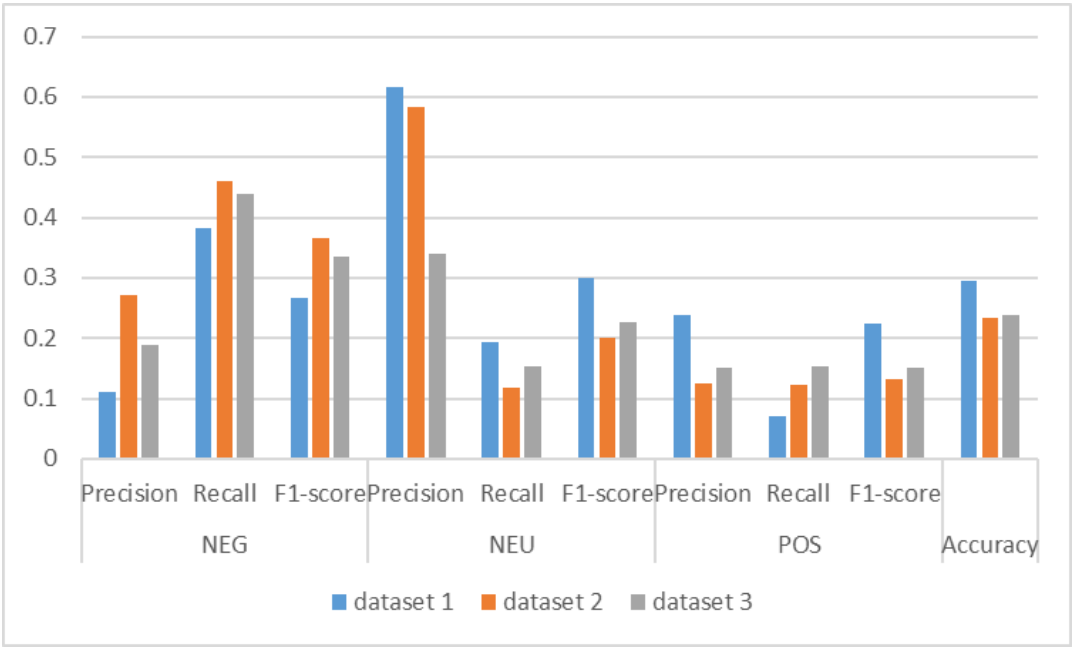

Figure 6.6 Bar chart showcasing the changes in evaluation metrics across all the evaluation metrics for the three datasets, following the incorporation of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA), highlighting the enhanced performance of VADER

6.3.1.2 Sentiment Analysis and Detecting sarcasm with BERT

The Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA) were also examined with BERT using the same three datasets. Each dataset has 1000 posts; they run on three different splits: 50:10:40, 60:10:30, and 70:10:20. The decision to incorporate several datasets and data splits supports the robustness and generalizability of the research findings. It ensures that the conclusions are applicable across a variety of sentiment analysis contexts.

There are four methods, in addition to using the classifier without any changes, to compare all the performances:

- Original: apply BERT to the original_post.
- Method 1: train BERT on original posts. Then, predict the sentiments using the original posts. After that, predict the sentiment of only_text and only_emojis. Next, apply SDA.
- Method 2: train BERT on original posts. Then, predict the sentiments using the original posts. After that, convert the emojis of only_emojis into short text using the Demojize() function in Python. Then predict the sentiment of only_text and only_emojis. Finally, apply the technique of detecting sarcasm.
- Method 3: convert emojis into short text of original_post and only_emojis using the Demojize() function. Next, train BERT on original posts. Then, predict the sentiments using the original posts. After that, predict the sentiment of only_text and only_emojis. Finally, apply detecting sarcasm.
- Method 4: use Emoji Dictionary (ED) to translate emojis in original_post and only_emojis. Next, train BERT on original posts. Then, predict the sentiments using the original posts. After that, predict the sentiment of only_text and only_emojis. Finally, apply detecting sarcasm.

Analysing Table 6.5, Table 6.6, and Table 6.7 provides a comparative overview of BERT's performance across different data splits (50:10:40, 60:10:30, and 70:10:20) for the COVID-19 Vaccine, Electric Cars, and Vegetarianism datasets using various methods, including the original configuration.

In Table 6.5, BERT's application to the COVID-19 Vaccine dataset shows Method 4 consistently outperforms the original and other methods across all data splits, with notable gains in all the metrics of POS and then NEU classes.

Analysing the Electric Cars dataset with different splits in Table 6.6, similar to the COVID-19 Vaccine dataset, Method 4 exhibits superior performance in the Electric Cars dataset, especially in the precision of the POS class and recall of the NEU class.

Table 6.7 assesses BERT's performance on the Vegetarianism dataset; it also shows Method 4's effectiveness, with notable advancements in all classes, especially the recall of the NEG class and the precision of the NEU class. This underscores its robustness across topics.

The increase in training data size generally correlates with better classification performance, especially notable in Method 4, which emphasises the advantages of having a bigger dataset for training models.

Table 6.5 BERT performance comparison of using ED, SDA, and Demojize function in COVID-19 Vaccine dataset with different splits

		50:10:40					60:10:30					70:10:20				
		Orig.	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**
NEG	**Prec.**	0.85	0.77	0.77	0.89	0.92	0.88	0.87	0.78	0.9	0.9	0.85	0.74	0.82	0.95	0.93
	Rec.	0.86	0.93	0.93	0.97	0.95	0.85	0.87	0.94	0.97	0.96	0.89	0.98	0.97	0.95	0.96
	F1	0.85	0.84	0.85	0.93	0.94	0.86	0.87	0.85	0.93	0.93	0.87	0.84	0.89	0.95	0.95
NEU	**Prec.**	0.6	0.59	0.59	0.85	0.8	0.73	0.74	0.71	0.89	0.89	0.8	0.79	0.79	0.85	0.9
	Rec.	0.61	0.59	0.59	0.72	0.76	0.57	0.57	0.52	0.71	0.71	0.63	0.58	0.61	0.89	0.92
	F1	0.6	0.59	0.59	0.78	0.78	0.64	0.65	0.6	0.79	0.79	0.71	0.67	0.69	0.87	0.91
POS	**Prec.**	0.45	0.5	0.62	0.75	0.8	0.45	0.47	0.57	0.76	0.79	0.62	0.5	0.89	0.75	0.86
	Rec.	0.43	0.07	0.12	0.61	0.79	0.62	0.6	0.26	0.7	0.76	0.64	0.03	0.48	0.73	0.83
	F1	0.44	0.13	0.2	0.67	0.8	0.52	0.53	0.36	0.73	0.78	0.63	0.06	0.63	0.74	0.79
Accuracy		0.74	0.73	0.74	0.86	0.89	0.76	0.77	0.75	0.88	0.88	0.8	0.75	0.82	0.9	0.92

Table 6.6 BERT performance comparison of using ED, SDA, and Demojize function in Electric Cars dataset with different splits

		50:10:40					60:10:30					70:10:20				
		Orig.	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**
NEG	**Prec.**	0.79	0.79	0.53	0.83	0.9	0.8	0.39	0.61	0.93	0.84	0.79	0.75	0.6	0.79	0.82
	Rec.	0.73	0.73	0.81	0.96	0.9	0.78	0.86	0.8	0.93	0.94	0.73	0.73	0.85	0.92	0.92
	F1	0.76	0.76	0.64	0.89	0.9	0.79	0.53	0.69	0.88	0.89	0.76	0.74	0.7	0.85	0.87
NEU	**Prec.**	0.7	0.7	0.7	0.84	0.78	0.59	0.64	0.58	0.9	0.86	0.72	0.73	0.73	0.85	0.85
	Rec.	0.59	0.59	0.55	0.8	0.84	0.53	0.48	0.51	0.79	0.79	0.71	0.69	0.69	0.86	0.85
	F1	0.64	0.64	0.61	0.82	0.81	0.55	0.55	0.54	0.84	0.82	0.71	0.71	0.71	0.86	0.85
POS	**Prec.**	0.56	0.56	0.54	0.82	0.83	0.6	0.2	0.61	0.79	0.84	0.64	0.64	0.82	0.84	0.93
	Rec.	0.69	0.69	0.37	0.74	0.77	0.67	0.01	0.5	0.79	0.8	0.7	0.68	0.54	0.7	0.83
	F1	0.62	0.62	0.44	0.78	0.8	0.63	0.02	0.55	0.79	0.82	0.67	0.66	0.65	0.76	0.88
Accuracy		0.67	0.67	0.58	0.83	0.84	0.66	0.45	0.6	0.84	0.84	0.71	0.7	0.69	0.83	0.87

Table 6.7 BERT performance comparison of using ED, SDA, and Demojize function in Vegetarianism dataset with different splits

		50:10:40					60:10:30					70:10:20				
		Orig.	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**	**Orig.**	**Mtd. 1**	**Mtd. 2**	**Mtd. 3**	**Mtd. 4**
NEG	**Prec.**	0.81	0.78	0.59	0.84	0.86	0.72	0.71	0.57	0.76	0.83	0.72	0.73	0.64	0.75	0.89
	Rec.	0.68	0.71	0.83	0.93	0.93	0.71	0.71	0.85	0.98	0.95	0.68	0.7	0.83	0.96	0.89
	F1	0.74	0.75	0.69	0.88	0.89	0.82	0.71	0.68	0.85	0.88	0.7	0.71	0.72	0.84	0.89
NEU	**Prec.**	0.26	0.27	0.28	0.52	0.52	0.33	0.33	0.33	0.38	0.6	0.38	0.38	0.38	0.5	0.5
	Rec.	0.31	0.28	0.31	0.45	0.52	0.36	0.36	0.36	0.23	0.41	0.2	0.2	0.2	0.4	0.4
	F1	0.29	0.27	0.3	0.48	0.52	0.35	0.35	0.35	0.29	0.49	0.26	0.26	0.26	0.44	0.44
POS	**Prec.**	0.86	0.86	0.91	0.95	0.94	0.84	0.84	0.88	0.94	0.93	0.82	0.83	0.86	0.97	0.92
	Rec.	0.9	0.89	0.75	0.92	0.91	0.83	0.83	0.7	0.87	0.9	0.89	0.89	0.8	0.88	0.94
	F1	0.88	0.88	0.83	0.93	0.93	0.84	0.83	0.78	0.91	0.92	0.85	0.86	0.83	0.92	0.93
Accuracy		0.8	0.8	0.74	0.89	0.89	0.77	0.76	0.71	0.85	0.88	0.78	0.79	0.77	0.87	0.89

Method 1 generally shows a slight gains over the original configuration but struggles to match the refinements seen with later methods. This suggests that while the initial modifications made in Method 1 provide some benefit, they are not as impactful in refining sentiment classification.

Method 2 exhibits varied performance, with slight advancements in some cases but not consistently across all classes and datasets. It indicates that the changes made in Method 2 offer limited enhancements to BERT's classification accuracy.

Method 3 shows a marked leap in performance, especially in the NEG and NEU classes. This suggests a more nuanced approach to emoji interpretation and sarcasm detection, leading to better overall classification accuracy.

Method 4 outperforms all other methods. By effectively incorporating ED, SDA, and the Demojize function, Method 4 enhances BERT's ability to interpret complex sentiment signals, making it the most effective strategy for improving sentiment classification accuracy.

Methods 1 and 2 in all datasets have the lowest performance. In these two methods, the emojis were replaced after training the classifier; this indicates that there is a need to treat the emojis before starting training, which was exactly the approach taken in Methods 3 and 4. However, in Method 3, the Demojize function is used, and in Method 4, Emoji Dictionary (ED) is used. The performances of Methods 3 and 4 are close to each other; both considerably enhance BERT's sentiment analysis capabilities. In spite of that, Method 4 stands out as the most effective, demonstrating the highest precision, recall, F1 scores, and overall accuracy.

As seen in Figure 6.7, Figure 6.8, and Figure 6.9, that represents the extent of change across various evaluation metrics in all datasets with different splits when Method 4 is applied compared to using BERT on original posts. The degree of change differs from one dataset to another. It has been noted that using this recommended method can assist in resolving the imbalanced dataset issue. There is an inverse relationship between the size of the class and the amount of performance enhancement; in other words, for every large class, there is little progress, and vice versa. For example, Dataset 3 shows the greatest gain in the NEU class, in spite of the fact that this dataset's NEU class percentage is only 7.25%, as Figure 6.10 illustrates. Another example is the POS class in Dataset 1; the percentage of the POS class in this dataset is only 16.75%. The smallest class in each dataset achieved the least gains: the

NEG class in dataset 1 (64.6%), the NEG class in dataset 2 (33.5%), and the POS class in dataset 3 (66%).

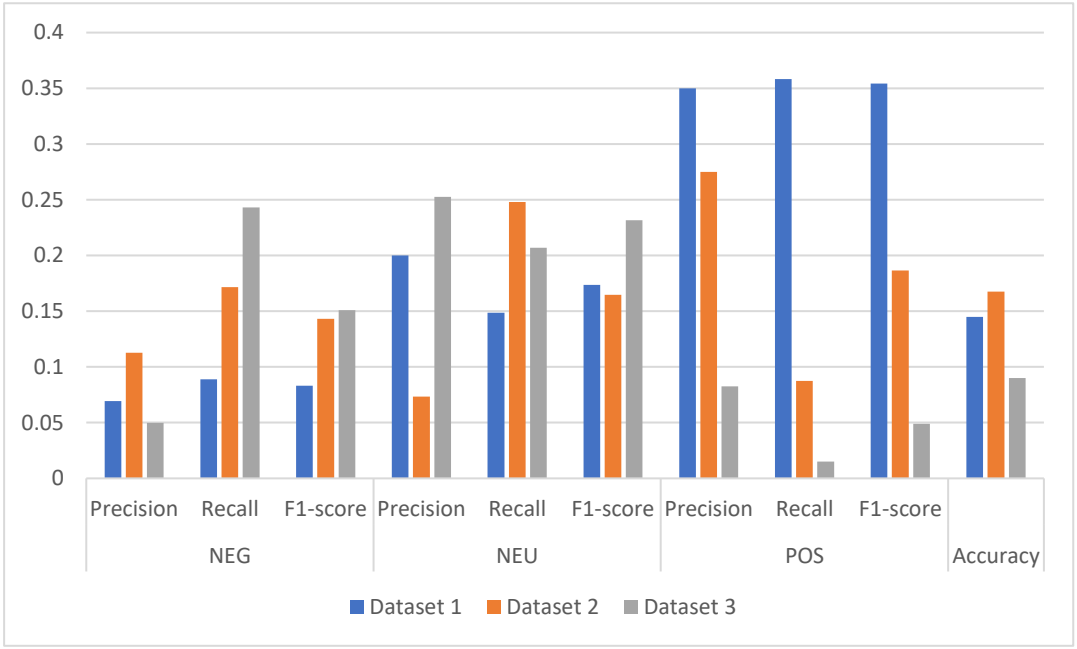

Figure 6.7 Bar chart illustrating the changes in evaluation metrics for Method 4 in compared with the original BERT with 50% training, 10% validating, and 40% testing

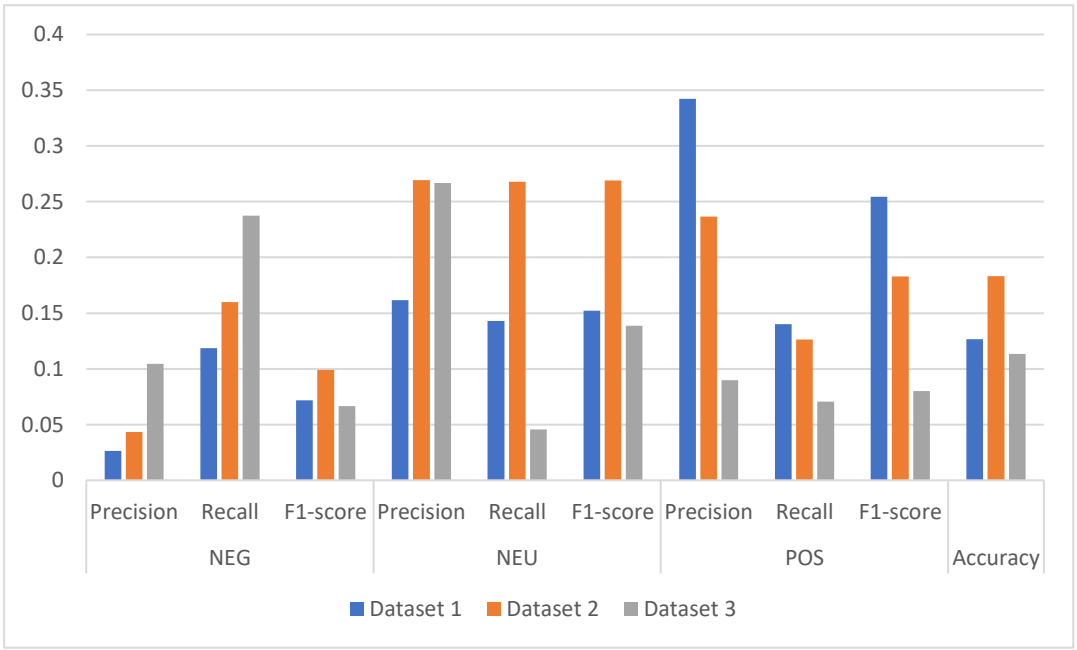

Figure 6.8 Bar chart illustrating the changes in evaluation metrics for Method 4 in compared with the original BERT with 60% training, 10% validating, and 30% testing

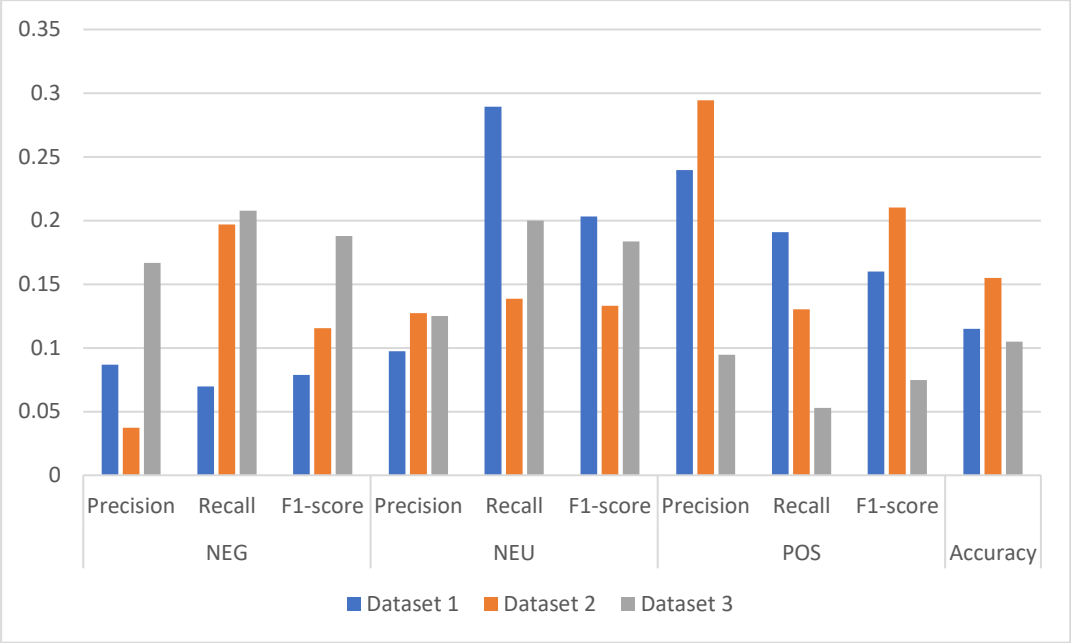

Figure 6.9 Bar chart illustrating the changes in evaluation metrics for Method 4 in compared with the original BERT with 70% training, 10% validating, and 20% testing

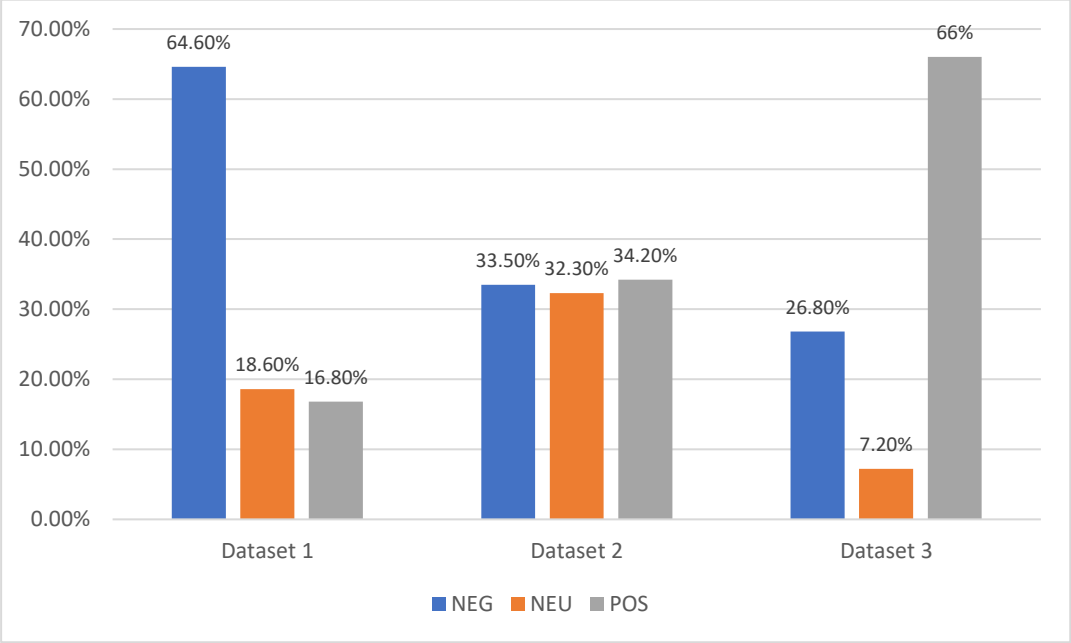

Figure 6.10 Bar chart represents the distribution of the posts sentiments (NEG, NEU, POS) of the three datasets that used for evaluating the impact of emoji in sentiment analysis

6.3.1.3 Discussion

Our results show that the performance of the VADER sentiment classification with an Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA) is higher than analysing the original text, as seen in Figure 6.11, which shows the average changes after using the proposed method in all datasets. The performance of all classes has improved; the average F-1 score gains for the

NEG, NEU, and POS classes are 32.31%, 24.25%, and 16.88%, respectively, and the accuracy has been enhanced by about 25.57% on average.

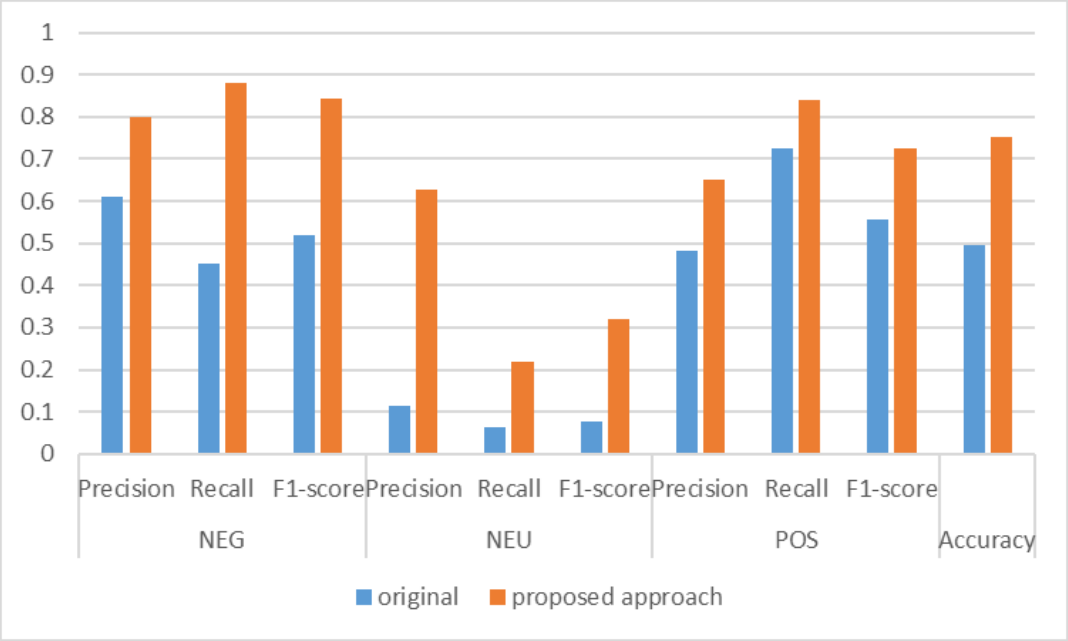

Figure 6.11 The average change in evaluation metrics resulting from the proposed method compared to the original VADER.

Our results show that the performance of the BERT sentiment classification with Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA) is higher than analysing the original text, as seen in Figure 6.12, which shows the average changes after using the proposed method in all datasets. The performance of all classes has improved; the average F-1 score gains of the NEG, NEU, and POS classes are 11.07%, 18.33%, and 17.25%, respectively, and the accuracy has been enhanced by about 13.34% on average.

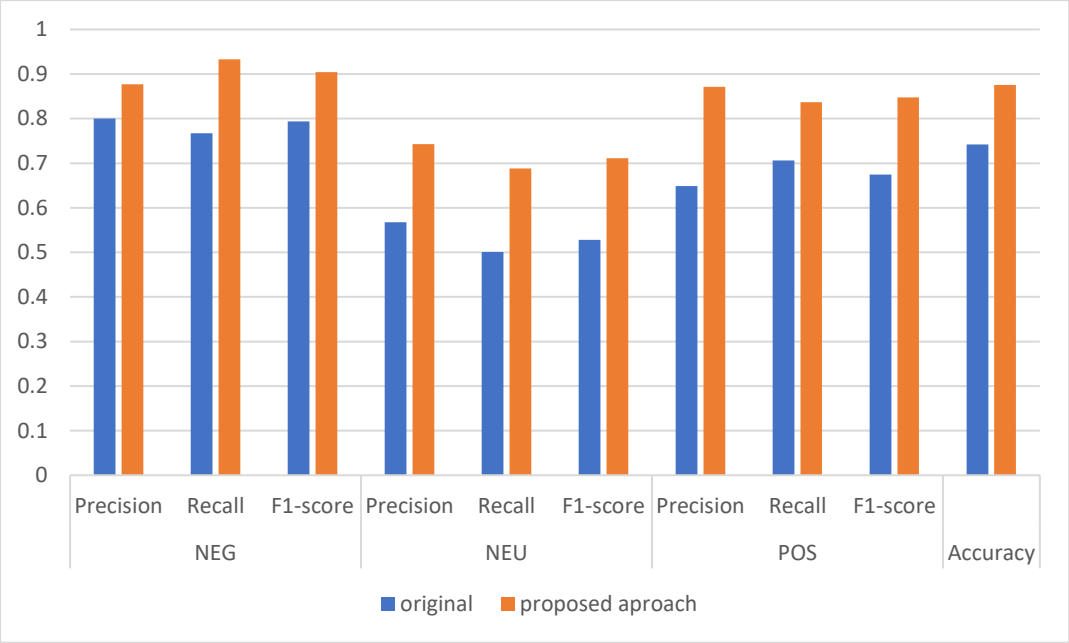

Figure 6.12 The average change in evaluation metrics resulting from the proposed method compared to the original BERT

As an overview, the suggested methodology, which includes including emojis, modifying them by using Emoji Dictionary (ED), and adding Sarcasm Detection Approach (SDA), enhanced the accuracy in all three datasets with VADER and BERT. The increase in VADER is more pronounced because, without any modifications, the performance of BERT in general is higher when compared with VADER.

6.3.2 Comparative Analysis

This section evaluates the performance of various sarcasm detection approaches across seven datasets described in Section 5.3. The comparative study focuses on two aspects: sentiment classification using VADER and sentiment classification using BERT. The approaches include:

3. Original Baseline: Applying VADER or BERT directly to the datasets without any modifications.
4. Logistic Regression: Predicting sarcasm using a logistic regression model and modifying sentiment classifications based on sarcastic predictions.
5. WELMSD: Using WELMSD for sarcasm prediction and modifying sentiment classifications accordingly.
6. ED + SDA: Incorporating the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) to refine sarcasm detection and sentiment classification.

The logistic regression and WELMSD models were trained using a combined sarcasm detection dataset comprising 53,325 posts labelled as sarcastic (sar) or non-sarcastic (non-sar). The sources of this dataset include:

- Twitter Sarcastic Classification Dataset: 3,468 tweets (Aslam, 2023)
- Irony and Sarcasm Detection Dataset: 3,817 tweets (Murphy, 2020)
- Tweets with Sarcasm and Irony Dataset: 43,240 tweets (John, 2020)
- iSarcasmEval Dataset: 1,400 posts (Farha, 2022)
- SarcasmDetectionUsingLogisticRegression: 3,468 posts (Ramos, 2021)

After training, the models were used to predict whether each post in the seven datasets was sarcastic or non-sarcastic, with the results stored in two new columns:

- Predicted_logRegression: Indicates whether the post was predicted as sarcastic (1) or not sarcastic (0) by the logistic regression model.
- Predicted_welmsd: Indicates whether the post was predicted as sarcastic (1) or not sarcastic (0) by the WELMSD model.

6.3.2.1 Implementation of Comparative Approaches

Comparison Using VADER

The sarcasm detection methods were integrated into the sentiment analysis workflow using VADER as follows:

1. Original Baseline: VADER was applied directly to the original posts to predict sentiment without any modifications.
2. ED + SDA: Emojis were processed using the Emoji Dictionary (ED), followed by sentiment classification with VADER. Sarcasm was detected using the SDA by identifying sentiment conflicts between text and emojis. If sarcasm was detected, the final sentiment was reclassified as negative.
3. Logistic Regression: VADER was applied to predict sentiment, and the logistic regression model was used to identify sarcastic posts. If a post was predicted as sarcastic (predicted_logRegression = 1), its final sentiment was changed to negative.
4. WELMSD: Similar to logistic regression, VADER was applied first, and WELMSD was used to identify sarcastic posts. If a post was predicted as sarcastic (predicted_welmsd = 1), its final sentiment was changed to negative.

Comparison Using BERT

The sarcasm detection methods were also integrated into the sentiment analysis workflow using BERT as follows:

1. Original Baseline: BERT was applied directly to the original posts to predict sentiment without any modifications.
2. Logistic Regression: BERT was applied to predict sentiment, and the logistic regression model was used to identify sarcastic posts. If a post was predicted as sarcastic (predicted_logRegression = 1), its final sentiment was changed to negative.
3. WELMSD: Similar to logistic regression, BERT was applied first, and WELMSD was used to identify sarcastic posts. If a post was predicted as sarcastic (predicted_welmsd = 1), its final sentiment was changed to negative.
4. SDA: BERT was applied to predict sentiment, and sarcasm was detected using the SDA. If sarcasm was detected, the final sentiment was reclassified as negative.
5. ED + SDA: Similar to SDA, but enhanced by incorporating the Emoji Dictionary (ED) to refine the sentiment classification of emojis before applying the Sarcasm Detection Approach (SDA).

6.3.2.2 Results of Comparative Analysis

Performance Using VADER

The comparison using VADER considers the following approaches:

1. Original: Standard VADER without any modifications.
2. VADER_Logistic: VADER combined with logistic regression for sarcasm detection. Sarcastic posts detected by logistic regression are reclassified as negative.
3. VADER_WELMSD: VADER combined with WELMSD for sarcasm detection. Sarcastic posts identified by WELMSD are reclassified as negative.
4. VADER_SDA: VADER enhanced with the Sarcasm Detection Approach (SDA), where sarcasm is detected based on sentiment conflicts between text and emojis.
5. VADER_ED_SDA: VADER combined with the Emoji Dictionary (ED) and SDA for improved sarcasm detection.

Table 6.8, Table 6.9, and Table 6.10 focus on general datasets, including Dataset 1 (COVID-19 Vaccine - General), Dataset 2 (Electric Cars - General), and Dataset 3 (Vegetarianism - General). Across these datasets, VADER_ED_SDA consistently delivers higher precision and F1-scores for the NEG and POS classes, indicating its robustness in capturing sarcasm-induced negativity and positive sentiments. VADER_WELMSD often achieves higher recall for the NEG class,

reflecting its capability to identify sarcastic content, though it sacrifices precision, leading to lower overall F1-scores. Neutral sentiment (NEU class) remains challenging across all approaches, with consistently low recall and F1-scores, highlighting the difficulty of detecting neutrality in text-heavy datasets.

Table 6.8 Performance comparison of sarcasm detection approaches using VADER on Dataset 1 (COVID-19 Vaccine - general)

Dictionary		**Original**	**VADER_Logistic**	**VADER_WELMSD**	**VADER_SDA**	**VADER_ED_SDA**
NEG: 1073 posts	**Prec.**	0.54	0.54	0.53	0.55	0.56
	Rec.	0.56	0.57	0.63	0.57	0.59
	F1	0.55	0.56	0.58	0.56	0.57
NEU: 1652 posts	**Prec.**	0.79	0.80	0.81	0.80	0.80
	Rec.	0.33	0.33	0.32	0.33	0.33
	F1	0.47	0.47	0.46	0.47	0.47
POS: 275 posts	**Prec.**	0.18	0.18	0.17	0.18	0.19
	Rec.	0.80	0.77	0.65	0.79	0.80
	F1	0.29	0.29	0.27	0.30	0.30
Accuracy		0.46	0.46	0.46	0.46	0.47

Table 6.9 Performance comparison of sarcasm detection approaches using VADER on Dataset 2 (Electric Cars - general)

Dictionary		Original	VADER_Logistic	VADER_WELMSD	VADER_SDA	VADER_ED_SDA
NEG: 598 posts	Prec.	0.51	0.51	0.41	0.52	0.52
	Rec.	0.47	0.53	0.68	0.48	0.49
	F1	0.49	0.52	0.51	0.50	0.51
NEU: 2094 posts	Prec.	0.85	0.86	0.88	0.85	0.85
	Rec.	0.43	0.43	0.37	0.43	0.43
	F1	0.58	0.57	0.52	0.58	0.58
POS: 308 posts	Prec.	0.19	0.18	0.17	0.18	0.19
	Rec.	0.84	0.78	0.61	0.82	0.83
	F1	0.31	0.29	0.26	0.30	0.31
Accuracy		0.48	0.49	0.46	0.48	0.49

Table 6.10 Performance comparison of sarcasm detection approaches using VADER on Dataset 3 (Vegetarianism - general)

Dictionary		Original	VADER_Logistic	VADER_WELMSD	VADER_SDA	VADER_ED_SDA
NEG: 761 posts	Prec.	0.66	0.62	0.54	0.64	0.66
	Rec.	0.47	0.49	0.57	0.49	0.51
	F1	0.55	0.55	0.56	0.56	0.57
NEU: 1473 posts	Prec.	0.76	0.77	0.78	0.77	0.77
	Rec.	0.41	0.41	0.38	0.41	0.41
	F1	0.53	0.53	0.51	0.53	0.53
POS: 766 posts	Prec.	0.40	0.40	0.40	0.40	0.40
	Rec.	0.87	0.83	0.78	0.85	0.85
	F1	0.55	0.54	0.53	0.54	0.55
Accuracy		0.54	0.54	0.53	0.54	0.55

Table 6.11, Table 6.12, and Table 6.13 focus on emoji-rich datasets, including Dataset 4 (COVID-19 Vaccine - Emoji Only), Dataset 5 (Electric Cars - Emoji Only), and Dataset 6 (Vegetarianism - Emoji Only). The integration of the Emoji Dictionary (ED) in VADER_ED_SDA shows substantial gains in these datasets. VADER_ED_SDA consistently achieves higher precision, recall, and F1-scores in the NEG and POS classes, demonstrating its ability to handle the interplay between emojis and text effectively. Other approaches, such as VADER_WELMSD and VADER_SDA, also show improved performance compared to Original VADER, but their effectiveness is limited compared to the balanced performance of VADER_ED_SDA. The NEU class continues to perform poorly, with low precision and recall across all methods, reflecting the inherent challenge of neutral sentiment classification in emoji-rich contexts.

Table 6.11 Performance comparison of sarcasm detection approaches using VADER on Dataset 4 (COVID-19 Vaccine - emoji only)

Dictionary		**Original**	**VADER_Logistic**	**VADER_WELMSD**	**VADER_SDA**	**VADER_ED_SDA**
NEG: 646 posts	**Prec.**	0.76	0.76	0.74	0.72	0.82
	Rec.	0.53	0.54	0.61	0.57	0.80
	F1	0.62	0.63	0.66	0.64	0.81
NEU: 186 posts	**Prec.**	0.13	0.13	0.15	0.13	0.13
	Rec.	0.08	0.08	0.08	0.07	0.08
	F1	0.10	0.10	0.10	0.09	0.10
POS: 168 posts	**Prec.**	0.26	0.26	0.25	0.28	0.42
	Rec.	0.69	0.68	0.55	0.64	0.65
	F1	0.38	0.38	0.35	0.39	0.51
Accuracy		0.47	0.48	0.50	0.49	0.64

Table 6.12 Performance comparison of sarcasm detection approaches using VADER on Dataset 5 (Electric Cars - emoji only)

Dictionary		Original	VADER_Logistic	VADER_WELMSD	VADER_SDA	VADER_ED_SDA
NEG: 335 posts	**Prec.**	0.49	0.50	0.47	0.41	0.60
	Rec.	0.40	0.46	0.67	0.45	0.70
	F1	0.44	0.48	0.55	0.43	0.65
NEU: 323 posts	**Prec.**	0.20	0.20	0.23	0.17	0.20
	Rec.	0.08	0.08	0.06	0.06	0.08
	F1	0.11	0.11	0.10	0.09	0.11
POS: 342 posts	**Prec.**	0.43	0.44	0.45	0.47	0.52
	Rec.	0.76	0.72	0.57	0.70	0.73
	F1	0.55	0.54	0.50	0.56	0.61
Accuracy		0.42	0.43	0.44	0.41	0.51

Table 6.13 Performance comparison of sarcasm detection approaches using VADER on Dataset 6 (Vegetarianism - emoji only)

Dictionary		Original	VADER_Logistic	VADER_WELMSD	VADER_SDA	VADER_ED_SDA
NEG: 268 posts	**Prec.**	0.58	0.57	0.50	0.46	0.67
	Rec.	0.43	0.45	0.55	0.49	0.75
	F1	0.49	0.50	0.53	0.47	0.71
NEU: 72 posts	**Prec.**	0.02	0.02	0.01	0.01	0.02
	Rec.	0.04	0.04	0.03	0.03	0.04
	F1	0.03	0.03	0.02	0.02	0.03
POS: 660 posts	**Prec.**	0.75	0.75	0.77	0.78	0.86
	Rec.	0.73	0.72	0.66	0.67	0.70
	F1	0.74	0.73	0.71	0.72	0.78
Accuracy		0.60	0.60	0.58	0.57	0.67

Table 6.14 presents results on the Combined Dataset (Dataset 7 - General and Emoji Only). This dataset aggregates both general and emoji-rich content, offering a comprehensive evaluation of each approach. VADER_ED_SDA maintains its competitive edge, achieving the highest F1-scores across all sentiment classes. Its ability to generalize across diverse contexts highlights the robustness of combining ED and SDA. While VADER_WELMSD performs well in recall for the NEG class, it lacks the balance of precision and recall exhibited by VADER_ED_SDA. Similar to the other datasets, the NEU class remains the weakest across all approaches, with minimal gains in performance.

Table 6.14 Performance comparison of sarcasm detection approaches Using VADER on Dataset 7 (Combined Dataset - general and emoji)

Dictionary		Original	VADER_Logistic	VADER_WELMSD	VADER_SDA	VADER_ED_SDA
NEG: 3518 posts	**Prec.**	0.56	0.56	0.49	0.54	0.61
	Rec.	0.50	0.53	0.64	0.52	0.62
	F1	0.53	0.54	0.56	0.53	0.61
NEU: 6421 posts	**Prec.**	0.75	0.75	0.77	0.75	0.75
	Rec.	0.38	0.37	0.33	0.37	0.38
	F1	0.50	0.50	0.46	0.50	0.50
POS: 2061 posts	**Prec.**	0.28	0.28	0.27	0.28	0.30
	Rec.	0.77	0.74	0.62	0.73	0.76
	F1	0.41	0.40	0.38	0.41	0.43
Accuracy		0.48	0.48	0.47	0.48	0.51

The results across all datasets underscore the importance of integrating advanced sentiment analysis tools with sarcasm and emoji-specific methodologies. VADER_ED_SDA consistently outperforms other approaches in the NEG and POS classes, particularly in emoji-rich contexts, by leveraging the interplay between emojis and text. VADER_WELMSD demonstrates strength in detecting sarcasm-induced negativity but often sacrifices precision, making it less balanced overall. The NEU class remains a challenge for all methods, indicating the need for further refinement in neutral sentiment classification. These findings highlight the potential of combining multiple techniques, such as ED and SDA, to address the complexities of sentiment analysis in diverse contexts.

Performance Using BERT

The comparison using BERT includes the following approaches:

1. Original: Standard BERT without any modifications.
2. BERT_Logistic: BERT combined with logistic regression for sarcasm detection. Sarcastic posts detected by logistic regression are reclassified as negative.
3. BERT_WELMSD: BERT combined with WELMSD for sarcasm detection. Sarcastic posts identified by WELMSD are reclassified as negative.
4. BERT_SDA: BERT enhanced with the Sarcasm Detection Approach (SDA).

5. BERT_ED_SDA: BERT combined with the Emoji Dictionary (ED) and SDA for improved sarcasm detection.

The results from general datasets are detailed in Table 6.15, Table 6.16, and Table 6.17, corresponding to Dataset 1 (COVID-19 Vaccine - General), Dataset 2 (Electric Cars - General), and Dataset 3 (Vegetarianism - General). Key findings include:

BERT_ED_SDA frequently demonstrates balanced performance in the NEG and POS classes, achieving competitive F1-scores and showcasing its robustness in handling sarcastic and emoji-related content. However, specific methods, such as BERT_WELMSD, may outperform it in recall for the NEG class in certain datasets, highlighting their ability to detect sarcasm-induced negativity. BERT_WELMSD excels in recall for the NEG class, indicating its capability to identify sarcastic patterns. However, its lower precision limits its overall F1-scores compared to BERT_ED_SDA. Neutral sentiment classification (NEU class) remains challenging for all approaches, with metrics consistently lower than for the NEG and POS classes. This highlights the inherent difficulty of identifying neutrality, especially in sarcastic content.

Table 6.15 Performance comparison of sarcasm detection approaches using BERT on Dataset 1 (COVID-19 Vaccine - general)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.83	0.82	0.77	0.81	0.83	0.83	0.81	0.77	0.81	0.82	0.82	0.81	0.77	0.81	0.72
	Rec.	0.81	0.81	0.82	0.81	0.81	0.76	0.76	0.78	0.78	0.8	0.79	0.79	0.81	0.79	0.84
	F1	0.82	0.81	0.79	0.81	0.82	0.79	0.78	0.77	0.79	0.81	0.81	0.8	0.79	0.8	0.78
NEU	**Prec.**	0.85	0.86	0.86	0.85	0.87	0.84	0.84	0.84	0.84	0.86	0.84	0.84	0.84	0.84	0.86
	Rec.	0.88	0.88	0.86	0.88	0.86	0.88	0.88	0.85	0.88	0.89	0.88	0.87	0.85	0.88	0.89
	F1	0.87	0.87	0.86	0.87	0.86	0.86	0.86	0.85	0.86	0.87	0.86	0.85	0.85	0.86	0.88
POS	**Prec.**	0.66	0.65	0.65	0.66	0.56	0.59	0.57	0.58	0.59	0.62	0.59	0.59	0.6	0.58	0.69
	Rec.	0.6	0.58	0.47	0.53	0.64	0.6	0.55	0.51	0.5	0.55	0.53	0.5	0.45	0.47	0.08
	F1	0.63	0.62	0.55	0.59	0.59	0.59	0.56	0.54	0.54	0.58	0.56	0.54	0.52	0.52	0.15
Accuracy		0.83	0.82	0.81	0.82	0.82	0.81	0.8	0.8	0.81	0.82	0.81	0.81	0.8	0.81	0.8

Table 6.16 Performance comparison of sarcasm detection approaches using BERT on Dataset 2 (Electric Cars - general)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.75	0.71	0.53	0.71	0.78	0.79	0.72	0.51	0.73	0.75	0.78	0.72	0.51	0.73	0.7
	Rec.	0.7	0.73	0.81	0.71	0.74	0.78	0.79	0.84	0.78	0.74	0.72	0.74	0.83	0.73	0.73
	F1	0.72	0.72	0.64	0.71	0.76	0.78	0.75	0.64	0.75	0.75	0.75	0.73	0.63	0.73	0.72
NEU	**Prec.**	0.9	0.9	0.91	0.9	0.92	0.91	0.91	0.91	0.91	0.9	0.89	0.89	0.91	0.89	0.91
	Rec.	0.93	0.93	0.83	0.93	0.91	0.91	0.9	0.8	0.91	0.92	0.91	0.9	0.79	0.91	0.88
	F1	0.91	0.91	0.87	0.91	0.92	0.91	0.91	0.85	0.91	0.91	0.9	0.89	0.85	0.9	0.89
POS	**Prec.**	0.66	0.68	0.74	0.65	0.59	0.62	0.61	0.65	0.58	0.62	0.56	0.56	0.56	0.53	0.52
	Rec.	0.6	0.52	0.4	0.5	0.71	0.61	0.53	0.42	0.49	0.54	0.54	0.5	0.39	0.45	0.59
	F1	0.63	0.59	0.52	0.56	0.65	0.61	0.57	0.51	0.53	0.58	0.55	0.53	0.46	0.48	0.55
Accuracy		0.85	0.84	0.78	0.84	0.86	0.85	0.84	0.77	0.84	0.84	0.83	0.82	0.76	0.83	0.82

Table 6.17 Performance comparison of sarcasm detection approaches using BERT on Dataset 3 (Vegetarianism - general)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.74	0.71	0.61	0.75	0.72	0.69	0.66	0.57	0.69	0.7	0.72	0.69	0.61	0.72	0.68
	Rec.	0.72	0.74	0.76	0.74	0.76	0.66	0.68	0.72	0.67	0.72	0.74	0.75	0.78	0.74	0.79
	F1	0.73	0.73	0.68	0.74	0.74	0.68	0.67	0.64	0.68	0.71	0.73	0.72	0.68	0.73	0.73
NEU	**Prec.**	0.77	0.79	0.78	0.77	0.8	0.77	0.78	0.77	0.77	0.79	0.81	0.81	0.81	0.81	0.8
	Rec.	0.79	0.79	0.72	0.79	0.8	0.76	0.75	0.68	0.76	0.79	0.79	0.78	0.72	0.79	0.76
	F1	0.78	0.79	0.75	0.78	0.8	0.76	0.76	0.72	0.76	0.79	0.8	0.8	0.77	0.8	0.78
POS	**Prec.**	0.66	0.66	0.69	0.67	0.72	0.64	0.64	0.64	0.65	0.69	0.69	0.68	0.69	0.69	0.69
	Rec.	0.65	0.63	0.61	0.65	0.69	0.68	0.66	0.62	0.68	0.67	0.71	0.67	0.64	0.71	0.65
	F1	0.66	0.64	0.65	0.66	0.7	0.66	0.65	0.63	0.66	0.68	0.7	0.68	0.66	0.7	0.67
Accuracy		0.74	0.73	0.7	0.74	0.76	0.71	0.71	0.67	0.72	0.74	0.75	0.75	0.72	0.75	0.74

The results from emoji-rich datasets are presented in Table 6.18, Table 6.19, and Table 6.20, corresponding to Dataset 4 (COVID-19 Vaccine - Emoji Only), Dataset 5 (Electric Cars - Emoji Only), and Dataset 6 (Vegetarianism - Emoji Only). Observations include BERT_ED_SDA consistently performs well across NEG and POS classes, leveraging emoji semantics to improve precision and F1-scores. It often achieves a more balanced performance compared to other methods. BERT_SDA performs adequately but generally falls short of BERT_ED_SDA, emphasizing the added value of incorporating the Emoji Dictionary (ED). The NEU class remains a challenge in emoji-rich datasets, with metrics reflecting the complexity of detecting neutrality in contexts heavily influenced by emojis.

Table 6.18 Performance comparison of sarcasm detection approaches using BERT on Dataset 4 (COVID-19 Vaccine - emoji only)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.86	0.86	0.83	0.76	0.93	0.86	0.86	0.82	0.75	0.91	0.82	0.82	0.8	0.74	0.82
	Rec.	0.84	0.84	0.85	0.95	0.93	0.87	0.87	0.89	0.93	0.97	0.92	0.92	0.93	0.95	0.97
	F1	0.85	0.85	0.84	0.84	0.93	0.86	0.86	0.85	0.83	0.94	0.87	0.87	0.86	0.83	0.89
NEU	**Prec.**	0.63	0.63	0.62	0.64	0.77	0.56	0.56	0.58	0.59	0.8	0.64	0.64	0.65	0.65	0.82
	Rec.	0.68	0.68	0.66	0.66	0.79	0.59	0.59	0.59	0.59	0.7	0.57	0.57	0.57	0.57	0.66
	F1	0.66	0.66	0.64	0.65	0.78	0.57	0.57	0.58	0.59	0.74	0.6	0.6	0.6	0.6	0.73
POS	**Prec.**	0.45	0.45	0.43	0	0.75	0.57	0.56	0.5	0.2	0.75	0.66	0.65	0.62	0.5	0.74
	Rec.	0.45	0.45	0.36	0	0.73	0.52	0.5	0.34	0.02	0.66	0.43	0.42	0.31	0.01	0.39
	F1	0.45	0.45	0.39	0	0.74	0.54	0.53	0.4	0.04	0.7	0.52	0.51	0.42	0.03	0.51
Accuracy		0.74	0.74	0.73	0.73	0.87	0.76	0.75	0.74	0.71	0.87	0.77	0.77	0.76	0.72	0.81

Table 6.19 Performance comparison of sarcasm detection approaches using BERT on Dataset 5 (Electric Cars - emoji only)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.8	0.77	0.62	0.8	0.82	0.79	0.76	0.57	0.77	0.75	0.77	0.74	0.58	0.41	0.89
	Rec.	0.74	0.76	0.83	0.74	0.91	0.71	0.74	0.79	0.74	0.94	0.76	0.78	0.81	0.97	0.93
	F1	0.77	0.76	0.71	0.77	0.86	0.75	0.75	0.66	0.76	0.84	0.76	0.76	0.67	0.58	0.91
NEU	**Prec.**	0.68	0.67	0.69	0.68	0.85	0.64	0.64	0.67	0.64	0.86	0.71	0.71	0.73	0.75	0.81
	Rec.	0.8	0.77	0.69	0.8	0.8	0.79	0.78	0.62	0.77	0.78	0.73	0.71	0.58	0.49	0.82
	F1	0.73	0.71	0.69	0.73	0.8	0.71	0.71	0.64	0.7	0.82	0.72	0.71	0.65	0.59	0.82
POS	**Prec.**	0.71	0.72	0.72	0.71	0.88	0.66	0.67	0.63	0.67	0.86	0.62	0.62	0.57	0	0.84
	Rec.	0.64	0.62	0.49	0.64	0.84	0.57	0.54	0.44	0.56	0.73	0.61	0.58	0.45	0	0.8
	F1	0.67	0.67	0.59	0.67	0.86	0.61	0.6	0.52	0.61	0.79	0.62	0.6	0.51	0	0.82
Accuracy		0.73	0.71	0.67	0.73	0.85	0.69	0.69	0.61	0.69	0.82	0.7	0.69	0.61	0.48	0.85

Table 6.20 Performance comparison of sarcasm detection approaches using BERT on Dataset 6 (Vegetarianism - emoji only)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.82	0.78	0.67	0.83	0.78	0.82	0.81	0.63	0.82	0.87	0.79	0.77	0.61	0.77	0.83
	Rec.	0.68	0.68	0.74	0.72	0.87	0.53	0.57	0.61	0.57	0.94	0.56	0.6	0.63	0.57	0.93
	F1	0.74	0.73	0.7	0.77	0.82	0.64	0.67	0.62	0.68	0.9	0.66	0.67	0.62	0.66	0.88
NEU	**Prec.**	0.5	1	0.67	0.5	0.6	0.33	0.38	0.38	0.33	0.5	0.75	0.75	0.75	0.75	0.53
	Rec.	0.13	0.13	0.13	0.13	0.4	0.14	0.14	0.14	0.14	0.23	0.1	0.1	0.1	0.1	0.28
	F1	0.21	0.24	0.22	0.21	0.48	0.19	0.2	0.2	0.19	0.31	0.18	0.18	0.18	0.18	0.36
POS	**Prec.**	0.82	0.82	0.83	0.83	0.94	0.78	0.79	0.79	0.79	0.93	0.78	0.79	0.78	0.79	0.93
	Rec.	0.95	0.95	0.88	0.95	0.93	0.94	0.93	0.85	0.93	0.96	0.95	0.94	0.85	0.94	0.94
	F1	0.88	0.88	0.86	0.89	0.94	0.85	0.85	0.82	0.85	0.95	0.86	0.86	0.81	0.86	0.93
Accuracy		0.81	0.81	0.79	0.82	0.88	0.77	0.78	0.74	0.78	0.9	0.79	0.79	0.73	0.78	0.89

The results from the Combined Dataset (Dataset 7 - general and emoji) are detailed in Table 6.21, aggregating results across general and emoji-rich contexts. BERT_ED_SDA consistently achieves strong overall performance, particularly in the NEG and POS classes, where it demonstrates balanced precision, recall, and F1-scores. BERT_WELMSD shows strength in NEG recall but struggles with precision, resulting in less balanced overall performance compared to BERT_ED_SDA. Similar to other datasets, the NEU class exhibits the weakest performance metrics across all methods, reaffirming the difficulty of neutral sentiment classification in complex contexts.

Table 6.21 Performance comparison of sarcasm detection approaches using BERT on Dataset 7 (Combined Dataset - General and Emoji Only)

		70:10:20					60:10:30					50:10:40				
		Orig.	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**	**Orig.**	**LogReg**	**WELMSD**	**SDA**	**ED&SDA**
NEG	**Prec.**	0.86	0.82	0.66	0.86	0.83	0.86	0.83	0.67	0.68	0.8	0.83	0.79	0.66	0.82	0.86
	Rec.	0.87	0.88	0.89	0.87	0.92	0.85	0.86	0.87	0.88	0.91	0.85	0.85	0.88	0.85	0.88
	F1	0.86	0.85	0.76	0.87	0.87	0.85	0.84	0.76	0.77	0.85	0.84	0.82	0.75	0.84	0.87
NEU	**Prec.**	0.91	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.93	0.91	0.91	0.91	0.91	0.9
	Rec.	0.92	0.82	0.82	0.92	0.93	0.92	0.92	0.83	0.92	0.92	0.91	0.9	0.81	0.91	0.92
	F1	0.92	0.86	0.86	0.92	0.92	0.92	0.91	0.87	0.92	0.92	0.91	0.9	0.86	0.91	0.91
POS	**Prec.**	0.78	0.8	0.8	0.79	0.9	0.78	0.78	0.79	0.75	0.84	0.77	0.77	0.78	0.77	0.85
	Rec.	0.74	0.57	0.57	0.74	0.68	0.76	0.73	0.6	0.34	0.66	0.73	0.69	0.59	0.73	0.76
	F1	0.76	0.66	0.66	0.76	0.78	0.77	0.75	0.68	0.47	0.74	0.75	0.73	0.67	0.75	0.8
Accuracy		0.88	0.8	0.8	0.88	0.88	0.87	0.87	0.8	0.81	0.87	0.86	0.85	0.79	0.86	0.88

Across all datasets, the following trends are evident, BERT_ED_SDA emerges as a robust approach, delivering competitive performance in NEG and POS classes. Its integration of emoji semantics and sarcasm detection methodologies ensures balanced results, particularly in emoji-rich datasets. BERT_WELMSD excels in recall for sarcasm-induced negativity but lacks precision, making its overall classification less balanced. The NEU class remains the most challenging for all approaches, reflecting the inherent complexity of neutral sentiment detection in sarcastic and emoji-laden content. These findings emphasize the need for integrating advanced language models, sarcasm-specific techniques, and emoji semantics to tackle the complexities of sentiment analysis in social media contexts. BERT_ED_SDA stands out as a robust and adaptable approach, particularly in nuanced scenarios involving sarcasm and emojis.

6.4 Summary

This chapter presented a comparative analysis of sarcasm detection approaches, focusing on the integration of the Emoji Dictionary (ED) with the Sarcasm Detection Approach (SDA) and their application to VADER and BERT classifiers. The SDA, developed to identify sarcasm through conflicts between text and emoji sentiments, demonstrated its effectiveness in improving sentiment classification accuracy, particularly in identifying sarcasm-induced negativity.

The chapter outlined the methodology, including the training of Logistic Regression and WELMSD on sarcasm-labelled datasets and the integration of ED with SDA. It explained the implementation of SDA with VADER and BERT, along with the use of comparative methods to evaluate performance across seven datasets, covering both general and emoji-rich contexts.

The results consistently showed improved performance when ED and SDA were integrated with VADER and BERT. The most notable advancements were observed in the recall of the NEG class, highlighting an enhanced capability to detect sarcasm-driven negative sentiments. While BERT_ED_SDA delivered the strongest overall performance, neutral sentiment classification (NEU class) showed limited changes. This is logical, as the methods prioritize detecting sarcasm and sentiment-rich emojis rather than neutral content.

These findings reinforce the value of combining emoji-specific dictionaries and sarcasm-aware methodologies in sentiment analysis, particularly for datasets rich in sarcasm and emojis. The integration of ED and SDA offers a robust framework for advancing sarcasm detection and sets a foundation for further research in this domain.

Chapter 7 Applying ED and SDA Across Multiple Datasets

7.1 Introduction

This chapter evaluates the performance of the proposed approach, Emoji Dictionary combined with Sarcasm Detection Approach (ED+SDA), compared to baseline methods in sentiment classification tasks. The evaluation leverages two sentiment classification tools, VADER and BERT, across seven datasets, encompassing three thematic categories—COVID-19 vaccine, electric cars, and vegetarianism—each represented in general and emoji-specific datasets, along with a combined dataset.

The analysis highlights the effectiveness of ED+SDA in improving the precision, recall, F1-score, and accuracy of sentiment classification. By incorporating both text and emojis into sentiment analysis and detecting sarcasm-induced sentiment conflicts, demonstrating its potential for robust sentiment interpretation in diverse contexts.

7.2 Experimental Setup

Datasets

Seven datasets were utilized in this evaluation:

5. COVID-19 Vaccine - General: Focused on general discussions regarding COVID-19 vaccines.
6. Electric Cars - General: Captured general opinions on electric cars.
7. Vegetarianism - General: Addressed sentiments related to vegetarianism.
8. COVID-19 Vaccine - Emoji: Analysed emoji-rich content on COVID-19 vaccines.
9. Electric Cars - Emoji: Examined emoji-based posts about electric cars.
10. Vegetarianism - Emoji: Focused on emoji-laden discussions on vegetarianism.
11. Combined Datasets: Combined all six datasets to evaluate the approach's scalability and generalizability.

Each dataset was preprocessed to align with the ED+SDA framework, which involved:

- Replacing emojis with corresponding sentiments using the Emoji Dictionary (ED).
- Identifying and resolving sentiment conflicts between text and emojis for sarcasm detection (SDA).

7.3 Results and Analysis

This section presents the results of evaluating the baseline and suggested approaches (ED+SDA) on the seven datasets, using VADER and BERT for sentiment classification. The analysis highlights enhancements in precision, recall, F1-score, and accuracy, emphasizing the advantages of incorporating ED+SDA.

7.3.1 Results Using VADER

For the COVID-19 Vaccine - general dataset, as shown in Table 7.1, ED+SDA marginally improves accuracy from 46% to 47%. Notable difference is observed in NEG sentiment, with precision increasing from 0.54 to 0.56, recall from 0.56 to 0.59, and F1-score from 0.55 to 0.57. These gains indicate better detection of negative sentiments, demonstrating the value of sarcasm detection and emoji integration.

Table 7.1 Performance comparison of baseline and suggested approaches (ED+SDA) in the COVID-19 Vaccine general dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 1073 posts	Prec.	0.54	0.56
	Rec.	0.56	0.59
	F1	0.55	0.57
NEU: 1652 posts	Prec.	0.79	0.80
	Rec.	0.33	0.33
	F1	0.47	0.47
POS: 275 posts	Prec.	0.18	0.19
	Rec.	0.80	0.80
	F1	0.29	0.30
Accuracy		0.46	0.47

For the Electric Cars - general dataset, Table 7.2shows an increase in accuracy from 48% to 49%. A difference in NEG sentiment detection are evident, with F1-score rising from 0.49 to 0.51. The consistent performance across NEU and POS classes highlights ED+SDA's balanced approach to sentiment classification.

Table 7.2 Performance comparison of baseline and suggested approaches (ED+SDA) in the Electric Cars general dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 598 posts	Prec.	0.51	0.52
	Rec.	0.47	0.49
	F1	0.49	0.51
NEU: 2094 posts	Prec.	0.85	0.85
	Rec.	0.43	0.43
	F1	0.58	0.58
POS: 308 posts	Prec.	0.19	0.19
	Rec.	0.84	0.83
	F1	0.31	0.31
Accuracy		0.48	0.49

For the Vegetarianism - general dataset, from Table 7.3, ED+SDA increases accuracy from 54% to 55%. The NEG class benefits the most, with an F1-score increase from 0.55 to 0.57. This refinement indicates enhanced recognition of nuanced negative sentiments related to vegetarianism.

Table 7.3 Performance comparison of baseline and suggested Approaches (ED+SDA) in the Vegetarianism general dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 761 posts	Prec.	0.66	0.66
	Rec.	0.47	0.51
	F1	0.55	0.57
NEU: 1473 posts	Prec.	0.76	0.77
	Rec.	0.41	0.41
	F1	0.53	0.53
POS: 766 posts	Prec.	0.40	0.40
	Rec.	0.87	0.85
	F1	0.55	0.55
Accuracy		0.54	0.55

For the COVID-19 Vaccine - Emoji Dataset, Table 7.4demonstrates a substantial accuracy jump from 47% to 64%, driven by a marked increase in NEG precision (0.76 to 0.82) and F1-score (0.62 to 0.81). These results showcase ED+SDA's ability to capture sarcasm-induced sentiment shifts in emoji-rich content.

Table 7.4 Performance comparison of baseline and suggested approaches (ED+SDA) in the COVID-19 Vaccine emoji dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 646 posts	Prec.	0.76	0.82
	Rec.	0.53	0.80
	F1	0.62	0.81
NEU: 186 posts	Prec.	0.13	0.13
	Rec.	0.08	0.08
	F1	0.10	0.10
POS: 168 posts	Prec.	0.26	0.42
	Rec.	0.69	0.65
	F1	0.38	0.51
Accuracy		0.47	0.64

For the Electric Cars - Emoji Dataset, as illustrated in Table 7.5, ED+SDA increases accuracy from 42% to 51%. Enhancements in NEG precision (0.49 to 0.60) and F1-score (0.44 to 0.65) highlight the proposed method's robustness in handling conflicting emoji-text sentiments.

Table 7.5 Performance comparison of baseline and suggested approaches (ED+SDA) in the Electric Cars emoji dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 335 posts	Prec.	0.49	0.60
	Rec.	0.40	0.70
	F1	0.44	0.65
NEU: 323 posts	Prec.	0.20	0.20
	Rec.	0.08	0.08
	F1	0.11	0.11
POS: 342 posts	Prec.	0.43	0.52
	Rec.	0.76	0.73
	F1	0.55	0.61
Accuracy		0.42	0.51

For the Vegetarianism - Emoji Dataset, Table 7.6reports an accuracy increase from 60% to 67%. The NEG F1-score rises substantially from 0.49 to 0.71, reflecting improved sarcasm detection in emoji-based vegetarianism posts.

Table 7.6 Performance comparison of baseline and suggested approaches (ED+SDA) in the Vegetarianism emoji dataset using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 268 posts	Prec.	0.58	0.67
	Rec.	0.43	0.75
	F1	0.49	0.71
NEU: 72 posts	Prec.	0.02	0.02
	Rec.	0.04	0.04
	F1	0.03	0.03
POS: 660 posts	Prec.	0.75	0.86
	Rec.	0.73	0.70
	F1	0.74	0.78
Accuracy		0.60	0.67

For the Combined Dataset, from Table 7.7, ED+SDA raises accuracy from 48% to 51%. Notable gains in NEG F1-score (0.53 to 0.61) demonstrate the approach's scalability and adaptability across diverse datasets.

Table 7.7 Performance comparison of baseline and suggested approaches (ED+SDA) in the Combined datasets using VADER

Dictionary		Original	VADER_ED_SDA
NEG: 3518 posts	Prec.	0.56	0.61
	Rec.	0.50	0.62
	F1	0.53	0.61
NEU: 6421 posts	Prec.	0.75	0.75
	Rec.	0.38	0.38
	F1	0.50	0.50
POS: 2061 posts	Prec.	0.28	0.30
	Rec.	0.77	0.76
	F1	0.41	0.43
Accuracy		0.48	0.51

7.3.2 Results Using BERT

In COVID-19 Vaccine - general dataset, Table 7.8reveals consistent accuracy across different train-test splits, with marginal differences in F1-scores across sentiments. In the NEG sentiment, recall improves notably, particularly in the 50:10:40 split (0.79 to 0.84), though at the cost of precision (0.82 to 0.72). The NEU sentiment benefits from consistent enhancements in precision and recall, leading to higher F1-scores across all splits. However, the POS sentiment sees limited gains, with a sharp drop in recall for the 50:10:40 split (0.53 to 0.08). Overall, accuracy remains steady, indicating that ED+SDA effectively enhances NEG and NEU classifications while highlighting challenges in capturing POS sentiments.

Table 7.8 Performance comparison of baseline and suggested approaches (ED+SDA) in the COVID-19 Vaccine general dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.83	0.83	0.83	0.82	0.82	0.72
	Rec.	0.81	0.81	0.76	0.8	0.79	0.84
	F1	0.82	0.82	0.79	0.81	0.81	0.78
NEU	Prec.	0.85	0.87	0.84	0.86	0.84	0.86
	Rec.	0.88	0.86	0.88	0.89	0.88	0.89
	F1	0.87	0.86	0.86	0.87	0.86	0.88
POS	Prec.	0.66	0.56	0.59	0.62	0.59	0.69
	Rec.	0.6	0.64	0.6	0.55	0.53	0.08
	F1	0.63	0.59	0.59	0.58	0.56	0.15
Accuracy		0.83	0.82	0.81	0.82	0.81	0.8

For the Electric Cars general dataset, Table 7.9 shows a refinement of BERT_ED_SDA in F1-scores for the NEG sentiment across all splits, with a notable gain in recall (e.g., 0.7 to 0.74 in the 70:10:20 split). The NEU sentiment maintains consistent precision and recall, leading to stable F1-scores. For the POS sentiment, advancements are seen in the 70:10:20 split (F1: 0.63 to 0.65), though subsequent splits show mixed results with minimal changes.

Table 7.9 Performance comparison of baseline and suggested approaches (ED+SDA) in the Electric Cars general dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.75	0.78	0.79	0.75	0.78	0.7
	Rec.	0.7	0.74	0.78	0.74	0.72	0.73
	F1	0.72	0.76	0.78	0.75	0.75	0.72
NEU	Prec.	0.9	0.92	0.91	0.9	0.89	0.91
	Rec.	0.93	0.91	0.91	0.92	0.91	0.88
	F1	0.91	0.92	0.91	0.91	0.9	0.89
POS	Prec.	0.66	0.59	0.62	0.62	0.56	0.52
	Rec.	0.6	0.71	0.61	0.54	0.54	0.59
	F1	0.63	0.65	0.61	0.58	0.55	0.55
Accuracy		0.85	0.86	0.85	0.84	0.83	0.82

For the Vegetarianism general dataset, in Table 7.10, BERT_ED_SDA exhibits balanced enhancements. The NEG sentiment benefits from increased recall, particularly in the 60:10:30 split (0.66 to 0.72), contributing to enhanced F1-scores. Similarly, NEU sentiments see consistent gains in precision, maintaining high recall and resulting in slight F1-score advancements. The POS sentiment demonstrates a slight boost in F1-scores across splits, reflecting improved recall and consistent precision. Accuracy rises moderately across all configurations, suggesting overall enhanced performance.

Table 7.10 Performance comparison of baseline and suggested approaches (ED+SDA) in the Vegetarianism general dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.74	0.72	0.69	0.7	0.72	0.68
	Rec.	0.72	0.76	0.66	0.72	0.74	0.79
	F1	0.73	0.74	0.68	0.71	0.73	0.73
NEU	Prec.	0.77	0.8	0.77	0.79	0.81	0.8
	Rec.	0.79	0.8	0.76	0.79	0.79	0.76
	F1	0.78	0.8	0.76	0.79	0.8	0.78
POS	Prec.	0.66	0.72	0.64	0.69	0.69	0.69
	Rec.	0.65	0.69	0.68	0.67	0.71	0.65
	F1	0.66	0.7	0.66	0.68	0.7	0.67
Accuracy		0.74	0.76	0.71	0.74	0.75	0.74

For the COVID-19 Vaccine Emoji Dataset, in Table 7.11, BERT_ED_SDA notably improves NEG classification, with substantial recall gains across splits (e.g., 0.84 to 0.93 in the 70:10:20 split). NEU sentiment also benefits from a notable rise in precision (0.63 to 0.77) and F1-scores. The POS sentiment shows marked gains in precision and recall for earlier splits, though the 50:10:40 split highlights challenges in recall (0.43 to 0.39). Overall, accuracy increases consistently, demonstrating the effectiveness of the suggested approach in emoji-rich datasets.

Table 7.11 Performance comparison of baseline and suggested approaches (ED+SDA) in the COVID-19 Vaccine emoji dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.86	0.93	0.86	0.91	0.82	0.82
	Rec.	0.84	0.93	0.87	0.97	0.92	0.97
	F1	0.85	0.93	0.86	0.94	0.87	0.89
NEU	Prec.	0.63	0.77	0.56	0.8	0.64	0.82
	Rec.	0.68	0.79	0.59	0.7	0.57	0.66
	F1	0.66	0.78	0.57	0.74	0.6	0.73
POS	Prec.	0.45	0.75	0.57	0.75	0.66	0.74
	Rec.	0.45	0.73	0.52	0.66	0.43	0.39
	F1	0.45	0.74	0.54	0.7	0.52	0.51
Accuracy		0.74	0.87	0.76	0.87	0.77	0.81

For the Electric Cars Emoji Dataset, in Table 7.12, BERT_ED_SDA delivers notable gains in NEG sentiment recall, considerably boosting F1-scores. NEU sentiment sees precision enhancements in all splits, enhancing F1-scores consistently. The POS sentiment also benefits from improved precision and recall, achieving higher F1-scores across splits. The overall accuracy rises markedly, emphasizing the capability of ED+SDA to handle sentiment conflicts in emoji-based datasets effectively.

Table 7.12 Performance comparison of baseline and suggested approaches (ED+SDA) in the Electric Cars emoji dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.8	0.82	0.79	0.75	0.77	0.89
	Rec.	0.74	0.91	0.71	0.94	0.76	0.93
	F1	0.77	0.86	0.75	0.84	0.76	0.91
NEU	Prec.	0.68	0.85	0.64	0.86	0.71	0.81
	Rec.	0.8	0.8	0.79	0.78	0.73	0.82
	F1	0.73	0.8	0.71	0.82	0.72	0.82
POS	Prec.	0.71	0.88	0.66	0.86	0.62	0.84
	Rec.	0.64	0.84	0.57	0.73	0.61	0.8
	F1	0.67	0.86	0.61	0.79	0.62	0.82
Accuracy		0.73	0.85	0.69	0.82	0.7	0.85

For the Vegetarianism Emoji Dataset, in Table 7.13, BERT_ED_SDA demonstrates enhancements in NEG sentiment recall, particularly in the 60:10:30 split (0.53 to 0.94), leading to a marked increase in F1-scores. While NEU sentiment precision improves slightly, challenges in recall persist, limiting F1-score gains. The POS sentiment benefits from consistent recall and precision enhancements, resulting in stronger F1-scores across splits. Accuracy sees noticeable gains, reflecting the improved classification performance of the suggested approach.

Table 7.13 Performance comparison of baseline and suggested approaches (ED+SDA) in the Vegetarianism emoji dataset using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.82	0.78	0.82	0.87	0.79	0.83
	Rec.	0.68	0.87	0.53	0.94	0.56	0.93
	F1	0.74	0.82	0.64	0.9	0.66	0.88
NEU	Prec.	0.5	0.6	0.33	0.5	0.75	0.53
	Rec.	0.13	0.4	0.14	0.23	0.1	0.28
	F1	0.21	0.48	0.19	0.31	0.18	0.36
POS	Prec.	0.82	0.94	0.78	0.93	0.78	0.93
	Rec.	0.95	0.93	0.94	0.96	0.95	0.94
	F1	0.88	0.94	0.85	0.95	0.86	0.93
Accuracy		0.81	0.88	0.77	0.9	0.79	0.89

For the Combined Dataset, in Table 7.14, BERT_ED_SDA shows consistent recall gains for the NEG sentiment across splits, boosting F1-scores (e.g., 0.84 to 0.87 in the 50:10:40 split). NEU sentiment maintains high precision and recall, resulting in stable F1-scores. For POS sentiment, slight precision gains contribute to higher F1-scores, particularly in the 50:10:40 split. Accuracy

remains steady across configurations, indicating robust performance adjustments across diverse datasets with the ED+SDA approach.

Table 7.14 Performance comparison of baseline and suggested approaches (ED+SDA) in the Combined datasets using BERT

		70:10:20		60:10:30		50:10:40	
		Original	BERT_ED_SDA	Original	BERT_ED_SDA	Original	BERT_ED_SDA
NEG	Prec.	0.86	0.83	0.86	0.8	0.83	0.86
	Rec.	0.87	0.92	0.85	0.91	0.85	0.88
	F1	0.86	0.87	0.85	0.85	0.84	0.87
NEU	Prec.	0.91	0.92	0.91	0.93	0.91	0.9
	Rec.	0.92	0.93	0.92	0.92	0.91	0.92
	F1	0.92	0.92	0.92	0.92	0.91	0.91
POS	Prec.	0.78	0.9	0.78	0.84	0.77	0.85
	Rec.	0.74	0.68	0.76	0.66	0.73	0.76
	F1	0.76	0.78	0.77	0.74	0.75	0.8
Accuracy		0.88	0.88	0.87	0.87	0.86	0.88

7.4 Discussion

This section synthesizes the insights gained from the results, emphasizing the impact of integrating the Emoji Dictionary and Sarcasm Detection Approach (ED+SDA) and exploring dataset-specific performance trends.

The integration of ED+SDA demonstrated consistent performance enhancements across both VADER and BERT. Notable increases in precision, recall, and F1-scores, particularly in detecting negative sentiments (NEG), reflect the effectiveness of ED+SDA in addressing sentiment conflicts in sarcastic content.

VADER exhibited substantial gains in NEG detection but struggled with consistent performance across datasets for NEU and POS categories. BERT achieved higher overall accuracy, benefiting from the ED+SDA integration.

In the general datasets, BERT consistently outperformed VADER, with the most notable gains observed in the electric cars dataset. This suggests that sarcasm and sentiment nuances were more effectively captured when deeper contextual understanding was applied.

Emoji datasets revealed the strongest performance increases with ED+SDA, particularly for VADER. The integration addressed the limitations of traditional lexicon-based methods, leveraging the contextual sentiment provided by emojis.

The combined dataset results illustrated the scalability of ED+SDA, with both tools benefiting from a larger, more diverse dataset. This underscores the robustness of the proposed approach across different domains and data types.

The main impact of Integrating ED+SDA is enhancing sarcasm detection by leveraging sentiment conflicts between text and emojis, ED+SDA provided a nuanced mechanism for identifying sarcasm, addressing a critical gap in traditional sentiment analysis methods. Moreover, the observed performance gains across diverse datasets indicate the versatility of ED+SDA, making it a valuable enhancement for both lexicon-based and machine learning-based sentiment analysis tools. ED+SDA's seamless integration with both VADER and BERT highlights its adaptability and potential for broader application in sentiment analysis frameworks.

7.5 Summary

The experiments in this chapter demonstrate the effectiveness of integrating the Emoji Dictionary and Sarcasm Detection Approach (ED+SDA) with sentiment classification tools (VADER and BERT) across multiple datasets. Key findings from this study include:

1. Performance Enhancements: ED+SDA consistently improved the performance of both VADER and BERT, with notable gains in precision, recall, and F1-scores for detecting negative sentiments (NEG). These results validate the effectiveness of incorporating sentiment conflicts as a proxy for sarcasm detection.
2. Tool-Specific Observations: While BERT generally outperformed VADER due to its contextual understanding, VADER demonstrated substantial gains in emoji-based datasets after ED+SDA integration, particularly in NEG sentiment detection.
3. Dataset Trends:
    - General datasets highlighted the robustness of ED+SDA in diverse contexts.
    - Emoji-based datasets showcased its ability to address the unique challenges posed by emoji-rich content, including sentiment conflicts and sarcasm.
    - The combined dataset results underscored the scalability and adaptability of the approach across various domains.
4. Implications for Sentiment Analysis and Sarcasm Detection: The integration of ED+SDA enhances traditional sentiment analysis frameworks by introducing a novel mechanism for identifying sarcasm through text-emoji sentiment conflicts. This approach addresses limitations in current methodologies, offering a more nuanced and comprehensive solution for sentiment classification tasks.

Chapter 8 Discussion

The results are analysed and explained in this chapter. The importance of the findings and how they match or differ from initial assumptions will be discussed, results' implications and how they add to the domain of existing knowledge are explored. Emphasise how these results contribute to what is already known in the field of study.

8.1 Preprocessing

For both VADER and BERT, the suggested preprocessing pipeline is the same: remove quotes, mentions, URLs, and hashtags. The effect of preprocessing becomes more pronounced based on the characteristics of the dataset. For example, if the posts have more hashtags or quotes, the effect of removing them becomes more visible. But even if the recommended preprocessing pipeline decreases some of the evaluation metrics in some cases, this is not dominant in all datasets. This pipeline has been chosen based on reasonable justifications and the results of the experiments; there is no guarantee that there are no exceptions in some values.

8.2 Optimization issues

This study established that 0.05 is the optimal threshold for running VADER by running two experiments.

By running BERT on three datasets with different dataset sizes and different dataset splits, it was found that larger datasets typically yield more accurate results since they allow the classifier to learn more quickly and recognise intricate patterns. Smaller training datasets would be more detrimental to extremely imbalanced datasets since they would provide insufficient data for the model to perform well. Not always does a larger dataset yield more accurate results; this would also depend on the quality of the dataset and whether it’s a good representative of the dataset topic and the classification task. At the same time, larger datasets demand more processing power; hence, this factor must be considered when choosing the optimal split. The optimal dataset split can be determined after experimenting with different splits and monitoring performance changes.

8.3 Understanding the impact of emojis

The effect of removing emojis in preprocessing for sentiment analysis varies based on the dataset's specific characteristics and the role emojis play in expressing sentiment. If the goal is

sentiment analysis relying only on text content, removing emojis during preprocessing can be beneficial to reduce noise and eliminate potentially irrelevant information. However, this study aims to utilise every available element that could enhance sentiment classification performance. Therefore, the introduction of the Emoji Dictionary (ED) and the Sarcasm Detection Approach (SDA) was proposed to leverage the presence of emojis and their increasing use in social media.

In this study, there is a notable enhancement in all evaluation metrics when they are used with VADER and BERT. All datasets show superior sentiment classification even when the Emoji Dictionary (ED) is applied alone, without the Sarcasm Detection Approach (SDA). This relates back to the quality of the Emoji Dictionary (ED). The enhancements noted are particularly relevant to posts incorporating emojis. The potential for better overall performance increases with the proportion of posts including emojis. The method's impact diminishes if the dataset in question includes very few emojis.

8.4 Tool Selection Considerations

The selection of sentiment analysis tools should be guided by specific requirements and considerations tailored to the objectives of the analysis.

If the primary goal is to accurately detect and interpret sarcastic expressions, VADER, with its lexicon-based approach, may excel at capturing sentiment nuances that are indicative of sarcasm.

If the sentiment analysis task involves a broad range of sentiments and expressions beyond sarcasm, a machine learning model like BERT, which has demonstrated strong generalization capabilities, might be preferred. BERT can learn complex patterns and context dependencies from data, making it adaptable for various sentiment analysis scenarios.

Considering the interpretability and explainability of the sentiment analysis model, VADER, being a lexicon-based model, provides more transparency in understanding how sentiment scores are assigned based on predefined lexical rules. This can be advantageous in scenarios where interpretability is crucial.

In assessing the computational resources required for the sentiment analysis tool, lexicon-based methods like VADER are often computationally less demanding compared to machine learning models like BERT. Resource efficiency is particularly relevant for applications with constraints on processing power and memory.

When considering the availability of labelled training data, machine learning models like BERT require substantial amounts of labelled data for training. If labelled data is limited, lexicon-based methods, such as VADER, may be more feasible.

When assessing the real-time processing capabilities of the sentiment analysis tool, lexicon-based methods often exhibit faster processing times compared to machine learning models. Real-time applications may benefit from tools with low processing latency.

If the sentiment analysis needs to accommodate multiple languages, evaluate the tool's ability to handle multilingual datasets. Some lexicon-based methods may have limitations in languages other than the ones they were initially designed for.

In summary, the nuanced performance of sentiment analysis tools, such as VADER and BERT, in specific aspects like sarcasm detection underscores the importance of aligning tool selection with the unique requirements and characteristics of the sentiment analysis task at hand. The choice should be informed by the specific goals, dataset properties, and contextual nuances of the application domain.

8.5 Research Limitation

While this research has considerably advanced sentiment analysis by incorporating emojis, developing the Emoji Dictionary (ED), and introducing the Sarcasm Detection Approach (SDA), it's important to recognise certain limitations.

Dataset Specificity: The research primarily focused on X datasets, and the findings may not generalise seamlessly to other social media platforms or text corpora. Future research is encouraged to test the proposed methodologies across varied data sources to assess their broader applicability.

Cultural and Linguistic Constraints: While the research primarily considered English-language posts, the effectiveness of the Emoji Dictionary's (ED) across different cultural and linguistic contexts remains to be explored. This opens an avenue for future research to investigate emoji usage and sentiment expression across languages and cultures.

Sarcasm Complexity: Sarcasm is inherently complex and context-dependent. While the Sarcasm Detection Approach (SDA) demonstrated promising results, future work should aim to achieve a deeper contextual and linguistic understanding for improved sarcasm detection.

Acknowledging these limitations invites further research to expand upon this study's contributions, enhancing sentiment analysis in diverse contexts and languages.

Chapter 9 Conclusion and Future Work

9.1 Research Overview

The culmination of this research journey has yielded valuable insights and advancements in the domain of sentiment analysis, leveraging VADER as a lexicon-based classifier and BERT as a machine learning model across three diverse X datasets. The primary objectives of this thesis were to enhance sentiment classification performance through the identification of an optimal preprocessing pipeline, address relevant issues, and harness the expressive power of emojis in sentiment analysis by developing an Emoji Dictionary (ED) to aid in interpreting emoji-laden sentiments and proposing a Sarcasm Detection Approach (SDA).

The exploration into preprocessing pipelines uncovered an understanding of the impact of various steps on sentiment analysis outcomes by evaluating preprocessing techniques. The goal was to identify a pipeline that balances noise reduction while preserving essential sentiment-bearing information. The findings provide practical insights into crafting preprocessing strategies tailored to the X data.

Incorporating emojis into sentiment analysis proved to be a pivotal aspect of this research. The creation of the Emoji Dictionary (ED) serves as a comprehensive reference. The Emoji Dictionary (ED) not only provides a standardised interpretation of emojis across diverse datasets but also presents an adaptive feature. This adaptability is crucial, as it enables adjustments based on specific dataset characteristics, including cultural nuances and demographic relevance. The dynamic nature of the Emoji Dictionary (ED) allows for a tailored approach, ensuring that the interpretation of emojis remains contextually relevant and aligns with the unique characteristics of each dataset, further enriching the depth of sentiment analysis.

Sarcasm, a prevalent form of expression on social media, posed a distinctive challenge within this research. The proposed Sarcasm Detection Approach (SDA), integrating both lexical and contextual features, aimed to recognise sarcastic intent by leveraging the existence of emojis and the Emoji Dictionary (ED) with the classifier. The model aimed to recognise sarcastic posts, demonstrating the applicability of the approach across different types of data. It is noteworthy to highlight that, in the comparative evaluation of the Sarcasm Detection Approach (SDA), VADER outperformed BERT in the context of sarcasm detection. This insightful observation can be attributed to the strengths of VADER in capturing sentiment nuances, especially in scenarios where sarcasm may appear in language that is subtle or nuanced. Although BERT demonstrated superior performance overall, the tailored focus on sarcasm within the Sarcasm Detection

Approach (SDA) showcased the nuanced and domain-specific advantages of leveraging VADER in conjunction with emojis and the Emoji Dictionary (ED). This nuanced performance distinction underscores the importance of selecting the most fitting tools for specific aspects of sentiment analysis.

The utilisation of three distinct X datasets added a layer of complexity and real-world applicability to the research. Testing and validating the methodologies across diverse datasets aimed to ensure the generalizability and robustness of the proposed approaches in capturing sentiments across different domains.

Reflecting on the collective findings, it becomes evident that the cooperation between the sentiment analysis approach, coupled with thoughtful preprocessing and emoji interpretation, substantially contributes to the advancement of sentiment analysis. The insights garnered from this research not only improve the accuracy of sentiment classification but also pave the way for more nuanced analyses, addressing the challenges posed by diverse, dynamic, and often ambiguous social media data.

In conclusion, this thesis underscores the importance of meticulous preprocessing, the inclusion of emojis, and the development of specialised models for nuanced aspects like sarcasm, directly addressing the research questions stated at the beginning. The findings clarify how dataset characteristics, preprocessing methods, and innovative methodologies such as the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) markedly influence the accuracy and effectiveness of sentiment analysis and sarcasm detection in X.

Regarding RQ 1, the research demonstrates that different sentiment analysis approaches yield variable results across datasets, underscoring the impact of dataset nature on analysis outcomes. In response to RQ 2, it is evident that preprocessing steps play an important role in enhancing the classification performance of sentiment analysis and sarcasm detection models, pointing to their essential role in preparing data for analysis. Lastly, addressing RQ 3, the introduction and application of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) have markedly improved classification performance, confirming the pivotal role of emojis in enriching sentiment analysis.

These insights contribute to both academic discourse and practical applications, such as social media monitoring and customer feedback analysis, marking a substantial advancement in understanding sentiment analysis complexities within X data. This research paves the way for further exploration and refinement in sentiment analysis, encouraging continued investigation into these critical factors and their interplay in enhancing sentiment analysis and sarcasm detection.

9.2 Future Work

Several directions for further investigation arise from the insights obtained from this study:

Multilingual Analysis: Extend the analysis to include multilingual datasets to understand how the proposed methodologies perform across different languages, cultures, and linguistic nuances.

Platform Variability: Investigate the transferability of the Emoji Dictionary (ED) and Sarcasm Detection Approach (SDA) to various social media platforms beyond X, considering the unique characteristics and communication styles of each platform.

Fine-tuning for Specific Domains: Explore the adaptability of the proposed methodologies to specific domains such as politics, finance, or entertainment, where sentiment analysis requirements may differ.

Dynamic Emoji Dictionary: Develop an automated mechanism to update and expand the Emoji Dictionary dynamically based on emerging emoji usage trends, cultural shifts, or modifications in the ways that people express their emotions.

9.3 Summary

This research makes substantial progress in sentiment analysis, particularly in areas such as social media monitoring and customer feedback analysis, where a nuanced understanding of sentiments is pivotal. The study has illuminated the practical utility of sentiment analysis tools, demonstrating the nuanced capabilities of VADER and BERT in sarcasm detection and how the incorporation of emojis can enhance model interpretability.

It's recommended to apply the proposed methodologies across a variety of datasets to explore their adaptability to different domains, languages, and cultural backgrounds, ensuring broader applicability and generalizability.

To maintain its utility, the Emoji Dictionary (ED) should be regularly reviewed and updated to align with evolving emoji use and cultural trends.

There's an encouragement for collaborative efforts to further refine and expand the Emoji Dictionary (ED), integrating insights from diverse communities and cultural perspectives to create more comprehensive and inclusive reference.

Pursuing interdisciplinary collaborations that bring together linguists, cultural experts, and domain specialists can enrich the methodologies of sentiment analysis with diverse perspectives.

In essence, this research not only marks a major enhancement in the field of sentiment analysis but also opens up vast opportunities for further investigations and methodological refinements. By addressing the limitations identified and embracing the recommended future work, there's a clear pathway towards the continued refinement of sentiment analysis methodologies and their applications in various contexts.

List of References

Agrawal, P., Askani, P., Nayak, R., & Mamatha, H. R. (2021). Computer Security Based Question Answering System with IR and Google BERT. *Proceedings of International Conference on Communication and Artificial Intelligence: ICCAI 2020*, 293–302.

Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, *8*(3), 27.

Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. *Chaos, Solitons & Fractals*, *140*, 110120.

Al-Shabi, M. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS*, *20*(1), 1.

Al-Shabi, M. A. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS*, *20*(1), 1.

Amalina, I. N., & Azam, Y. (2020). Cultural Interpretation of Emoji in Malaysian Context. *European Proceedings of Social and Behavioural Sciences*.

Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., & Manicardi, S. (2016). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. *KDWeb*.

Arifiansyah, R. (2024). Sentiment Analysis and Marketing Mix in Twitter Conversations About Mixue. *Ilomata International Journal of Management*, *5*(1), 133–156.

Arista, D., Sibaroni, Y., & Prasetyo, S. S. (2024). Sentiment Analysis on Twitter (X) Related to Relocating the National Capital using the IndoBERT Method using Extraction Features of Chi-Square. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(1), 403–411.

Aslam, F. (2023). *Twitter sracastic classification dataset*. https://www.kaggle.com/datasets/farhanch67/twitter-sracastic-classification-dataset/data

Bamman, D., & Smith, N. (2015). Contextualized sarcasm detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 574–577.

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in twitter sentiment analysis. *Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10*, 615–624.

Barbieri, F., Ballesteros, M., & Saggion, H. (2017). Are emojis predictable? *ArXiv Preprint ArXiv:1702.07285*.

Bipin Gupta, A. G. (2019). Logistic Regression Method for Sarcasm Detection of Text Data. *International Journal of Computer Applications*.

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1–6.

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, *47*(7), 545–566.

Burgers, C., van Mulken, M., & Schellens, P. J. (2012). Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics*, *44*(3), 231–242.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21.

Campbell, J. D., & Katz, A. N. (2012). Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, *49*(6), 459–480.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340–358.

Chen, Y., Yuan, J., You, Q., & Luo, J. (2018). Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. *Proceedings of the 26th ACM International Conference on Multimedia*, 117–125.

Churches, O., Nicholls, M., Thiessen, M., Kohler, M., & Keage, H. (2014). Emoticons in mind: An event-related potential study. *Social Neuroscience*, *9*(2), 196–202.

ÇILGIN, C., Metin, B. A. Ş., BİLGEHAN, H., & Ceyda, Ü. (2022). Twitter sentiment analysis during covid-19 outbreak with vader. *AJIT-e: Academic Journal of Information Technology*, *13*(49), 72–89.

Daniel, T. A., & Camp, A. L. (2020). Emojis affect processing fluency on social media. *Psychology of Popular Media*, *9*(2), 208.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267–3276.

DeLancey, J. (2020). *Pros and Cons of NLTK Sentiment Analysis with VADER*. https://www.codeproject.com/Articles/5269447/Pros-and-Cons-of-NLTK-Sentiment-Analysis-with-VADE

Dhola, K., & Saradva, M. (2021). A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 932–936.

Dixon, S. (2023). *Share of tweets containing emojis from July 2016 to July 2021*. Emojipedia. https://www.statista.com/statistics/1367443/share-of-tweets-containing-emojis/

Drus, Z., & Khalid, H. (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science*, *161*, 707–714.

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *122*, 16.

Farha, I. A. (2022). *iSarcasmEval*. https://github.com/iabufarha/iSarcasmEval/blob/main/README.md

Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. In *Sentiment Analysis and Ontology Engineering* (pp. 1–24). Springer.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *ArXiv Preprint ArXiv:1708.00524*.

John, N. (2020). *Five classifiers in sklearn*.

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)*, 437–442.

Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 161–169.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol*, *15*(1–2), 5–27.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12), 2009.

Gupta, S., Singh, A., & Ranjan, J. (2021). Emoji score and polarity evaluation using CLDR short name and expression sentiment. *Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020) 12*, 1009–1016.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26–32.

Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 929–932.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1).

Jianqiang, Z. (2015). Pre-processing boosting Twitter sentiment analysis? *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 748–753.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, *50*(5), 1–22.

Kaur, M., Cargill, T., Hui, K., Vu, M., Bragazzi, N. L., & Kong, J. D. (2024). A Novel Approach for the Early Detection of Medical Resource Demand Surges During Health Care Emergencies: Infodemiology Study of Tweets. *JMIR Formative Research*, *8*, e46087.

Kelly, R., & Watts, L. (2015). Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NaacL-HLT*, *1*, 2.

Khairnar, J., & Kinikar, M. (2013). Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, *3*(6), 1–6.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research*, *57*(8), 1012–1025.

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS One*, *10*(12), e0144296.

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI)*, *9*(4), 372.

Kumar, H. M. K., & Harish, B. S. (2020). A New Feature Selection Method for Sentiment Analysis in Short Text. *Journal of Intelligent Systems*, *29*(1), 1122–1134.

Kumar, P., & Sarin, G. (2022). WELMSD–word embedding and language model based sarcasm detection. *Online Information Review*, *46*(7), 1242–1256.

Kumar, T. S. S., Pandian, K. A., Aara, S. T., & Pandian, K. N. (2021). A reliable technique for sentiment analysis on tweets via machine learning and bert. *2021 Asian Conference on Innovation in Technology (ASIANCON)*, 1–5.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Lovell, R. E., Klingenstein, J., Du, J., Overman, L., Sabo, D., Ye, X., & Flannery, D. J. (2023). Using machine learning to assess rape reports: Sentiment analysis detection of officers'"signaling" about victims' credibility. *Journal of Criminal Justice*, *88*, 102106.

Malde, R. (2020). *A Short Introduction to VADER*. https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53

Md Saad, N. H., Mei Nie, F., & Yaacob, Z. (2023). Exploring sentiment analysis of online food delivery services post COVID-19 pandemic: grabfood and foodpanda. *Journal of Foodservice Business Research*, 1–25.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Mejova, Y., Weber, I., & Macy, M. W. (2015). *Twitter: a digital socioscope*. Cambridge University Press.

Muntinova, T. (2023). From Hashtags to Strategy: Twitter's Influence on Business Decisions. *Theoretical Hypotheses and Empirical Results*, *5*.

Murphy, R. (2020). *Irony-and-Sarcasm-Detection*. https://github.com/ronanmmurphy/Irony-and-Sarcasm-Detection/blob/main/data.txt

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–5.

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, *1*(4), 7.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, *1*(1), 61–67.

Obeleagu, O. U., Abass, Y. A., & Adeshina, S. (2019). Sentiment analysis in student learning experience. *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, 1–5.

Othman, S., Alshalwi, S., & Caushi, E. (2022). Sentiment of emojisets: how emoji sequences improve sentiment cognition. *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 72–79.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Patil, A., & Gupta, S. (2015). *A Review on Sentiment Analysis Approaches*. Multicon-W.

Pavalanathan, U., & Eisenstein, J. (2015). Emoticons vs. emojis on Twitter: A causal inference approach. *ArXiv Preprint ArXiv:1510.08480*.

Peacock, D. C., & Khan, H. U. (2019). Effectiveness of social media sentiment analysis tools with the support of emoticon/emoji. *16th International Conference on Information Technology-New Generations (ITNG 2019)*, 491–494.

Pohl, H., Domin, C., & Rohs, M. (2017). Beyond just text: semantic emoji similarity modeling to support expressive communication??? *ACM Transactions on Computer-Human Interaction (TOCHI)*, *24*(1), 1–42.

Pokhriyal, H., & Jain, G. (2024). Sarcasm Detection: Acknowledging Misleading Content in Social Media Using Optimised Wilson's Technique and Gumbel Mechanism. In *Harnessing Artificial Emotional Intelligence for Improved Human-Computer Interactions* (pp. 197–221). IGI Global.

Prasad, A. G., Sanjana, S., Bhat, S. M., & Harish, B. S. (2017). Sentiment analysis for sarcasm detection on streaming short text data. *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, 1–5.

Ramos, J. (2021). *SarcasmDetectionUsingLogisticRegression*. https://github.com/JomariRamos/SarcasmDetectionUsingLogisticRegression

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 704–714.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). *On stopwords, filtering and data sparsity for sentiment analysis of twitter*.

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, *36*, 101003.

Shiha, M., & Ayvaz, S. (2017). The effects of emoji in sentiment analysis. *Int. J. Comput. Electr. Eng.(IJCEE.)*, *9*(1), 360–369.

Singh, P. K., & Husain, M. S. (2014). Methodological study of opinion mining and sentiment analysis techniques. *International Journal on Soft Computing*, *5*(1), 11.

Smeureanu, I., & Bucur, C. (2012). Applying supervised opinion mining techniques on online user reviews. *Informatica Economica*, *16*(2), 81.

Subramanian, J., Sridharan, V., Shu, K., & Liu, H. (2019). Exploiting emojis for sarcasm detection. *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, 70–80.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Tepavčević, J., Blešić, I., & Vučenović, D. (2023). Sentiment analysis of reviews of spa hotels in Serbia. *International Scientific Conference on Economy, Management and Information Technologies*, *1*(1), 59–64.

Trivedi, S. K., & Singh, A. (2021). Twitter sentiment analysis of app based online food delivery companies. *Global Knowledge, Memory and Communication*.

Villasor, H. D. B., & Baradillo, D. G. (2024). NATURAL LANGUAGE PROCESSING EMPLOYING SENTIMENT ANALYSIS ON THE PUBLIC VOICE OF FILIPINOS DURING CRISIS SITUATIONS. *EPRA International Journal of Multidisciplinary Research (IJMR)*, *10*(1), 80–89.

Wallace, B. C., Kertz, L., & Charniak, E. (2014). Humans require context to infer ironic intent (so computers probably do, too). *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 512–516.

Walther, J. B., & D'addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, *19*(3), 324–347.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing twitter" big data" for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 587–592.

Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2017). Emojinet: An open service and api for emoji sense discovery. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 437–446.

Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using support vector machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 333–337.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, *89*.

강아미. (2021). *VADER 와 성향 점수를 이용한 텍스트 분류*.

Appendix A Emojis with survey results, statistics, and decisions

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
1		face with tears of joy	0.4404			1		0.8	11	2	51	Keep
2		rolling on the floor laughing	0.4939						5	1	32	Keep
3		clapping hands	0	0.3	0.3	1	0.3	1	17	2	1	Replace it with ‘proud’
4		loudly crying face	-0.4767					-1	8	0	47	Keep
5		face with rolling eyes	-0.1	-0.2			-0.6	-0.2	0	0	26	Keep
6		person facepalming	0	-0.3	-0.4		-0.7	-0.4	0	0	24	Replace it with ‘frustration’
7		thumbs up	0	0.4	0.5		0.4	1	19	5	6	Replace it with ‘approval’
8		face with symbols on mouth	0	-0.7	-0.8	-0.5	-0.7	-1	0	0	6	Replace it with ‘outburst of anger’
9		pouting face	0	-0.5	-0.9	-1	-0.7	-0.8	0	0	7	Replace it with ‘angry’
10		red heart	0.6369		0.9			1	34	4	2	Keep
11		fire	-0.34	0	-0.1	0	0.3		7	9	2	Remove
12		collision	-0.3612	0	-0.1			0	2	7	1	Remove

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
13	😳	flushed face	0	-0.1			-0.2	-0.4	1	0	15	Replace it with ‘embarrassment’
14	🥴	woozy face	0	-0.1	-0.2	-0.5	-0.3	-0.4	0	2	9	Replace it with ‘confused’
15	💀	skull	0	-0.1	-0.3		-0.3		0	0	10	Replace it with ‘frustration’
16	▶️	play button	0.34	0	0		0	0	1	4	1	Remove
17	🤨	face with raised eyebrow	0	-0.1		-0.4	-0.2		0	0	11	Replace it with ‘suspicion’
18	⚠️	warning	-0.34	0		0.5		-0.7	0	1	0	Keep
19	😅	grinning face with sweat	0.3612	0	0	-0.3	-0.3		4	3	3	Remove
20	😩	weary face	-0.2732						2	0	10	Keep
21	😢	crying face	-0.4767						1	0	6	Keep
22	💙	blue heart	0.6369			0.4			6	1	0	Keep
23	☠️	skull and crossbones	0	-0.1	-0.3			-0.4	0	0	3	Replace it with ‘danger’
24	🔹	small blue diamond	0.34	0	0			0	0	1	0	Remove
25	😱	face screaming in fear	-0.7003						2	1	4	Keep
26	🖕	middle finger	0	-0.7	-0.9	-0.5	-0.8	-0.8	0	0	2	Replace it with ‘threat’
27	⏰	alarm clock	-0.34	0	0		0	0-	1	4	0	Remove

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
28	✨	sparkles	0.3182	0					14	11	2	Keep
29	😒	unamused face	0	-0.2	-0.2	-1		-0.6	0	0	8	Replace it with ‘skepticism’
30	😎	smiling face with sunglasses	0.4588	0					8	1	0	Keep
31	💩	pile of poo	0	-0.3	-0.6	-1	-0.3	-0.4	0	0	5	Replace it with ‘aggression’
32	😊	smiling face with smiling eyes	0.7184						14	0	0	Keep
33	✌	victory hand	0.4939						9	2	0	Keep
34	😉	winking face	0	0.2		0.5	0.2	0.5	9	1	2	Replace it with ‘flirtation’
35	🎉	party popper	0.4019						1	4	0	Keep
36	😔	pensive face	0.0772	-0.2		-0.7		-0.7	0	0	8	Replace it with ‘sad’
37	😬	grimacing face	-0.34						1	0	4	Keep
38	💔	broken heart	0.2732	-0.6	-0.1	-1	-0.2	-0.8	0	0	6	Replace it with ‘grief’
39	😆	grinning squinting face	0.3612	0		1			0	1	2	Keep
40	🤞	crossed fingers	0	0.3	0.2	1			3	0	1	Replace it with ‘wishing for luck’
41	😵‍💫	dizzy	-0.2263	0	0			0	3	2	1	Remove
42	😵	knocked-out face	-0.2263	0					0	1	1	Keep

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
43	🥰	smiling face with hearts	0.8074						21	5	1	Keep
44	👎	thumbs down	0	-0.4	-0.4	-0.5	-0.6	-0.6	0	0	2	Replace it with ‘dislike’
45	🥳	partying face	0.3818						6	1	0	Keep
46	😁	beaming face with smiling eyes	0.4588						9	2	2	Keep
47	😍	smiling face with heart-eyes	0.4588						36	3	2	Keep
48	😀	grinning face	0.3612						1	0	1	Keep
49	👋	waving hand	0.4939	0					0	2	1	Keep
50	🙂	slightly smiling face	0.4033			1			2	0	2	Keep
51	😑	expressionless face	0	-0.2	-0.1		-0.4	-0.3	0	0	3	Replace it with ‘annoyance’
52	😔	disappointed face	-0.4767						0	0	1	Keep
53	😤	face with steam from nose	0	-0.6	-0.2			-0.7	0	0	4	Replace it with ‘highly annoyed’
	🗽	Statue of Liberty	0.5267	0	0		0		0	2	1	Remove
54	😌	relieved face	0.3818	-0.2				0.7	1	2	4	Keep
55	👌	OK hand	0.7297						9	4	0	Keep

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
56		credit card	0.3818	0	0		0		0	2	0	Remove
57		angry face	-0.5106						0	0	1	Keep
58		purple heart	0.6369						2	0	1	Keep
59		grinning face with smiling eyes	0.6705						3	1	0	Keep
60		small orange diamond	0.34	0	0		0	0	1	1	0	Remove
61		face vomiting	0	-0.4	-0.1		-0.7	-0.7	0	1	2	Replace it with ‘Ugh’
62		heart suit	0.6369						2	0	0	Keep
63		stop sign	-0.296	0	0	0	0		0	0	0	Remove
64		grinning face with big eyes	0.3612						1	0	1	Keep
65		frowning face	-0.34			-1			0	0	0	Keep
66		tired face	-0.4404						0	0	4	Keep
67		face with hand over mouth	0.4939	0				-0.4	0	3	3	Keep
68		nauseated face	0	-0.4	-0.2		-0.6	-0.3	0	0	2	Replace it with ‘Ugh’
69		sad but relieved face	0.3291	-0.4	0	-1		-0.3	0	0	0	Replace it with ‘Upset’
70		green heart	0.6369			0.5			28	4	0	Keep

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
71	🤗	hugging face	0.4215					1	9	1	0	Keep
72	😊	smiling face	0.4588						7	0	0	Keep
73	💞	two hearts	0.6486			0.5			9	0	1	Keep
74	🥇	1st place medal	0.4767	0	0.2		0	1	1	1	0	Keep
75	😕	confused face	-0.3182			-0.5			0	0	4	Keep
76	🤩	star-struck	0	0.4	0.3	1	0.6	0.8	11	0	1	Replace it with ‘amazing’
77	🙋	person raising hand	0.4939	0	0.1		0.2		2	1	0	Keep
78	✍	writing hand	0.4939	0	0.1	0	0	0	1	4	1	Remove
79	💠	diamond with a dot	0.34	0	0		0	0	0	0	0	Remove
80	😰	anxious face with sweat	-0.25						0	0	0	Keep
81	🔷	large blue diamond	0.34	0	0		0	0	0	1	0	Remove
82	🥴	lying face	-0.5267	0		0			0	0	0	Keep
83	✋	raised hand	0.4939	0	0		0		0	1	0	Remove
84	😖	confounded face	0	-0.3	-0.2	-0.6		-0.2	0	0	1	Replace it with ‘irritation’
85	🤓	nerd face	-0.296	0	-1	-0.5	0.2		0	0	1	Remove

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
86	😇	smiling face with halo	0.4588						3	1	0	Keep
87	😓	downcast face with sweat	-0.4215						0	0	0	Keep
88	😈	angry face with horns	-0.5106						0	0	0	Keep
89	😯	astonished face	0.3818	0	-0.1			-0.6	0	0	1	Remove
90	😣	persevering face	0	-0.3	-0.2	-1	-0.6	-0.3	0	0	2	Replace it with ‘struggle’
91	😶	face without mouth	0	-0.2	-0.2			-0.3	0	0	3	Replace it with ‘frustration’
92	😘	face blowing a kiss	0.4215						2	2	0	Keep
93	💛	yellow heart	0.6369			0.5			3	0	1	Keep
94	🤕	face with head-bandage	0		-0.1		-0.6	-0.7	0	0	1	Replace it with ‘hurting’
95	🤍	white heart	0.6369	0.6	0.5				2	0	0	Keep
96	😈	smiling face with horns	0.4588	0	-0.6	-1		-0.6	0	1	0	Replace it with ‘trouble’
97	🔆	bright button	0.4404	0		0	0	0	0	1	0	Remove
98	🎁	wrapped gift	0.4404	0	0	0.5	0	0.8	1	0	0	Remove
99	🖤	black heart	0.6369			1			4	0	0	Keep
100	😋	face savoring food	0	0.3	0.2		0.4	0.5	32	2	0	Replace it with ‘delicious’

rank	emoji	description	VADER score	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	POS	NEU	NEG	Decision
101		drooling face	0	0.3	0.1	0.7	0.5	0.5	19	2	2	Replace it with ‘delicious’
102		sparkling heart	0.7506						3	0	0	Keep
103		growing heart	0.7096						3	0	0	Keep
104		revolving hearts	0.6486						0	0	0	Keep
105		cut of meat	-0.2732	0	0	0	0	0	1	0	0	Remove
106		kiss mark	0.4215	0		1			2	2	0	Keep
107		orange heart	0.6369						4	0	0	Keep
108		brown heart	0.6369						0	0	0	Keep
109		heart exclamation	0.6369			0			1	0	0	Keep
110		smiling cat with heart-eyes	0.4588			0			1	0	0	Keep