Characteristics of opinions in the societal and non-societal domains

Loitongbam Gyanendro Singh[+], Sanasam Ranbir Singh[*]

[+]School of Electronics and Computer Science, University of Southampton, United Kingdom,
[*]Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, India.

Contributing authors: [+]gyanendrol9@gmail.com, [*]ranbir@iitg.ac.in;

Abstract

With the increasing availability of user opinions on the web, understanding the distinct nature of opinions in societal and non-societal contexts becomes crucial for opinion mining and sentiment analysis tasks. Societal topics, encompassing social unrest, terrorist acts, and government policies, differ significantly from non-societal topics like product reviews, movie reviews, and restaurant reviews. Given the regional specificity of societal issues and the lack of sentiment-annotated resources for them, this paper highlights the need to comprehend the differences in opinions between these domains for effective sentiment analysis. Through statistical text and network analysis, it investigates word usage, sentiment word association, and homogeneity in societal versus non-societal contexts. The study also explores graph-based analysis as a novel approach to sentiment analysis, considering its advantage in easily expanding context through the addition of nodes, as opposed to the complexity of inserting relevant tokens in text. The findings suggest that while non-societal sentiment resources might not be directly applicable to societal domains, graph-based analysis offers promising avenues for sentiment analysis in diverse societal topics.

Keywords: Sentiment analysis, Opinion mining, Text analysis, Graph analysis

1 Introduction

Research on sentiment analysis has gained much importance as user-generated content and social media platforms have grown rapidly since early 2000 [1, 2]. Its main goal is to discern sentiments in opinionated text, often requiring extensive annotated datasets and domain expertise. While ample annotated data exists for non-societal domains like product and movie reviews, societal topics such as social unrest and government policies lack standardized datasets. This gap stems from the vast scope and regional variance of societal issues. Sentiment analysis is also highly domain-specific, with sentiments varying across domains [3–8]. Bridging the resource gap between societal and non-societal domains necessitates understanding opinion characteristics and vocabulary associations across these domains.

Limited research has been conducted on analyzing the characteristics of opinion in societal versus non-societal contexts. Karamibekr and Ghorbani [4] have highlighted the contrast in feature engineering between these domains. They observed simpler dynamics in product reviews compared to the complexity in societal discussions, which encompass diverse expressions and sub-topics. This complexity is evident in the varied linguistic styles, with product reviews often focusing on features using adjectives, and societal discussions expressing sentiments through verbs and discussing broader impacts and sub-topics [9]. To overcome the limitations of language-dependent tools, this study extends the exploration of opinion characteristics in societal and non-societal domains by implementing language-independent statistical methods. These methods provide a quantitative approach to understanding the patterns and relationships in data, independent of linguistic intricacies. By quantifying the strength of associations between words and clustering them based on their co-occurrence, the study offers a systematic and objective way to compare the structure of opinions across different domains. This statistical perspective enables a more data-driven analysis, providing a more comprehensive understanding of sentiment dynamics in varying contexts.

In this study, we delve into the nature of opinions in societal and non-societal domains through statistical text and network analysis across various datasets. Initially, text analysis evaluates word distribution in the corpus, applying Zipf's[1] and Heap's[2] laws [10] to examine adherence to the Principle of Least Effort[3]. The semantic word associations across different domains are explored using Pointwise Mutual Information (PMI) [11], assessing vocabulary similarities. Additionally, the study measures opinion perplexity through language models to analyze corpus similarity and homogeneity. In the graph-based analysis, the corpus is transformed into a word co-occurrence graph. Here, the clustering coefficient method assesses the strength of word relations, identifying weak or strong ties. The graph's structure is further examined by identifying connected

[1] A principle stating that the frequency of any word is inversely proportional to its rank in the frequency table

[2] Describes the number of distinct words in a text as a function of the text length

[3] https://en.wikipedia.org/wiki/Principle_of_least_effort

Table 1: Characteristics of the Experimental Datasets

Dataset	Pos	Neg	Neu	Total	Topics	Domain
`Societal`	17,304	19,869	9705	46,878	Kashmir Unrest, Pathankot Attack, Surgical Strike, GSTN[4], Demonetization, Uri Attack, Paris Agreement, Syria Crisis	Social Issue
– Kashmir Unrest	1363	3638	947	5948	–	Social Issue
– Pathankot	1044	3722	1039	5805	–	Social Issue
– Surgical Strike	2116	3278	2191	7585	–	Social Issue
– GSTN	11852	6409	4823	23084	–	Social Issue
– Demonetization	653	1540	126	2319	–	Social Issue
– Uri Attack	126	416	205	747	–	Social Issue
– Paris Agreement	83	149	147	379	–	Social Issue
– Syria Crisis	67	717	227	1011	–	Social Issue
`SemEval-2016`	1296	2491	276	4063	Atheism, Climate Change, Feminist Movement, Hillary Clinton, Legalization of Abortion	Social Issue
`Sentiment-140`[$]	799978	800024	–	1600002	Consumer reviews discussion	Product Review
`Amazon`[!]	2000000	2000000	–	4000000	Consumer reviews discussion	Product Review
`Movie Review`[+]	1000	1000	–	2000	Movie reviews discussion	Movie Review

[$] Dataset downloaded from http://help.sentiment140.com/for-students/
[!] Dataset downloaded from https://www.kaggle.com/bittlingmayer/amazonreviews
[+] Dataset downloaded from https://www.cs.cornell.edu/people/pabo/movie-review-data/

components and analyzing scale-free network characteristics through the node degree distribution's power-law exponent. This comprehensive approach provides deeper insights into how opinions are structured and associated in varying domains.

The experimental analysis reveals distinct differences between societal and non-societal datasets in terms of sentiment vocabulary overlap and linguistic characteristics. Notably, societal datasets demonstrate unique word associations and linguistic traits compared to non-societal ones. Additionally, network analysis shows that societal datasets, unlike non-societal ones, adhere to scale-free network properties, suggesting real-world network structures. This indicates the potential of network representation in enhancing sentiment analysis in societal domains, alongside text-based methods. In summary, this study has the following observations:

- Non-societal datasets show strong word associations within their sentiment vocabularies, but these associations are not evident with societal datasets.
- Societal and non-societal datasets differ significantly in their linguistic characteristics.
- Graph analysis shows that societal datasets follow scale-free network properties, which allows them to capture complex, hierarchical relationships in data, providing valuable insights not present in non-societal datasets.

2 Experimental Setup

2.1 Datasets

To study the characteristic of opinions in societal and non-societal datasets, an in-house curated `Societal` and SemEval-2016 challenge datasets are considered as societal datasets while the online available customer review datasets namely product reviews posted in Amazon[5], Twitter[6], and movie reviews [1] posted in

[5] www.amazon.com
[6] http://help.sentiment140.com/for-students/

IMDb[7] are considered as non-societal datasets. Table 1 shows the characteristics of the datasets considered in this study.

2.1.1 Dataset preparation - `Societal` dataset

This section discusses the curation process of the in-house dataset named `Societal`. We manually identified popularly used event-specific hashtags in order to collect tweets[8] of the events from Twitter. Using the Twitter Streaming API[9], we were able to crawl 50,300 tweets. Two annotators have been assigned to these tweets to annotate the sentiment (i.e., positive, negative, or neutral). The languages of interest for annotating tweets are English and code-mixed Hindi and English. Both the annotators are fluent in both English and Hindi. As a guideline for annotation, the annotators are briefed to annotate the tweets based on textual content, without considering event context such as entities engaged, tweet author information, and so on. For example, people who support the event Surgical strike may express positive sentiment tweets. However, those who opposed the event can also express negative sentiment tweets. Since the event is about attacking people, tweets with such characteristics are annotated as negative sentiments. The annotators agree on the exact sentiment of 46,878 out of 54,550 tweets, with an 82.35 Kappa coefficient. According to the annotator's judgment, majority of the tweets on societal topics have sentiment polarity while only a few tweets are objective, i.e., a few tweets with neutral sentiment. The majority of tweets with disagreement are a consequence of the annotators' judgment of neutral sentiment. The same characteristics have also been reported in the study of Maynard et al. [12]

2.1.2 SemEval-2016

This dataset was created as the challenge dataset for the SemEval-2016 Stance detection task by Saif et al. [13]. The authors performed sentiment analysis on this dataset and achieved the best performance up to 76.4 F-macro scores by leveraging an inhouse curated sentiment lexicon [14] as features. This thesis work considers using this lexicon for word correlation and association analysis.

2.1.3 Amazon product reviews

McAuley et al. [15] curated this dataset for product recommendation tasks based on product reviews and ratings. The product reviews are based on laptops, movies, and books available on the Amazon website[10]. This dataset has been used for various text-classification [16] and sentiment classification [17, 18] tasks.

[7] https://www.imdb.com/
[8] Opinionated text in Twitter
[9] http://docs.tweepy.org
[10] https://www.amazon.com

2.1.4 Sentiment-140

Go et al. [19] curated this dataset for distant supervision sentiment analysis of tweets using emoticons. The dataset was filtered using phrases based on product or movie names such as Visa, Star Trek, Nike, etc.

2.1.5 Movie reviews

This dataset was curated from the Internet Movie Database (IMDb)[11] by Pang et al. [1] for sentiment analysis. This dataset was also used in Maas et al. [20] study for word representation learning on the sentiment analysis task.

2.2 Text analysis methods

The objective of the text-based analysis study is to understand the characteristics of word usage and corpus similarity across societal and non-societal domains.

2.2.1 Word distribution analysis

According to the Principle of Least Effort, human nature desires the maximum benefit for the least effort (word usages). The statistical characteristic of word distribution across datasets is investigated using Zipf's and Heap's laws [10] to determine if the considered corpora follow natural phenomena or the vocabularies of the corpus keep evolving due to numerous user associations.

Zipf's Law states that the rank r of a word with frequency f in the corpus approximately follows the equation:

$$f(r) \propto cr^z \tag{1}$$

where c is a constant number and r is the rank based on the frequency, denoted as *f(r)* and z is approximately equal to 1. That is, the second rank word has half the occurrences of the first rank word, the third rank term has one-third of the first, and so on. A log-log graph plot of a term's frequency as a function of its rank is identically a line with slope $z = -1$, as provided by the power-law equation:

$$log(f(r)) = log(c) + zlog(r). \tag{2}$$

Heap's Law represents vocabulary size M as a function of collection size:

$$M = kT^b \tag{3}$$

where T is the total number of words occurrences in the collection, k and b are parameters. According to Heaps' law, as more text instances are accumulated, the possibilities of uncovering a widespread vocabulary from which the individual tokens are derived decreases. The motivation for Heap's law is that the

[11] https://www.imdb.com/

simplest possible relationship between collection size and vocabulary size is linear in log-log space, as in Zipf's Law. The heaps law for corpus `Reuters-RCV1` gives a slope of 0.49 and intercept = 1.64[12].

2.2.2 Association of words across domains analysis

Pointwise Mutual Information (PMI) [11] is used to analyze the semantic associations of words across various corpora [2, 21]. PMI is a quantitative measure of the co-occurrence of an event (presence or absence), such as the presence of a word in a corpus or the co-occurrence of tokens in a corpus. Mutual Information (MI) may also be used to assess how much information the presence and absence of a term contributes to the corpus under consideration. MI is the expected value or average of the PMI scores for the presence or absence of a word in the corpus. This study considers analyzing the semantic associations of the words over the considered corpora using the PMI method. Equation 4 defines the mathematical formula for finding PMI of a term t appearing in a corpus c.

$$PMI(t;\ c) = \log \frac{P(t/c)}{P(t)} \tag{4}$$

where $P(t/c)$ is the conditional probability of token t appearing in corpus c. $P(t)$ is the probability of token t in the considered corpora. PMI can also be used to find the semantic orientation of two tokens in a corpus. Equation 5 defines the mathematical formula for finding PMI of a term t_1 co-occuring with term t_2 in a corpus.

$$PMI(t_1, t_2) = \log \frac{P(t_1, t_2)}{P(t_1)P(t_2)} \tag{5}$$

where $P(t_1, t_2)$ defines the probability of tokens t_1 and t_2 co-occur, $P(t_1)$ and $P(t_2)$ is the probabilities of individual tokens in a corpus. The ratio of the PMI score defines the statistical dependency of the two tokens in a corpus.

The strength of word association with sentiment lexicon can be analyzed using the PMI score of words co-occurring with sentiment polarized words in a corpus [21]. The strength of word association with sentiment lexicon is calculated as follows:

$$SOA(w_i) = \sum_{\forall w_p \in Positive\ set} PMI(w_i, w_p) - \sum_{\forall w_n \in Negative\ set} PMI(w_i, w_n) \tag{6}$$

Here the *Positive* and *Negative* sets are the group of words from a publicly available sentiment lexicon of the respective sentiments. Word w_i is said to have positive semantic orientation when the score of $SOA(w_i)$ is positive otherwise it is said to have negative semantic orientation.

[12] http://nlp.stanford.edu/IR-book/html/htmledition/heaps-law-estimating-the-number-of-terms-1.html

2.2.3 Homogeneity and similarity of corpus analysis

A corpus is similar to itself (homogeneous) if the language in the corpus does not vary. Likewise, a corpus is comparable to another corpus if the language constructs are similar [22]. A language model can be used to estimate the likelihood of language constructs within a corpus or between corpora. The language model is a statistical model that assigns probabilities to words and sentences using probability distributions learned from training corpora. Sentences that are real and syntactically aligned to the training corpus of the language model will have a high probability score. We acknowledge that perplexity measures a model's prediction ability, which does not directly correspond to text similarity or homogeneity, as it is influenced by factors such as corpus size and topic diversity. However, we chose perplexity because it reflects how well a language model generalizes to unseen data from another corpus. A lower perplexity score on an external corpus suggests shared linguistic patterns and vocabulary usage between the training and external corpora, indicating potential homogeneity. This is because a model trained on a corpus with similar linguistic structures, topics, and styles is more likely to predict unseen data from another corpus accurately. Therefore, while perplexity is an indirect measure, it provides valuable insights into the extent to which two corpora exhibit similar language characteristics. In a statistical n-gram-based language model (n = 3 in this study), the probability of a sequence of words ($\mathbf{W} = (w_1, w_2, ..., w_N)$) can be defined as:

$$P(STRT, STRT, w_1, w_2, \cdots, w_N, END) = \prod_{k=1}^{N+1} P(w_k \mathsf{w\_–k\text{-}1''}, \cdots, w\_k - n - 1'') \tag{7}$$

where (w_{-1}, w_0) and w_{N+1} are the $STRT$[13] and END tags added to every sentence while training the language model.

Various studies have considered perplexity as an intrinsic evaluation metric for assessing language model [22, 23]. A language model (LM) with a lower perplexity score determine a better language model. Perplexity of a language model can be define as:

$$PP(W) = 2^{-\frac{1}{N} log_2 P(W)} \tag{8}$$

By measuring the perplexity of the language models while keeping the language model constant, we can assess the homogeneity and similarity of corpora.

The *homogeneity of a corpus* can be determined by training a language model over the corpus and evaluate the language model perplexity over the same corpus's testing set. A corpus is not homogeneous if the perplexity score is high, indicating that the language used in the corpus varies significantly. On the other hand, the *similarity of corpora* can be estimated by training a language model on one corpus and evaluating the perplexity on the testing set of another

[13] n-1 number of $STRT$ tags are added at the beginning of the sentence.

corpus. A corpus is not similar if the average perplexity score is high, indicating that the language used in one corpus differs from the language used in another.

2.3 Graph analysis methods

The characteristics of the datasets are analyzed from a network analysis perspective by representing each dataset in a graph structure. This analysis aims to understand the word relations regardless of the language construct used in the corpora. If the words are strongly clustered, it indicates that their relationship follows a regular syntactic convention. If the relations are disjoined or weakly clustered, it indicates that word relations are not uniform and possibly from various languages or topics.

2.3.1 Representing corpus in a graph structure

The language we use to express ourselves may be represented as a network of words connected through grammatical relationships. On social media platforms such as Twitter, users often use hashtags and mentions to convey meta-information like sentiment, emotion, topic, or entity, and to draw the attention of mentioned users to their opinions [24, 25]. Previous studies have shown that the multilayer graph structure outperforms other graph structures in representing opinions, such as a randomly generated graph with the same nodes [26] and a dependency graph representation [27], demonstrating its robustness and effectiveness in sentiment analysis tasks. Inspired by these studies [26, 27], this study considers a multi-layer network $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathcal{L})$ with $\mathcal{L} = 3$ layers to represent opinions. This network captures the relationships among keywords (K), hashtags (H), and mentions (M) in a language-independent graph structure. The network consists of directed and undirected edges to capture the co-occurrence and sequential characteristics of K, H, and M in a tweet. An edge $e_{x,y} \in \mathbf{E}$ is directed if x and y occur sequentially in a tweet where i) $x, y \in K$, ii) $x \in K$ and $y \in H \cup M$, or iii) $x \in H \cup M$ and $y \in K$. An edge $e_{x,y} \in \mathbf{E}$ is undirected if $x, y \in H \cup M$ co-occur in a tweet. An example of the multi-layer network for the tweet "*Historic day for the Nation, #GST bill passed in Lok Sabha. #Congratulations to the nation,salute 2the vision of #PM @narendramodi ji*" is shown in Figure 1. This multi-layer network have three types of intra-layer associations $\mathbf{A} = \{\mathbf{A}^K, \mathbf{A}^H, \mathbf{A}^M\}$ and five types of bipartite associations $\mathbf{B} = \{\mathbf{B}^{HM}, \mathbf{B}^{MK}, \mathbf{B}^{HK}, \mathbf{B}^{KM}, \mathbf{B}^{KH}\}$ where $\mathbf{A}^i \in \mathcal{R}^{N^i \times N^i}$ is the adjacency matrix in layer $i \in \{K, H, M\}$, $\mathbf{B}^{i,j} \in \mathcal{R}^{N^i \times N^j}$ is the inter-layer relation between layer i and layer j, and N^i is the number of nodes in the respective layers. This network can also be viewed as one flattened representation in form of supra-adjacency matrix S, with total nodes $N = \mathsf{–V''\hat{}H} + \mathsf{–V''\hat{}M} + \mathsf{–V''\hat{}K}$,

$$\mathbf{S}_{N\times N} = \begin{bmatrix} \mathbf{A}^H & \mathbf{B}^{HM} & \mathbf{B}^{HK} \\ \mathbf{B}^{MH} & \mathbf{A}^M & \mathbf{B}^{MK} \\ \mathbf{B}^{KH} & \mathbf{B}^{KM} & \mathbf{A}^K \end{bmatrix} \tag{9}$$

Tweet:*Historic day for the Nation, #GST bill passed in Lok Sabha. #Congratulations to the nation,salute 2the vision of #PM @narendramodi ji*

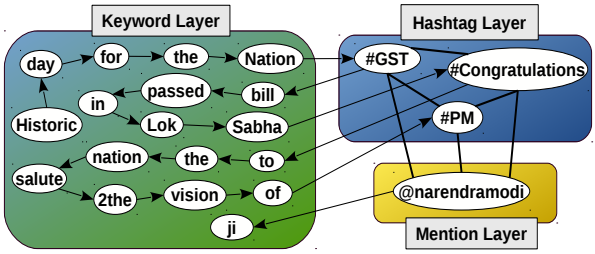


Fig. 1: An example of representing a tweet to a heterogeneous multi-layer network structure.

The intra-layer associations **A**s are on the main-diagonal, and the cross-layer connections **B** are on the off-diagonal elements of **S**. Further, $\mathbf{A}^K, \mathbf{B}^{HK}, \mathbf{B}^{KH}, \mathbf{B}^{MK}, \mathbf{B}^{KM}$ are asymmetric matrices and other matrices of **S** are symmetric. In similar fashion a tweet or a collection of tweets can be represented as a multi-layer network.

2.3.2 Clustering Coefficient

Clustering Coefficient (CC) is a measure of how strongly nodes in a network are clustered. It assesses the ego network[14] property to estimate the likelihood of a node being associated with another. The CC is computed by measuring the density of the subgraphs that remain connected after eliminating ego and the edges that are incident on ego. The CC can be categorized into two versions, namely global and local. The global version depicts the network's overall clustering, whereas the local version depicts the cohesiveness of individual nodes. This study aims to evaluate if the word associations in the graph are of weak or strong ties using the average estimates of local clustering coefficients for selected sentiment-oriented seed nodes in the graph. Given a graph $G = (V, E)$ with V nodes and E edges, the local clustering coefficient of a node (C_i) can be define as:

$$C_i = \frac{\{\mathsf{e}_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}}{k_i(k_i - 1)} \tag{10}$$

where Ni and ki denote the set of neighboring nodes and the number of neighboring nodes of ego i, respectively. The average clustering coefficient is the average of the local clustering coefficient scores of the sentiment seed nodes in the graph G.

[14]A subgraph based on the connection of one central node known as the ego in a graph.

Table 2: Slopes and intercepts of Zipf and Heap plots.

	Zipf		Heap	
Dataset	Slope	Intercept	Slope	Intercept
Societal	−0.651	−5.591	0.646	2.312
SemEval 2016	−0.351	−2.495	0.787	1.087
Sentiment140	−0.478	−5.322	0.714	1.812
Amazon	−0.777	−9.167	0.691	3.150
Movie	−0.966	−7.911	0.513	3.684

2.3.3 Connected components

A connected component (or simply component) is a network subgraph that is disconnected from other components. In a network, there can exist multiple components. Among the components, there exists a giant component where a significant amount of the nodes in the network are connected. The purpose of this study is to investigate if word associations in vocabularies are isolated or clustered, regardless of whether the associations are weak or strong. If the network has many components, it implies that the word associations in the individual components are related to a comparable syntactic word convention.

2.3.4 Scale free network analysis

A scale-free network is defined as one that asymptotically follows a power-law degree distribution. Any real-world network can be interpreted as power-law degree distributions, such as follower-followee networks in social networks like Twitter and Instagram, airway and railway routes, and so on. Since the language we use to express ourselves is a network of words linked together through syntactic relationships, in this study, we would like to investigate if the opinions follow a scale-free network property. The degree distribution of a network having k nodes can be defined as follows:

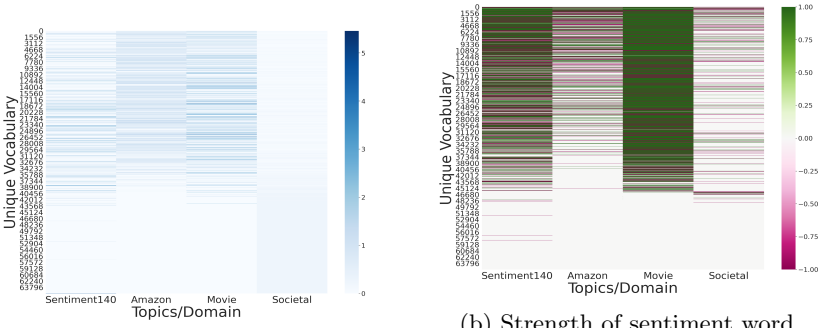
$$P_{deg}(k) = k^{-\gamma} \tag{11}$$

where γ is a parameter typically in the range $2 < \gamma < 3$ for a scale-free network. The function $P_{deg}(k)$ decays slowly as the degree k increases.

3 Observations

3.1 Text analysis

The study commences with an analysis of word distribution in the corpus, employing text-based analysis and generating Zipf's and Heap's plots. Table 2 summarizes the slopes and intercepts of these plots for both societal and non-societal datasets (i.e., `Societal`, SemEval-2016, Sentiment140, Amazon, and Movie reviews). The findings reveal intriguing trends: the Movie review dataset closely adheres to the Principle of Least Effort, as evidenced by the Zipf's plot slope near −1 and the Heap's plot slope approximating 0.5. These



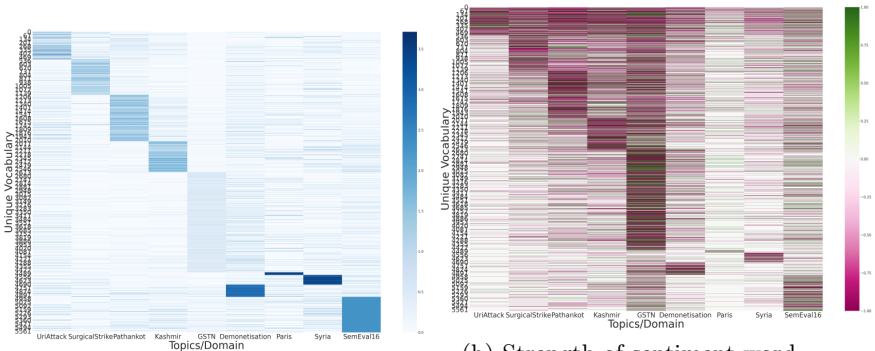
(a) Pointwise mutual information

(b) Strength of sentiment word association

Fig. 2: Heatmap plot of word vocabularies information in societal and non-societal datasets.

characteristics suggest a structured writing style. Conversely, the Amazon review dataset and Twitter datasets, namely `Societal`, SemEval-2016, and Sentiment140, exhibit steeper slopes in Zipf's plots, indicating noisy opinions characterized by misspellings, creative writing, and slang usage. These datasets only minimally follow the Principle of Least Effort. Furthermore, the slopes of Heap's plots for these corpora surpass 0.5, signifying incomplete coverage of the corpus's vocabulary. Among these datasets, it becomes evident that Movie reviews distinguish themselves with a more structured writing style compared to the others.

The subsequent analysis delves into word associations related to topics and sentiment using Pointwise Mutual Information (PMI) and the Strength of Association (SOA). Figure 2 presents heatmap plots of PMI and SOA scores for the most frequently occurring tokens across both societal and non-societal datasets. Figure 2 (a) showcases tokens with high information content in societal and non-societal datasets. Remarkably, there is minimal overlap of informative tokens between these two domains, suggesting distinct meanings. Informative tokens in non-societal datasets share similar informative content, further highlighting their differentiation from societal datasets. Furthermore, Figure 2 (b) illustrates the strength of association between these informative tokens and a seed sentiment lexicon. Notably, informative tokens in non-societal datasets exhibit a stronger association with sentiment lexicon words compared to the societal dataset. For instance, tokens like *ModiPunishesPak*, *IndiaStrikesBack*, *UriAttack*, *DeMonetisation*, and *KashmirUnrest* (which are less sentiment expressive) have higher information content in the societal dataset. In contrast, tokens like *beautiful*, *hate*, *best*, and *soulful* (which are more sentiment expressive) possess high information content in the non-societal datasets. These findings shed light on the nuanced differences in word associations and sentiment expressions between societal and non-societal contexts.



(a) Pointwise mutual information

(b) Strength of sentiment word association

Fig. 3: Heatmap plot of word vocabularies information of societal topics.

Table 3: Corpus homogeneity and similarity of corpora using perplexity score.

	Dataset	Societal	Sentiment140	Amazon	Movie
LM	Societal	16.32 (±2.07)	20.09 (±3.90)	17.33 (±0.91)	17.38 (±0.40)
	Sentiment140	20.21 (±2.24)	17.38 (±3.74)	16.25 (±1.19)	16.98 (±0.55)
	Amazon	20.26 (±2.20)	16.30 (±4.19)	15.37 (±1.11)	16.33 (±0.53)
	Movie	20.30 (±2.18)	16.38 (±4.15)	16.52 (±0.95)	15.50 (±0.52)

* LM: Language model

Given that the `Societal` dataset encompasses a diverse array of topics such as *Uri attack*, *Pathankot attack*, *Surgical strike*, and more, this study extends its investigation to explore word similarities associated with these topics. In Figure 3, we present a heatmap visualization depicting the Pointwise Mutual Information (PMI) and Strength of Association (SOA) scores for the most frequently occurring tokens within the `Societal` dataset, encompassing this wide range of topics. Figure 3(a) provides insights into how each topic exhibits distinct word associations that potentially offer better topic representation based on the PMI distribution. Notably, topics with similar themes, such as *Uri attack*, *Pathankot attack*, *Surgical strike*, and *Kashmir unrest*, share similar word associations. Furthermore, Figure 3(b) reveals that a majority of tokens within these topics are notably linked with negative emotions. In topics related to the Indian context, the vocabulary demonstrates a semantic orientation akin to sentiment tokens. This analysis highlights that the vocabulary used in the `Societal` dataset exhibits a weak semantic orientation compared to consumer review datasets. Additionally, Figure 3(b) underscores that topics with related themes share a similar vocabulary characterized by the same semantic orientation towards sentiment tokens within the `Societal` dataset.

To assess the homogeneity and similarity of the corpora, an intrinsic evaluation of language models (LMs) is conducted using perplexity scores[15], employing

[15] https://en.wikipedia.org/wiki/Perplexity

Table 4: Characteristics of the type of network representation of societal and non-societal datasets

	Societal	SemEval-2016	Sentiment-140	Amazon	Movie
Unique Vocabulary	50,184	11,468	605,284	2,669,763	39,969
Hashtags	10.55%	22.13%	1.44%	0.35%	0.05%
Mentions	11.05%	9.97%	51.16%	0.15%	0.03%
Keywords	78.40%	67.90%	47.39%	99.50%	99.93%
Edges	238,818	56,049	2,825,303	40,008,960	470,718
Degree$_{max}$	15,259	11,062	66,739	2,115,792	12,486
Degree$_{mean}$	15.753	23.267	282.284	1670.221	28.465
Degree$_{min}$	1	2	1	1	1
CC	100	10	11	13	1
GC	99.45%	99.67%	11.03%	79.25%	100.00%
Power_law$_{exponent}$	1.790	1.755	1.292	1.245	1.320

* **CC**: Connected Component, **GC**: Percentage of nodes belonging to Giant CC

a 10-fold cross-validation methodology. The homogeneity of each corpus is gauged by calculating the average perplexity score across its ten LMs. Since the LMs are trained using a 10-fold cross-validation approach, corpus similarity is determined by averaging the perplexity scores of the ten LMs trained on one corpus over the ten testing sets of another corpus. Table 3 presents the average perplexity scores of the language models for each corpus across their respective testing sets. Notably, the diagonal components of the table reveal that the Amazon product (15.37) and Movie (15.50) reviews datasets exhibit lower average perplexity scores compared to the `Societal` (16.32) and Sentiment140 (17.38) datasets. This suggests that the Amazon and Movie reviews datasets demonstrate greater homogeneity than the `Societal` and Sentiment140 datasets. Comparing the similarity of the `Societal` dataset to the others, it is evident that the LMs' average perplexity scores across these datasets, namely Sentiment140 (20.09), Amazon (17.33), and Movie reviews (17.38), are higher than the perplexity scores within their respective datasets (16.32). This implies that the `Societal` dataset differs significantly from these non-societal datasets, with the Sentiment140 corpus displaying the most pronounced dissimilarity. Similarly, when utilizing LMs trained on the Sentiment140 dataset, the perplexity score over the Sentiment140 dataset (i.e., 17.38) surpasses that of the Amazon (i.e., 16.25) and Movie (i.e., 16.98) datasets. This suggests that the Sentiment140 dataset shares more similarities with the Amazon and Movie reviews datasets. However, the `Societal` dataset exhibits a higher perplexity score than the Sentiment140 dataset, indicating differences in the language constructs used. Furthermore, employing LMs trained on the Amazon (15.37) and Movie (15.50) reviews datasets, the perplexity score over the `Societal` dataset exceeds 20, while the Sentiment140 dataset registers a perplexity score of roughly 16.30. This underscores that the language constructs utilized in the `Societal` dataset significantly diverge from those in the non-societal datasets.

Table 5: Average clustering coefficient of sentiment tokens in the word graph

Datasets	Positive	Negative
Societal	0.140 (±0.23)	0.141 (±0.22)
SemEval-2016	0.302 (±0.35)	0.312 (±0.35)
Sentiment140	0.290 (±0.23)	0.302 (±0.22)
Amazon	0.462 (±0.20)	0.472 (±0.19)
Movie	0.439 (±0.30)	0.473 (±0.31)

3.2 Graph-based analysis

In this section, we delve into the characteristics of the considered corpora from a network analysis perspective, utilizing a graph structure representation (as discussed in Section 2.3.1). One notable advantage of transforming tweets into a graph structure is its ability to circumvent the need for language-specific analysis. Table 4 provides a summary of various network properties, including node statistics, the number of connected components, and the number of nodes within giant connected components, across all corpora examined in this study. The statistics reveal that opinions expressed on Amazon and IMDb (movie reviews) platforms employ fewer hashtags and mentions compared to those on Twitter (`Societal`, SemEval-2016, and Sentiment140). This discrepancy could be attributed to the varying popularity of hashtags and mentions on these platforms at the time of dataset curation.

Furthermore, the Twitter datasets exhibit a substantial number of connected components, with `Societal` boasting the highest count. With the exception of product review datasets (Sentiment-140 and Amazon reviews), nearly all nodes within the considered datasets belong to giant connected components—an advantageous characteristic for the analysis of real-world social and information networks. Moreover, the $\text{Powerlaw}_{exponent}$ score for `Societal`, SemEval-2013, and SemEval-2016 approximates 2, signifying adherence to scale-free network features[16]. This observation underscores how a limited number of tokens (or nodes) are predominantly interconnected with the remaining nodes, a pattern commonly encountered in real-world social and information networks. This analysis sets the stage for a plethora of social network analysis studies that can be conducted using this tweet graph.

In addition, this study delves into node properties by employing local clustering coefficient measures to assess the strength of association between the considered sentiment lexicon and the tweet graph. Table 5 provides an overview of the average clustering coefficient scores for sentiment words across the datasets under consideration. Interestingly, the Amazon and Movie review datasets exhibit higher average clustering coefficients, exceeding 0.4, in contrast to the other datasets. This suggests that sentiment words find more coherent utilization on these platforms than on Twitter. Within the Twitter datasets, the `Societal` dataset registers the lowest average clustering coefficient, measuring at 0.14. This observation implies a notable disparity between the language

[16] https://en.wikipedia.org/wiki/Scale-free_network

employed in the `Societal` dataset and that represented by the sentiment lexicon.

4 Conclusion

This study conducts a comprehensive analysis, employing both text and graph-based methods, to delve into the intricacies of opinions within societal and non-societal datasets. Notably, social media datasets, particularly Twitter, do not follow the Principle of Least Effort in text-based statistical analysis, underscoring the distinct nature of Twitter opinions. Furthermore, the Pointwise Mutual Information (PMI) analysis unveils robust term associations among customer review datasets, in stark contrast to the minimal associations observed in the `Societal` dataset, accentuating its uniqueness. Within the societal domain, similar topics exhibit distinct traits and strong term connections. The prevalence of hashtags in Twitter datasets, relative to customer review domains, suggests their widespread use for expressing opinions on Twitter. The corpus similarity analysis highlights the divergence of the `Societal` dataset from non-societal datasets, emphasizing the heterogeneity of Twitter opinions. Additionally, network analysis uncovers scale-free network properties in the `Societal` and SemEval datasets, mirroring real-world network structures, signifying the potential of network representation in enhancing sentiment analysis. Collectively, these findings underscore the necessity for tailored sentiment analysis approaches based on dataset domain and characteristics.

References

[1] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 10, pp. 79–86 (2002)

[2] Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002). Association for Computational Linguistics

[3] Pang, B., Lee, L., *et al.*: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)

[4] Karamibekr, M., Ghorbani, A.A.: Sentiment analysis of social issues. In: Proceedings of the International Conference on Social Informatics (SocialInformatics), pp. 215–221 (2012)

[5] Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)

[6] Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, ??? (2015)

[7] Giachanou, A., Crestani, F.: Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR) **49**(2), 28 (2016)

[8] Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A., Benevenuto, F.: Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science **5**(1), 23 (2016)

[9] Karamibekr, M., Ghorbani, A.A.: A structure for opinion in social domains. In: 2013 International Conference on Social Computing, pp. 264–271 (2013). IEEE

[10] Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT press, ??? (1999)

[11] Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational linguistics **16**(1), 22–29 (1990)

[12] Maynard, D., Bontcheva, K.: Challenges of evaluating sentiment analysis tools on social media. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), pp. 1142–1148 (2016). https://www.aclweb.org/anthology/L16-1182

[13] Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT) **17**(3), 26 (2017)

[14] Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)

[15] McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 165–172 (2013)

[16] Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. Advances in neural information processing systems **28**, 649–657 (2015)

[17] Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 437–442 (2014)

[18] Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S.: Explicit factor models for explainable recommendation based on phrase-level sentiment

analysis. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 83–92 (2014)

[19] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford **1**(12), 2009 (2009)

[20] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150 (2011)

[21] Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) **21**(4), 315–346 (2003)

[22] Kilgarriff, A., Rose, T.: Measures for corpus similarity and homogeneity. In: Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, pp. 46–52 (1998)

[23] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, USA (2008)

[24] Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1031–1040 (2011)

[25] Singh, L.G., Anil, A., Singh, S.R.: She: Sentiment hashtag embedding through multitask learning. IEEE Transactions on Computational Social Systems **7**(2), 417–424 (2020)

[26] Singh, L.G., Mitra, A., Singh, S.R.: Sentiment analysis of tweets using heterogeneous multi-layer network representation and embedding. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8932–8946 (2020)

[27] Singh, L.G., Singh, S.R.: Sentiment analysis of tweets using text and graph multi-views learning. Knowledge and Information Systems, 1–21 (2024)