Analysing language use on social media for identifying malicious activities.

by

Pranav Bhandari

Master of Science(Computer Science)

Supervisors

1. Dr. Mehwish Nasim,

2. A/Prof. Greg Falzon

Submitted in partial fulfillment of the requirements for the degree of Master of Science (Computer Science)

College of Science and Engineering

Flinders University

Adelaide, Australia

October, 2022

Contents

List of Figures

List of Tables

List of Abbreviations

1. BERT: Bidirectional Encoder Representations from Transformers
2. CNN: Convolution Neural Networks
3. CSV: Comma Separated Values
4. EDA: Exploratory Data Analysis
5. LDA: Latent Dirichlet Allocation
6. LDA: Linear Discriminant Analysis
7. MFA: Moral Foundation Axis
8. MFD: Mortal Foundation Dictionary
9. NLP: Natural Language Processing
10. PCA: Principal Component Analysis
11. RNN: Recursive Neural Networks
12. SVM: Support Vector Machine
13. TF-IDF: Term Frequency-Inverse Document Frequency
14. TM: Topic Modeling

Declaration of Originality

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university.
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

_____

Pranav Bhandari

_____

Date

Executive Summary

There is a plethora of work in the domain of AI and machine learning to use facial features to detect emotion. An open question is whether algorithms can correctly model emotional markers in text, particularly in order to distinguish sensational or provocative text from other styles of writing? Can the types and style of language reliably indicate motivation and can AI systems be developed to grasp the difference in language use by malicious actors and common users?

This thesis analyzes various publicly available social media data sets to understand the language use on social media and identify malicious text. The data sets include Twitter data collected on the topic of bushfires, suspicious tweets collected from Twitter, and a threat data set collected from Youtube. Of particular interest is whether language use among groups with different opinions is similar or whether we can find some fine-grain differences. For instance, precise mechanisms that drive polarization on social networks are not fully understood. Language use amongst polarized groups may shed further light on the mechanisms of polarization in particularly in the online realm. In addition to style, content, and topic modeling, another potential mechanism of polarization is moral language use. This is because moral language tends to justify opinions by referring to uncompromising moral foundations.

This work was partially supported by the Defense Innovation Partnership grant titled "A framework for addressing design challenges in wargames". The findings of this thesis can be useful in modeling social media emulators for wargames. In recent years non-traditional wargames have gained popularity, particularly for modeling information operations, spread of diseases, and understanding financial outcomes.

Acknowledgement

I would like to express my sincere gratitude and appreciation towards my supervisors Dr. Mehwish Nasim and A/Prof Greg Falzon for their valuable supervision and guidance throughout the project. Furthermore, I would also like to thank Dr. Brett Wilkson for coordinating and providing necessary guidelines during the project. I am also grateful to my parents, relatives and friends who motivated me to do and achieve more. I feel obliged to mention and thank all the contributors of the data sets used in the thesis.

Finally, I pay my deep sense of gratitude towards Defense Innovation Partnership grant titled "A framework for addressing design challenges in wargames" for partially supporting the thesis.

Abstract

Although advances in natural language processing techniques have made significant contributions to the field of text mining with promising results, various problems are encountered in contextualizing the text to the level of performance comparable to humans. This thesis deals with various aspects of Natural Language Processing, to discover the underlying patterns and classification of the text, and contextualizing the natural language data. The thesis leverages the use of different methods to analyze the colossal amount of text present on social media to extract different intentions and behaviors in the text. Three different datasets are considered for the thesis, each of which presents malicious activities in different domains that are subjected to various levels of experimentation. The first data set deals with the spread of misinformation, the second with suspicious activities on Twitter, and the third with the spread of threat online. These datasets were chosen because they contained the mixture of both malicious and non-malicious activities which helped to differentiate the behavior of each party. The initial process starts with the Exploratory Data Analysis(EDA) of the data where various methods, such as sentiment and polarity analysis, and frequently used words are used with various illustrations to generate insights from the data. The EDA resulted in useful insights that categorized the distinct features of each of the categories(labels) from each other in the data set. In addition, experiments such as word analysis with various techniques allowed us to customize the themes of the different categories present in the data. Following the EDA, Topic Modeling for each of the datasets is performed where the underlying topics are extracted by combining the K-means clustering with Principal Component Analysis. This resulted in the discovery of different topics in the datasets that could be studied individually for different purposes. Furthermore, the moral inclination of all the documents in the corpus is discovered using the Moral Foundation Dictionary and FrameAxis methods. The documents are categorized into *vice and virtue* domains of five different moral foundation axes, and the results are analyzed. The divergence in moral scores for the individual category in each of the data sets indicated that the use of moral language is highly subjective to the topic and context of discussions. Finally, experiments are performed with the state-of-the-art NLP model called BERT, with fine-tuning of parameters to achieve an accuracy of 97% for the first data set and 98% for the second and third data sets.

1 Introduction.

The extensive amount and availability of text-based information in the online sphere is an opportunity as well as a challenge in various ways. While the purpose of social networks was to bring people and communities together for healthy discussions and information sharing, the advent of social networks also resulted in the fast spread of information disorder, causing threats to societies and individuals, and allowing the fast spread of extremism and criminal activities. With the development of social media, the world has seen very sophisticated ways of scamming people to get their personal information. Hence, it is imperative to quickly identify and understand the context of a piece of text, particularly the one that is shared on social media. Advancement in the field of Artificial Intelligence also widened the scope of languages, their interpretation, and their applications. Recent advances in AI based algorithms have shown excellent results on image-based data for the classification of emotions. However, understanding motivation, context or intent in textual data is still an active area of research.

This thesis deals with various aspects of Natural Language Processing(NLP) with respect to detecting intent (also referred to as malicious content) using 3 different datasets comprising Twitter and YouTube data. NLP is the technique by which the human language is represented and analyzed using different computational theories. This analysis of the text by machines to extract semantic and structural meaning is challenging and requires its comprehensive understanding [23]. NLP deals with various problems, including variation in order of words or phrases, derivation, inflection, and extraction of information and relationships [63]. Although various statistical methods are derived in NLP, the general rule followed is the process of inferring patterns from the set of predefined data called *training* data and generalizing the interpretation to similar fields. Apart from the classification of the text, other techniques which have been manifested as crucial parts of NLP have been presented in the thesis. First, an exploratory data analysis was performed for the three different data sets used. Furthermore, a different approach to topic modeling has been used to demonstrate the underlying topics in the corpus. Moreover, this thesis also focuses on mapping the moral valence of documents in the dataset to evaluate their nature.

In text analysis, the main techniques that are followed include classification, categorization, and clustering. Among these, the classification of texts plays a crucial role in the discovery of knowledge from the corpus [89]. Classification is the process of categorizing the document into their respective labels. Each of the data sets used in the thesis has its own labels in which the text are classified. This can be used as a base to train different models, which help to classify future input. There are two different methods applied for classification according to [89], which are the statistical approach and the machine learning approach. Statistical methods are mainly based on mathematical approaches and use various mathematical functions that form the basis for classification. However, these methods are inefficient for larger data sets and cannot be used when greater accuracy is required because they reach performance limits as sample sizes grow and problems increase in complexity. Therefore, with the wide range of variety, velocity, and volume of data, machine learning models and procedures are highly preferred in the current classification scenario. Several parametric and non-parametric classifiers, such as the Naive Bayes classifier, logistic regression, support vector machine, decision trees, and others, are used for classification tasks. In this thesis, we explore the behavior of an established classification algorithm called BERT [92, 28] in the three data sets present and observe its effectiveness across domains following the promising results it has produced in other similar contexts [26].

Initially, Exploratory Data Analysis is carried out to gain useful insights. Exploratory data analysis is considered a crucial step in data science. In the contemporary scenario where the corpus size is large, manual inspection and the generation of insights from the data are challenging [69]. In the case of Natural Language Processing, it plays a major role, which will be evident in the subsequent sections. Exploratory data analysis is a specific term used in the case of exploring data sets with computational and statistical tools and adds a significant contribution to this thesis because the datasets presented are relatively new to the field and accommodate the possibility of further investigation and exploration. Although there are various methodologies for exploring the data, it is considered an art to extract the most important information from a given data set. Exploratory data analysis is often the first step in data analysis and is used to gain a maximum understanding of the fundamental structure of the data set [70]. Various techniques have been

applied in the thesis, including the sentiment and polarity of tweets, the most frequently used words, the length of sentences, and wordcloud visualization. Individual labels present in the corpus are also analyzed along with the whole corpus so that the contribution of each category could be discovered.

Topic modeling is computed following the Exploratory Data Analysis phase. Topic Modeling is a set of techniques used to detect words and phrases in documents and group them into respective topics [49]. In simple words, topic modeling is used to discover the topics present in the documents. Topic modeling is important because it helps to discover the underlying facts and figures in a simple way that we might not be able to figure out just by skimming the documents, mostly because in real-world scenarios the corpus is large [18]. This is an unsupervised learning technique that is used to classify topics and their respective terms that contribute the most to the topic. Therefore, this classification has assumptions on distribution and statistical mixture, in which the distributional assumption refers to that similar words constitute similar topics and the statistical mixture deals with the topic concern that a document is made up of a variety of topics [68]. Topic modeling can be used to identify variables and features in documents.

Mapping the moral valence of the document is a crucial part of this thesis. Documents [1] are processed through a framework that determines and maps their moral valence, which is then carefully studied. Roy and Goldwasser [75] explain in their article that the political and social scenario in the modern world has significantly transitioned to social networks. This is mainly because it allows for real-time feedback from people and gets their emotions so that their response is known. When these messages are transmitted to the readers, everyone has their own language framework and focus points [61]. This framework is responsible for conveying biases and might also affect the understanding and opinion of the readers. In simpler words, in context, to make an aspect or issue important, these characteristics are highlighted [32, 21]. But the definition is not limited to the description of political and social events and is used in various other fields such as marketing [56, 40], public health [36], and several other domains [46]. This article includes data sets from various domains that incorporate the views and thoughts of various people. Therefore, to understand this moral sentiment or emotion, a theoretical framework called Moral Foundation Theory was formed [75]. Interesting discoveries have been made in the field, such as the relationship between these five moral foundations and various political ideologies. Furthermore, intriguing relationships between genders, religions, personalities and many more [39] have been discovered that contribute significantly to understanding the context. Hence mapping the moral valence of say a tweet(document) is the categorization of the documents into different spaces from which the moral judgement of the documents can be conferred. In this thesis, three different datasets are used, which will be described in detail in the following sections.

1.1 Focus of the thesis

The contribution of the thesis is in four different domains of NLP, as described above. They can be summarized as follows:

1. Firstly, the thesis contributes to understanding the nature of the documents with the help of Exploratory Data Analysis (EDA). Multiple tools were used to analyze the polarity of sentences, senti-

[1] Documents refer to all individual observations present in the datasets that are independently labeled, e.g. tweets.

ment analysis, and word frequency distributions.

2. The underlying topics are generated for analysis of the focus of the documents in the corpus using the combination of K means clustering with Principal Component Analysis (PCA) and topic modeling. Hence, the thesis contributes to the study by discovery and analysis of the underlying topics in the datasets used.

3. The moral valence for all the documents in the datasets is mapped, and through a comparison of all the results, the results are performed. The discovery related to moral valence mapping was made in different scenarios and contrasted to understand the moral principles of all the documents.

4. The thesis also contributes to the discovery of the best-suited classification method in the context of the data sets used and achieves a high level of accuracy in the classification.

This thesis provides a significant contribution to the growing area of social cyber security. This work analyzes various forms of malicious-text through a variety of cross-disciplinary NLP techniques. This thesis thoroughly compared and contrasted three types of dataset related to malicious activities. The insights gained from this work can serve as a foundation for more sophisticated work on intent classification.

The progression of the sections for the thesis is as follows. The second section presents a review of the literature on all of the experiments and methodologies used in this thesis. However, due to the presence of large numbers of individual topics and focuses, additional topics, namely EDA, topic modeling, mapping moral valence, and classification, are individually sectioned, where the methodology and results are combined and presented for each.

2 Literature review

This section of the literature review comprises various subsections, each containing its particular topic of discussion. Since this thesis deals with the different aspects of Natural Language Processing (NLP) and a range of approaches like topic modeling, exploratory data analysis, mapping moral valence, and understanding the details of classification, the literature review section is sub-sectioned accordingly. Three different data sets have been used in the article, each with a different theme and context. These datasets are relatively new and accessible for a wide range of explorations, and other insights can be generated.

2.1 Exploratory Data Analysis(EDA)

Exploratory data analysis explores the data sets given so that meaningful insights can be generated from them. Data analysts generally become familiar with their data using this method, which is a handy tool [76] and *helps to maximize the benefit of the data* [50]. It is important because this tool helps to discover patterns and information that are difficult to explore by general reading. This is done using descriptive statistics and visualization methods and is usually easy to read and understand. EDA explores a wide range of possibilities, such as producing or testing hypotheses. They can also be used to find hidden underlying patterns in the data set.

EDA has been utilized in a wide variety of settings, such as to analyze reviews provided by Toyota Camry users [6, 85]. Several problems, such as unwanted noise, heating/cooling, and uncomfortable seating, were discovered in the communications. Similarly, Majumder et al. [57] analyze the customer review data set to determine whether the various hypotheses or initial assumptions predicted were true or not. However, the limitations are not only for a sector; the techniques were used to improve the call center experience in [7]. [5] describe in their article the use of exploratory data analysis to explore and use the data to produce quality hypotheses for quality improvement projects. Tukey [90] in his article argued that exploratory data analysis should be prioritized before deciding on a model and confirmation.

2.1.1 Need for exploratory data analysis

EDA was used in the case of statistical problems to generate hypotheses by visualizing the different solutions [90]. Later, EDA was also offered for other purposes, such as quality improvements [27]. Since data analysis is an integral part of NLP, exploratory data analysis presents methods that can guide the exploration of data insights [90]. Other methods, such as Confirmatory Data Analysis [33], are commonly used. However, the significant difference between CDA and EDA is that CDA is used to confirm the hypothesis. On the contrary, EDA is used to explore or produce theories from documents and is the first step towards data exploration [5].

2.1.2 Methods used for exploratory data analysis

As seen previously, EDA is a versatile and essential tool in the context of the present NLP scenario. A variety of methods, such as sentiment analysis and word frequency analysis, are used for EDA in different

scenarios. Sentiment analysis is an EDA technique that is used in many Natural Language Processing applications [76, 71, 22]. Sentiment analysis refers to the determination of the emotion or attitude within a text and is defined below.

- Sentiment analysis is used to investigate the emotion or attitude toward an entity. The polarity of the documents is analyzed, determining the negativity or positivity, and hence correct actions can be taken against it. The polarity distribution also helps to determine the heart aspect of the texts and classify it into a range of emotions [71].

Other methods such as the N-gram, where the most commonly used bigrams, trigrams, etc., can be determined to understand the most commonly used terms in the corpus [71]. It can also be inferred that although there is a set of methods for EDA, the same techniques are generally not applicable for all scenarios. The need to customize these techniques according to research objectives and type of data set is crucial, and the user can use any suitable technique for data exploration instead of a fixed methodology [62]. Therefore, selecting the appropriate EDA approach is highly content-specific.

Representations for EDA are usually done using visual diagrams such as histograms, box plots, bar charts, pie charts, scatter plots, and many others in the papers discussed for univariate and bivariate data. Understanding these diagrams, their shape, and determining the relevant information from these diagrams are crucial and must be completed following the correct procedures [62]. Although EDA is considered an important tool, Jebb et al. [50] mention in their article that the lack of appropriate published material on the topic was felt. Some reviews of the original work presented by Tukey [90] were found, but most do not provide separate sections or importance for this part. Therefore, there is no proper data exploration for the data analysis phase. It was also found that even when these exploratory data analyzes were performed, they were packaged into the final confirmed product, making it difficult to visualize them in the articles [50].

2.1.3 Challenges in performing EDA.

Several challenges associated with EDA are presented in the article [50]. First, the credibility of the results must be considered. Sometimes, the result obtained might be the product of a cursory analysis or sampling error, which leads to problems in the later phase of the project. Therefore, strong and final conclusions cannot be drawn through this process and the results should be subjected to careful analysis. However, since the objective of the EDA is predefined as an exploratory tool, expectations and scope must realistically be bounded. Other problems that are evident as discussed in [17] are that the probability of finding genuine effects is lower, that the found effects may not fully align with the original effects, and finally that the explored effects are exaggerated. Furthermore, attesting the results obtained from EDA is a difficult task because they are entirely dependent on the process that is used to explore the data. If the methods are faulty, the results produced may be faulty, which is a major concern [62]. In context to this thesis, all the datasets used are relatively new and benchmarking is a challenge, and hence the process used must be carefully analyzed.

2.2 Topic Modeling

Manual exploratory literature review [10] is a time-consuming and inefficient process. Researchers usually need to go through a large volume of papers and separate the themes of these papers individually. This is limited not only to the literature review, but also to any domain where a large corpus containing several documents is to be analyzed. Topic modeling is an alternative approach to the analysis of documents in the data set and the formation of groups that have a similar theme or context [49]. The latent semantic structure of the text is derived from the corpus without prior knowledge or annotation of the data set in topic modeling [49]. Topic modeling can be used in various domains such as image classification and recognition, text classification, topic evaluation of different documents [68].

One of the problems with rigid cluster analysis is that the members of the clusters are fixed, which means that an element can only be a member of one of the clusters. This rigid classification may sometimes result in the loss of essential information. Topic modeling addresses this problem using the concept of mixed membership models[14]. The mixed membership model refers to the fact that every bit of the document has some relevance to every topic.

2.2.1 Need for Topic Modeling

Online corpus are usually large in size and may not be annotated. Many hidden topics might be present in a data source that cannot be leveraged if not known. Therefore, topic modeling is needed to discover these hidden topics that carry specific themes [10]. Furthermore, data sources are a mixture of a variety of topics, and it is not always true that all topics are a mixture and should be used for research [18]. Hence, topic modeling helps to analyze a particular topic in depth and exclude those that are not of our interest.

2.2.2 Methods for Topic Modeling

The topic modeling process has changed significantly over the years. Initially, topic modeling was performed using various clustering methods and text mining. Then I transitioned to other methods, such as latent topic models and other neural network approaches. However, the problem in the field is that there are inconsistent results for all long and short texts that caused noise and even resulted in inconsistent results [25]. This resulted in problems with bench marking and comparing results in various domains. Probabilistic graphical models are the most popular methods in the field of topic modeling [18]. Documents are viewed as a collection of words that contain underlying topics of various probabilities [99]. The first topic modeling technique that was widely used was the Latent Semantic Indexing technique, which is still used in some cases [31]. Topic modeling can also be viewed as a data compression technique, where the variance of the topic is maximized, and the collection of data in each topic is compressed.

This is the common principle of another popular topic modeling technique, called principal component analysis (PCA) [1]. However, it is not ideal to assume that all topics present in a document are mutually exclusive and that there is a distinct line between all topics in the corpus that led to the development of Latent Dirichlet Allocation (LDA) [31]. Latent Dirichlet Allocation is used for topic modeling in various articles such as [55, 59] for topic modeling in microblogs and extracting opinions from tweets.

Another important topic modeling technique is linear discriminant analysis (LDA). Unlike PCA, LDA is a supervised technique. The main motto of LDA is to maintain the separability of classes between topics while preserving the original information of the data set [66]. To obtain optimal results, eigendecomposition can be used for LDA, but it might be costly in terms of time and memory, especially for higher-dimensional datasets. Therefore, LDA can be combined with other techniques to obtain optimal results. LDA applications are wide ranging and can fit large datasets [66] and small datasets [79], but are preferred in cases where high dimensionality is faced because it is also a tool to reduce dimensionality.

2.2.3 K-Means for Topic Modeling

Although the above mentioned methods are commonly used for topic modeling, other methods such as K means clustering have also been shown to be efficient in the field [99, 77]. Document clustering is closely related to topic modeling, and one can benefit from the other when performed together. Clustering facilitates the extraction of both local topics in the clusters and global topics between documents, which is also an objective of topic modeling [96]. Rashid et al. describe in their paper [74] that clustering provides a new perspective for topic modeling and might produce better results depending on the scenarios. Sapul et al. [77] used a clustering method combined with latent dirichlet allocation, which improved the performance of topic modeling and provided better results while discovering the latest trend in various tweets. This improvement in results was due to the enhancement of the set of clustering features of the K-means.

Different types of clustering methods commonly used are partition clustering, hierarchical clustering, density-based clustering, grid-based clustering, and many more.

The number of K clusters that are initially determined by the users is given as input. Initially, the optimal value of K is determined using the elbow method [87]. Furthermore, each of the data points is considered and placed in the determined number of clusters. The centroids of the clusters are computed and iterated to find an ideal centroid by reducing the error distance between each point and the chosen centroid so that it does not change any further. Simply put, the algorithm for the *K-means* algorithm is as follows:

Algorithm 1 :- K means clustering.

1. Initially, select the number of K clusters and initialize the centroid.
2. Clusters of K numbers are formed by adjusting all data points to the centroid that are the closest.
3. Recalculate the centroid for each of the formed clusters.
4. Repeat steps 2 and 3 until the centroid does not change.

One of the problems with K-means clustering is that the value of K will be determined by the user. Therefore, determining an optimal number for clustering is a challenging task [77]. The elbow method is used to determine optimal clusters in various cases in different domains regardless of the context [87] . The concept of combining the K-means with the elbow method is that, as the value of the K cluster increases, the error increases or decreases until a stable point is reached. For example, as K moves from 2 to 3 or

from 3 to 4, and so on until the upper limit of the window is reached, a drastic drop in value is observed, which plots as an elbow. This means that better data modeling would not be received when adding any more clusters to the group. Therefore, from this graph the optimal number of clusters that are suitable for the corpus can be observed. The percentage of variance explained is looked at using the elbow method.

Algorithm 2 :- Determining the elbow points.

1. Initialize the value K = 1.
2. Start.
3. Increase the value of K by 1 in each iteration.
4. Calculate the sum of squared errors according to Equation E mentioned in 4. This is also known as the cost of the optimal solution.
5. Observe a point where the cost or error changes drastically.
6. This is the point that shows the true value of K.
7. End.

The thesis bases the evaluation of the topic on the K-means method combined with the elbow method, which is detailed in further sections.

2.2.4 Problems in Topic Modeling.

Some of the common problems in topic modeling presented in [73] are the lack of information and topics that are determined by a few documents only due to the short texts in the data set. The consistency of words is an important factor in determining the topic modeling of the topics, which means that words have a certain pattern to follow. Often words that represent a certain domain occur together within a topic rather than outside the topic. Another problem mentioned is that the statistical information of the words that are semantically related but rarely occur in the data set is challenging. Finally, one of the biggest challenges is the appearance of multiple topics in a single document [66].

2.3 Mapping the moral valence of documents.

Our moral values influence and define the choices and decisions made daily. In recent years, the concept of morality has also grown in the field of natural language processing [47]. The main concept behind mapping moral values is to determine the words used and analyze them to find morality. The moral foundation theory proposed by Haidit and Josheph [41] is used as a basis to determine moral valence scores in many articles.

Several methods are used to determine the moral values of tweets. Interest in extracting the sentiment and context of language from the late 1960s with the Harvard General Inquirer dictionary [84].Oscar et al. [9] mention in their paper that methodologies and the area of research have only been extended ever since.

Moral values are also used dominantly because these are considered to be closest to the people's lives, and narratives because they relate to attitude and disposition. The later phases of the long history of moral valence faced challenges due to the presence of various structures such as irony and sarcasm.

2.3.1 Evaluation of the Moral Valence.

Evaluation of moral values outset using the moral foundation dictionary [41], which was operated jointly with word counting procedures [34] and linguistic inquiry software, and was initially used to find moral values between different cultural groups. Furthermore, this method was used to obtain the moral value of news in the New York Times for the 12-year period [24]. Around 8000 tweets were also analyzed by Teernstra et al. in [88] relevant to political scenarios. Methods such as visual diagrams, bigrams, and other techniques were compared with simple machine learning techniques , to which the conclusion was drawn that machine learning techniques work better compared to simple counting of the selected moral words in the corpus since they resulted in higher accuracy.

Unsupervised and supervised methods are common in the field [11]. Supervised models are mainly based on and rely on external knowledge and frameworks. The main goal is to optimize these frameworks and increase the efficiency of the results. Deep learning methods are used to determine moral values, and these methods use metadata related to the corpus, such as tweet time, news topics and political affiliations, for the calculation [11].

Unsupervised methods, on the other hand, rely on systems where external framing annotations are not supported. The FrameAxis technique [51], which is also used in this thesis, is based on microframes. Two different measures are proposed, *Intensity and Bias*, which capture the structure of the document. This structure is influenced by the word contribution in the document. The weighted average of words towards the semantic axis was considered in [51] for the calculation of Intensity and Bias. One particular advantage of using this technique is that the risk of spurious correlation is limited because of the use of microframes, which might be introduced when pre-trained word embedding is used.

However, the FrameAxis method faces some challenges as a result of bias in various language models or word embedding. Different types of bias, such as racial or gender, have been introduced and recorded in pre-trained word embedding [15] that could create an opportunity to look at prejudice and stereotypes in society, but also could introduce an error in capturing the microframe bias. Although different methods have been applied to reduce these biases, the FrameAxis method does not depend on the a specific word embedding and hence can be benefited from future advances in word embedding that produce minimum bias [51].

2.3.2 The Moral Foundation Dictionary.

Haidt and Josheph [41, 42] suggest that humans support a small number of moral values. There are five moral foundations that can summarize morality and opinion in the text, and two sets associated with them, i.e. *vice and virtue*, make up the moral foundation axis. The five main components of *Moral Foundations* are summarized in Table 1.

The sentence-based approach is mostly followed in the field of mapping moral valence because of its simplicity and clarity. The Moral Foundation Dictionary (MFD) was used in [38] to discover moral rhetoric in general. It was also applied to discover the political discourse of tweets over a period of years. The distributed dictionary (DDR) representation was later proposed by Garten et al. [37] based on psychological dictionary and semantic similarity. This method was used in charitable giving to determine moral values that provided a promising result. Later, Garten et al. [37] presented the extended version of MFD, where language representations were incorporated by demographic embedding. The extended version of MFD outperforms its previous versions of MFD because it relies on crowd-sourced and content-driven data. This can therefore be used consistently against various domains to study moral intuitions of political, social, and communicative effects [44].

Foundation(vice/virtue)	**Descriptions**
Subversion/ Authority	Undermining the power is subversion and the right to give orders is authority
Harm/ Care	Anger and frustration towards who are spreading harm and injustice and understanding and empathy towards the vulnerable and those victimized.
Cheating/ Fairness	Fair and reliable people ans systems are treated with gratitude and cooperation
Betrayal/ Loyalty,Ingroup	pique for not being loyal / the desire to be in groups
Degradation/ Sanctity, Purity	degradation for things that are degrading and purity for pure or holy things.

Table 1: The five major categories of the *Moral Foundation Dictionary* described in [41, 42]. Each of the moral foundation are sub-divided into their vice and virtue category and these can be extracted from the text.

2.3.3 The FrameAxis Approach.

FrameAxis technique proposed by Kwak et al. in [51] is specifically designed considering the problems faced in the field and approached in a modular approach. This is an unsupervised method that is used to characterize text based on various *microframes*. The operation of each microframe is supported by antonym pairs such as *love-hate*, and for each of these antonym pairs, FrameAxis obtains *intensity and bias*.

As discussed above, the five different moral axes are represented using the vice and virtue domains [41]. All these moral axes (with antonyms) are represented in table 1. One advantage of using the FrameAxis approach is that, to demonstrate the resemblance between the word and the semantic axis, this approach takes advantage of word embedding.The following section describes the method used in the FrameAxis method, as described by Kwak et al. [51]. Initially, the set of microframes is compiled, followed by the computation of words that contribute to the microframework. Finally, the intensity and bias are computed for each category of the moral framework according to the MFD.

1. **Building Predefined Microframes.**

   In a word vector space, the "semantic axis" is initially defined, which is a microframe [8]. This semantic axis is a vector from a word to its antonym. Taking into account that the pole words or antonyms say w^+ (for positive words like good) and w^- (for negative words like bad), the vector of the semantic axis defined above is given as v_f and computed as w^+ - w^-. The semantic axis or the microframe is represented by $_f$. As this framing process is highly dependent on antonym pairs, it is important to incorporate a large number of antonym pairs to obtain better results. The antonym pairs were initially extracted from WordNet [60]. 1621 pairs of anatonyms were extracted for predefined microframes [51].

2. **Bias and Intensity calculations.**

   A pair of antonyms, as previously stated, w^+ and w^- define the microframe or semantic axis [8]. The contribution of each of the word to the microframe is the major factor that determines the bias and intensity scores from the documents. Similarity metrics are used to determine the contribution of a word w in the microframe or semantic axis f. This similarity is calculated between the microframe vector given by v_f and the word vector given by v_w. There are various methods to calculate the similarity between the vectors, and among them, cosine similarity is used because of its simplicity and effectiveness, which is given as:

$$c_f^w = \frac{v_w . v_f}{||v_w||\ ||v_f||} \tag{1}$$

   Once the similarity is computed, bias and intensity can be calculated using these figures. The formula for bias is given as follows:

$$B_f^t = \frac{\sum_w (n_w c_f^w)}{\sum_w n_w} \tag{2}$$

   where $w \in t$ and n_w denote the number of occurrences of w in t.

   For a corpus t given in the microframe f, we need to determine the bias in t. Bias is the aggregation of c_f^w to f for each of the words in t. This approach to aggregation shares the basic conceptual value with the Nelson et al. expectancy value model [67]. This is used to determine the attitude of an individual towards an issue. Word embedding is performed in the process of calculating bias, and the attribute of the corpus is each of the words represented. The feature of the attribute is the frequency

of the word in the corpus, and to evaluate this attribute, we look at the contribution of the word towards the attribute.

On the other hand, intensity determines the weight or presence of a microframe in the document. Therefore, mathematically, considering all words in t, we consider the contribution of the word calculated before, that is, c_f^w and calculate its second moment in the microframe f. For example, in a data set that describes threat, if the dominance of *violence - non-violence* is prevalent and many categories-related words are used, it can be concluded that the *violence - non-violence* microframe is authoritative in the corpus. The important thing to note is that this intensity is not dependent or associated with the microframe bias axis [51]. The calculation is given below.

$$I_f^t = \frac{\sum_w n_w (c_f^w - B_f^T)^2}{\sum_w n_w} \tag{3}$$

Here, B_f^T is the bias across the corpus T and is considered the baseline microframe and is used to calculate the second moment. The idea behind sqaring the term is that, for any words that are far from the base of bias and closer to the poles, they will subscribe more to the intensity baseline. Considering the equation of intensity presented in equation 3, it can be inferred that to get a higher scores for intensity, the majorly influencing factor is cosine similarity of a word to the microframe and the bias on the entire corpus T. As the difference between these two entities increases, the intensity value is likely to increase as well, because it is the differences between these values..

3. **Vice and virtue determination for the semantic axis**

   In the previous steps, FrameAxis [51] is followed, which calculates the relevance of the words to the semantic axis that was previously defined. The semaxis approach [8] was initially used following intensity and bias calculations based on similarity metrics. Using the five basic moral foundations and adding the domains of vice and virtue, *moral foundation axis* is created. This article promotes the use of an extended moral foundation dictionary, which contains a set of human-related terms [44]. Initially, each of the words is taken and categorized into one of the five moral groups based on the probability value by choosing the highest. Once the highest probability value is determined and the class is chosen, the sentiment analysis tool is used to categorize the word from vice (representing negative) to virtue (representing positive). For example, a moral entity *care*, its vice and virtue domain is represented as *care.vice* and *care.virtue*.

2.3.4 Challenges in mapping the Moral Valence.

Although framing is an intriguing topic and can produce insightful information as a result, there are several problems with respect to this field. The inherent vagueness in operationalizing the structure is one of the most challenging and open questions in the field [78, 83]. Another challenge is that, since the framework requires a large manual effort for most of its cycle, such as choosing the correct topic and its attributes and finding the appropriate domain of interests, this invites a lot of hurdles, especially when dealing with a large number of documents present in the corpus [21]. Various methods have been used to analyze and solve

these problems, such as recognizing political ideology presented in the articles [82, 12]. Although various approaches are proposed, the problem is that all of these studies are highly dependent on smaller data sets and contain fixed predefined ideas of moral foundations. [51].

2.4 Classification.

Classification is a dominant subject in the contemporary scenario where various novel techniques and their combinations are introduced or existing techniques are refined. Different methods such as LSTM, RNN's have been firmly established in the field of Machine Learning for purposes such as language modeling and translation [89].

2.4.1 Why Attention models for classification?

Efforts have been made to expand the boundaries of recurrent neural networks and the encoder-decoder architecture, as described in the paper by Vaswani et al [92]. Furthermore, the paper also describes the workings of the encoder-decoder architecture. The input representation $(x_1, ...x_n)$ is mapped by the encoder to the form $z = (z_1, ...z_n)$. The decoder generates the result of $(y_1, ...y_n)$ for the input z that is produced by the encoder. This is an autoregressive technique in which the output of the previous stage is taken as an extra input to generate the next output. The encoder used in the model architecture contains *multi-head self-attention* mechanisms and feed-forward networks. In the decoder part, apart from the two *attention* in the model.

Vaswani et al. [92] define attention as a function that performs mapping between queries and key-value pairs with the output where all these quantities are vectors. The queries, keys, and values were focused to create multihead attention instead of creating a single attention function that increased the performance. One aspect of this is that it helps to make possible a contextual representation of the text. Attention-based models are also faster than recurrent-based models. Convolution-based models are expensive compared to recurrent-based models [20], and have almost the same level of complexity as the combination of self-attention model and point-wise feedforward networks [92].

2.4.2 Unsupervised feature based and fine tuning approach.

Devlin et al. [28] explain in their paper that the trend of using pre-trained models is wide and where both neural and non-neural methods are used. Rather than using scratch-trained embedded procedures, the use of pre-trained embedding is more reliable and efficient [91]. Left-to-right language modeling objectives are used for the pre-trained word embedding along with the separation of correct words from the incorrect words, which has been generalized for sentence embedding and paragraph embedding [28]. Contextual representation is an integral part of word embedding, where semantic and syntactic information must be clearly represented [54]. For this, different methods such as BiLSTM [97] have been used, where each word is embedded in such a way that it depends on the context. For a sequence of words represented as $[w_1, w_2, w_3, ...w_n]$, the bidirectional LSTM works in such a way that the forward LSTM considers the sequence from w_1 to w_n, while the backward LSTM considers the sequence from w_n to w_1. Finally, the

concatenation of hidden representations of forward and backward networks provides the final representations that are context-specific. Similar concepts of extracting context-sensitive information from text by parsing them in the left-to-right and right-to-left directions have been used in ELMo presented in [58]. Although the concept of concatenation provides satisfactory results [97], these models are not considered deeply bidirectional, but feature-based [28].

Although the approach of transfer learning has dominated the field of computer vision, NLP still requires customization and training from scratch [45]. Although only a limited number of papers were discovered in the case of pre-training and fine-tuning approach, the impacts made were significant in the field. Dai and Le [26] present the concept of pre-training and to enhance the results in the later stages where various tasks, such as classification, can be performed in their paper which can be generalized across various domains. This technique uses unlabeled data and hence can use abundant training data to excel at sequence learning. They argue that these parameters that are learned from the pre-training phase can later be used in specific supervised tasks. The major benefit of this is that the pre-trained model can be used across various domains where only a few parameters are needed to train according to the context. However, the model presented by Dai and Le [26] required a large data set for overfitting, which was administered and improved in the model presented by Howard and Ruder [45] that demonstrated state-of-the-art results in small data sets. Fine-tuning has been successfully demonstrated in similar tasks, such as classification, question answering, and sentiment analysis, but has also failed in unrelated tasks [45].

This thesis aims to use data sets that are mostly related to malicious behavior that is spread on social networks. We are interested in determining the modality of language used by people involved in spreading any kind of malevolent activity, whether it is spreading misinformation, any kind of suspicious activity, threats, and violence of any type. We believe that the collection of relevant data and the derivation of a statistical and computational solution should be accompanied by the proper exploration, understanding, and analysis of the contextualization to elevate the efficiency of the primary objective. Hence we conduct our research on the basis of the above-mentioned methods which are EDA, topic modeling, mapping the moral valence, and classification.

3 Exploring data sets

With the objective of exploring and analyzing the language use in various social platforms that contain or spread malicious intent, different datasets were analyzed. After the analysis of different datasets, three different datasets were chosen for the thesis that cover different domains but had a common method of classification, i.e. malicious or non-malicious. The first data set presented by Weber et al. [94] aims to address the spread of misinformation related to the Australian bushfires of 2019-2020. The Suspicious tweet data set [2] is the second data set that targets suspicious activities on Twitter such as cyberbullying and hate. Finally, the Threat corpus presented by Hammer et al. [43] contains content related to various disagreements and feuds related to culture, religion, and county [3]. The purpose of using these dataset is that they help to identify malicious behavior concerning various domain which is the objective of the thesis. This thesis first describes the three data sets that have been used in the experiments.

3.1 The Bushfire data set(#arsonemergency data set)

Weber et al. [94] present the data set on the social media discussions during the bushfires. This was a crucial topic of discussion in 2019-2020 in Australia and around the world. The bushfire data set is the collection of tweets of people's opinions gathered during the 2019-2020 bushfire disaster in Australia. Data were collected around a trending hashtag "ArsonEmergency", indicating that arson was the cause of bushfires. There are 27,546 instances of pre-set data in CSV format. These bushfires caused significant damage to the habitats, properties, and lifestyle of people with a drastic increase in temperature pollution.

The two main communities described in [94] are *supporters and opposers*. Another unaffiliated community was also present in the data, whose affiliation was hard to discern. The article describes that *supporters* preferred the arson narrative more and used different hashtags and news to spread more information on social networks. In contrast, *opposers* focused more on presenting facts and figures and links to credible sources and retweeted more. One of the narratives of the article was to understand and contain the spread of misinformation on social networks. In the current environment, social networks are a major source of information and influence on the general public. Therefore, when misinformation is spread amongst the public, this can lead to negative impressions and cause severe problems.

Furthermore, when this kind of data and impressions are used for decision making, the risk of making a wrong decision increases. Data presented in Weber et al. [94] explain that various narratives were built around discussions that the bushfire was the result of arson and that climate change had very little or no role in it. Therefore, it is urgent to contain the spread of misinformation.

In the context of this thesis, the interest lies in the peculiarities in language used by both communities.

[2] `https://www.kaggle.com/datasets/munkialbright/suspicious-tweets`

[3] The Threat dataset is a collection of various cultural and religious conflicts between different group of people. This thesis is not concerned with any of these topics and does not promote these activities in any way. The only purpose to use the dataset is to analyze the behavior of threat comments compared to the non-threat comments, which has been presented by masking the original words used in the dataset. Also, the data are non-identifiable and no information about the source is divulged. Although the computational procedures applied for the dataset are the same as those for the other two datasets, the relating figures are not presented for this dataset.

3.1.1 Exploratory Data Analysis on The Bushfire data set

The preparation of the data set was initially completed. This step included cleaning up data in many forms. The intention is to deal with the chunks in the corpus that provide only the most valuable and ideally distinct information about each topic group. Initial preparation included removing the user tweeter handle and cleaning the punctuation, digits, links, emojis, retweet symbols, and other noise. In most analyses, only supporters and opponents are considered. This is because the agenda of the unaffiliated group is ambiguous and cannot be distinctly separated from other categories.

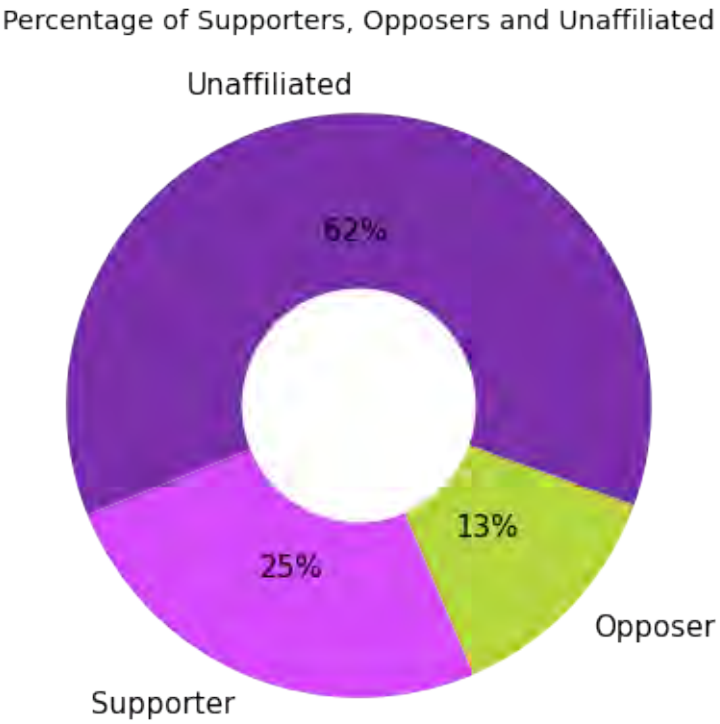

Figure 1: Distribution of Supporters, Opposers and Unaffiliated across the corpus in a pie chart. It is observable that the unaffiliated group covers a major portion of the corpus but the focus is majorly on the Supporters and Opposers, and the Unaffiliated group will be discarded for the evaluation purposes.

Sentiment analysis dives deeper into the user's opinion. Sentiment analysis ranges from -1 to 1, where -1 denotes the highest number for negative sentiments, while 1 denotes the highest number for positive sentiments. When sentiment analysis on the tweets present in the corpus was performed, most were found to be -0.25 to +0.25. This is justifiable because most people are discussing a topic and mainly present facts and figures, rather than any kind of personal feelings or sentiments. However, a significant number of tweets have positive and negative sentiments attached to them. Figure 2 demonstrates more positive sentiments than negative ones, as we observe from the center to the left and right directions.

In context to the #arsonemergency data set, there are three different categories of people in the data set, that is, *supporters, opposers, and unaffiliated*. Therefore, the descriptions can be analyzed and it can be seen whether they have a positive or negative analysis. The polarity of descriptions for supporters and

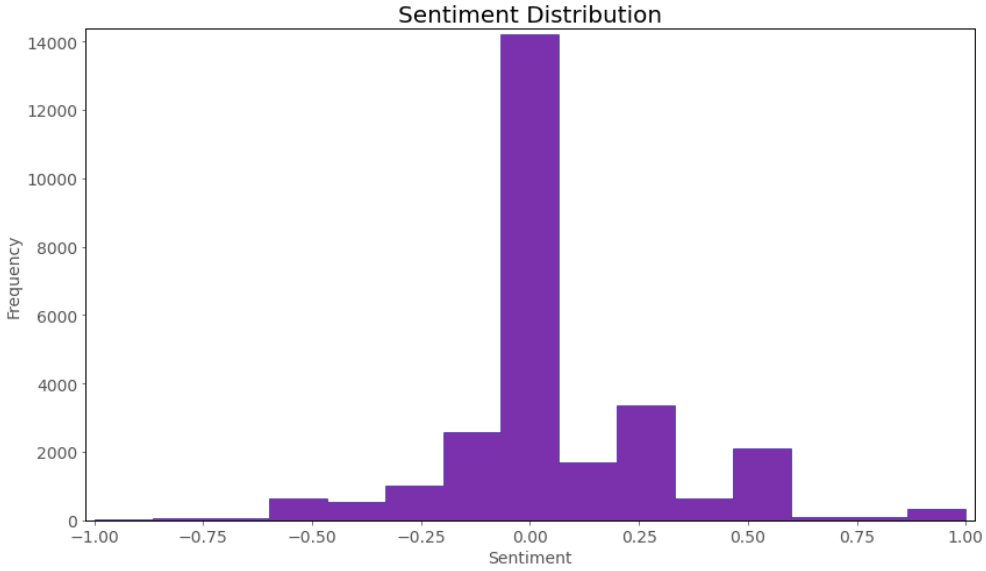

Figure 2: Histogram representing the sentiment distribution across the #arsonemergency dataset.

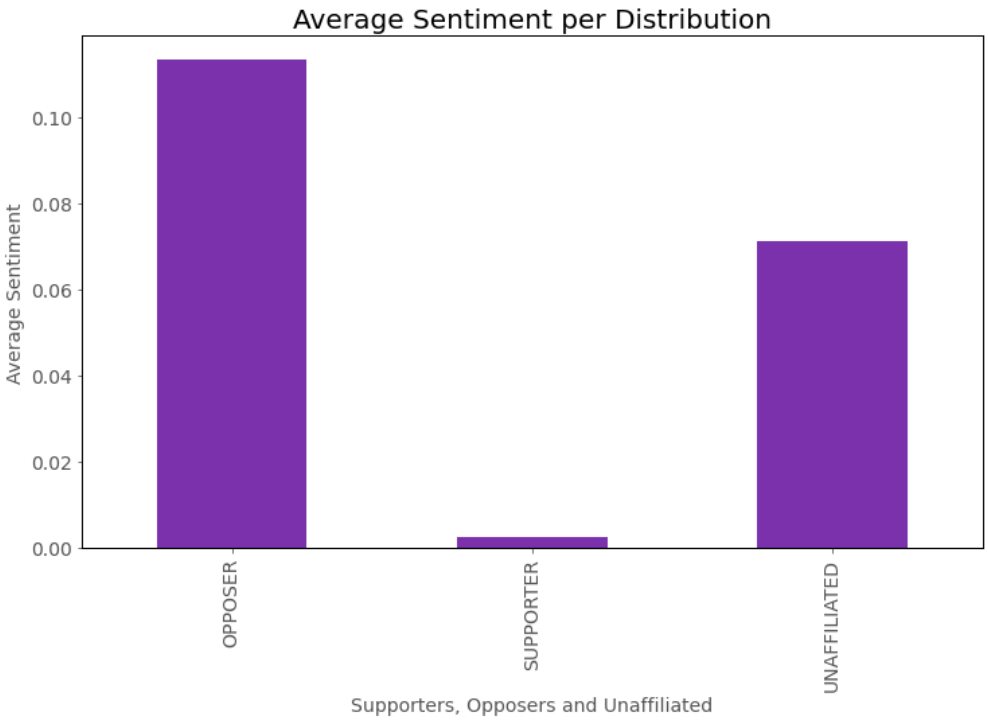

Figure 3: Average sentiment analysis in each category. The average sentiment for the opposer category is significantly higher than the supporter category. The average sentiment for the unaffiliated lies between opposers and supporters, but it is significantly higher than the supporter category.

opposers is expected to be different from or opposite, but the opinion cannot be held firm. Furthermore, the polarity of the unaffiliated group remains to be seen. Figure 3 shows that, compared to supporters of arson theory, opposers have a more positive attitude toward comments. The difference is significantly higher for the opposing category. This might be due to their aim to spread positive information while containing the

flow of misinformation among the public. In the case of the unaffiliated group, it is seen that the average sentiment is significantly higher than that of the supporter category. This shows that the average sentiment is comparable to the opposer category, and the thoughts of the unaffiliated category are similar to the opposing category. The average sentiment for the supporters of arson theory is minimum among all of the groups and close to the neutral sentiment. This contradicts the initial assumption that since the corpus is a collection of debates, the sentiments expressed must be negative (for at least one of the two groups, if not both).

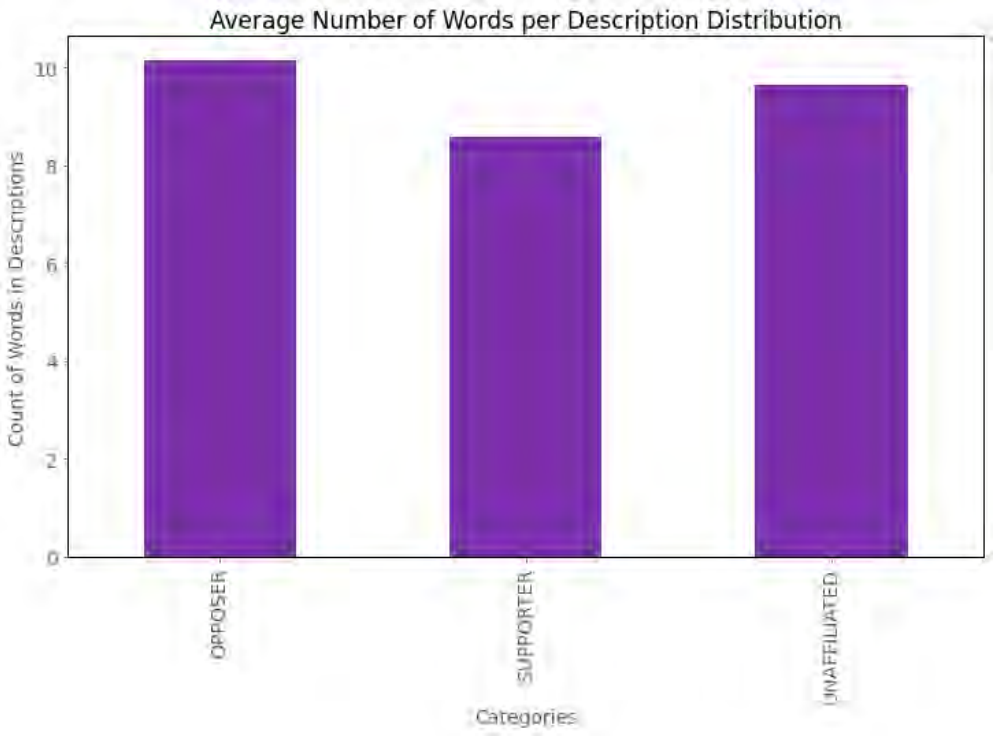

Figure 4: Average Word counts in each of the category

Figure 4 represents the average number of words in each category. In the analysis of these average results, it is seen that the opposing community explained more and used more words compared to the supporting community. One reason might be that the opposers mainly based their tweets on explaining the facts citing official sources and validating their points with specific reasons. The unaffiliated category has a word length similar to that of the opposing category, which is also accompanied by sentiment scores similar to those shown in Figure 3.

When analyzing the frequently used words in the category *opposer*, it can be seen that they used terms such as “disinformation”, “misinformation”, “bot”, “lightning”, and “climate emergency” more often. It can be inferred that they emphasized more than misinformation was spread, and “climateemergency” is the kind of topic we are looking for in the bushfire, whereas the story comes different when we analyze the supporter category. Most of the tweets emphasize common words such as “arsonemergency”, “arson”, and “arsonist”. This shows that the focus is particularly on the theory of arson and the arsonist. “arsonemergency” has been used the most in both categories because all tweets and retweets followed the same hashtags. Overall, it can

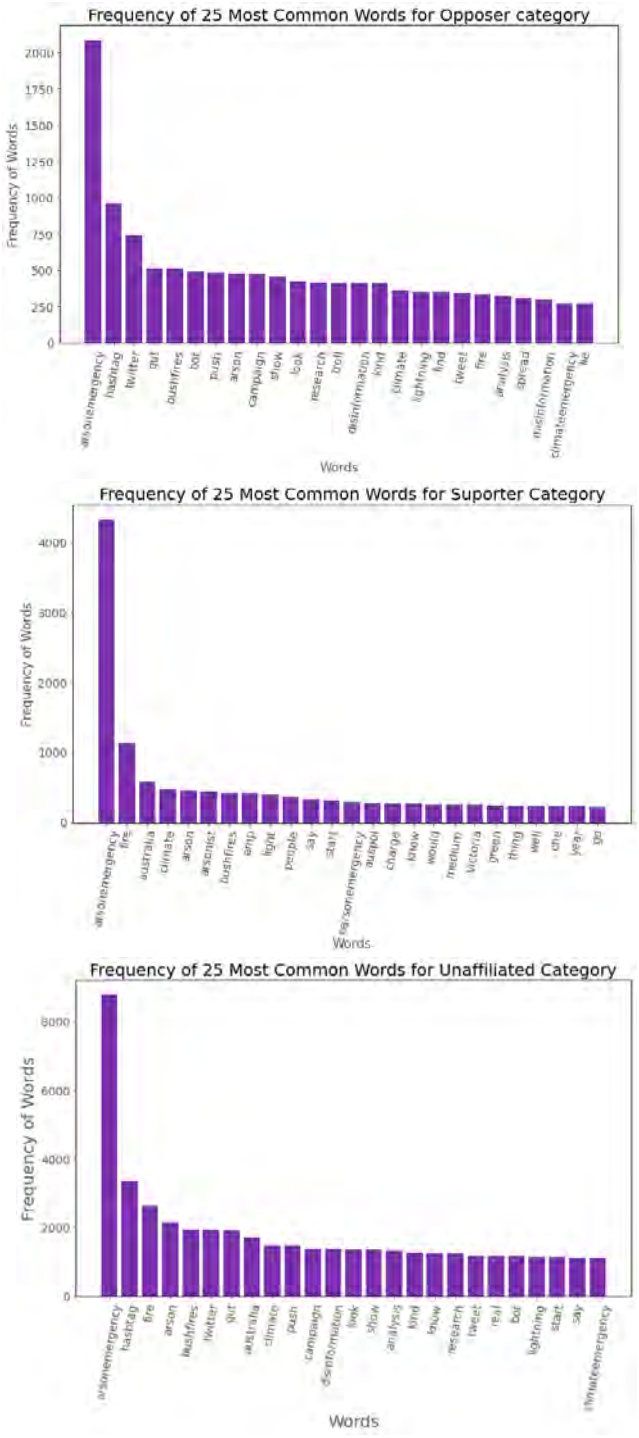

Figure 5: Most Frequently used words in the Opposer, Supporter and Unaffiliated categories. It can be noted that the top frequency words for the supporter and opposer categories are different, but for the unaffiliated category, it is not clearly distinguishable.

be seen that opposing parties follow diverse contexts compared to supporters and base their facts on various topics rather than following a single stream. When analyzing the unaffiliated category, the most common words have different combinations. It contains all the words such as arsonemergency", "disinformation", "arson", and "climateemergency". This shows that the unaffiliated topic is an amalgamation of both the

supporter and opposer categories and has no distinct features. This shows that this domain can introduce significant noise into the training process.

Figure 6: Wordcloud representation of the bushfire data set. The wordcloud generated is based on two of the three categories i.e. the supporters and opposers only. Words with the highest frequency are highlighted in larger fonts.

WordCloud representation has two main benefits in data visualization and exploration. First, wordclouds are visually pleasing and easy to understand, even for people of all domains. The other is that wordclouds represent important and most dominant data from the data set, indicating that they contribute more to context generation. The bigger words represent higher dominance in the corpus. “arsonemergency” and “bushfires” are the most common words, as they were used in the hashtags in most of the documents. In addition to these words, words like “climate”, “arson”, “disinformation”, and ,“climateemergency”. have been used significantly as well. This means that topics such as climate emergencies and arson and arsonists were discussed more frequently. Other words related to the spread of misinformation, trolling, conspiracies, and lightning are commonly found in the corpus.

3.2 The Suspicious tweets data set.

The Suspicious tweets[4] data set is an open-source kaggle data set with two major categories, that is, suspicious and non-suspicious tweets. With the dominance of social networks in the current scenario, it

[4] https://www.kaggle.com/datasets/munkialbright/suspicious-tweets

is exploited with both positive and negative intentions. If used correctly, social networks can be used to educate people in various contexts and spread positive information. They are easy to use and provide an expressive way of communicating. However, the problem arises when social media is used for a variety of illegal and inhuman tasks, such as cyberbullying, terrorism, theft, harassment, and many more. This data set contains a collection of various tweets that are classified as suspicious and non-suspicious.

3.2.1 Exploratory Data Analysis of the Suspicious tweet data set.

Initially, the preprocessing of the data set was carried out. As the corpus is a collection of tweets, people tend to use different contractions and slang words. The contractions were generated to their normal form. Retweet tags, along with the mentions of the account names, were also removed. The corpus was then cleaned by removing digits, punctuation, and emojis. The words were lemmatized to their original form, and the stopwords were removed.

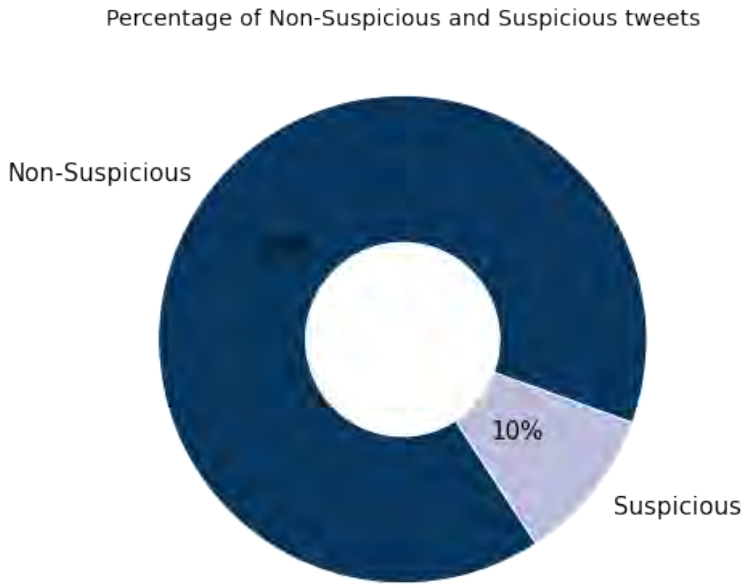

Figure 7: Pie chart determining the suspicious and non suspicious tweets present in the corpus. A moderate imbalance of data is seen in the data set.

Initially, we start with the statistics of the categories present in the corpus. As seen in Figure 7, 10% of the data belong to the suspicious category, while 90% of the data belong to the non-suspicious category. This shows that the dominance of the non-suspicious category is large compared to the threat category. This also implies that there is a moderate imbalance in the data. Therefore, to mitigate the class imbalance, different methods, such as downsampling and upweighting, might be required.

Figure 8 is a histogram showing the sentiment distribution of the suspicious tweet corpus. It is evident that most of the documents in the corpus have a sentiment distribution between -0.10 and +0.10. It can be inferred that most documents have positive and negative sentiments, but they are not extreme. The histogram also shows that more documents range towards positive sentiments than negative ones.

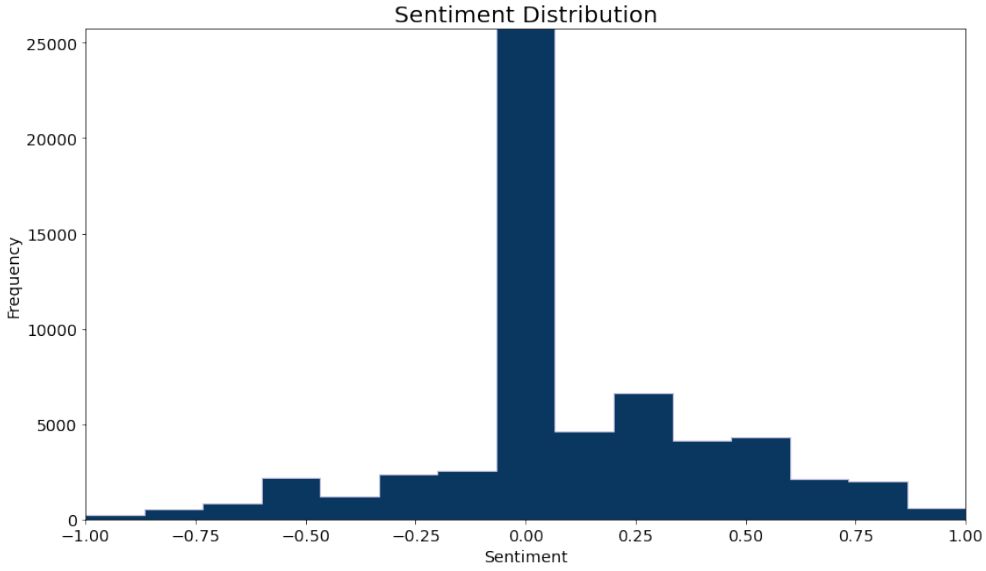

Figure 8: Histogram representing the sentiment distribution across the suspicious tweets corpus.

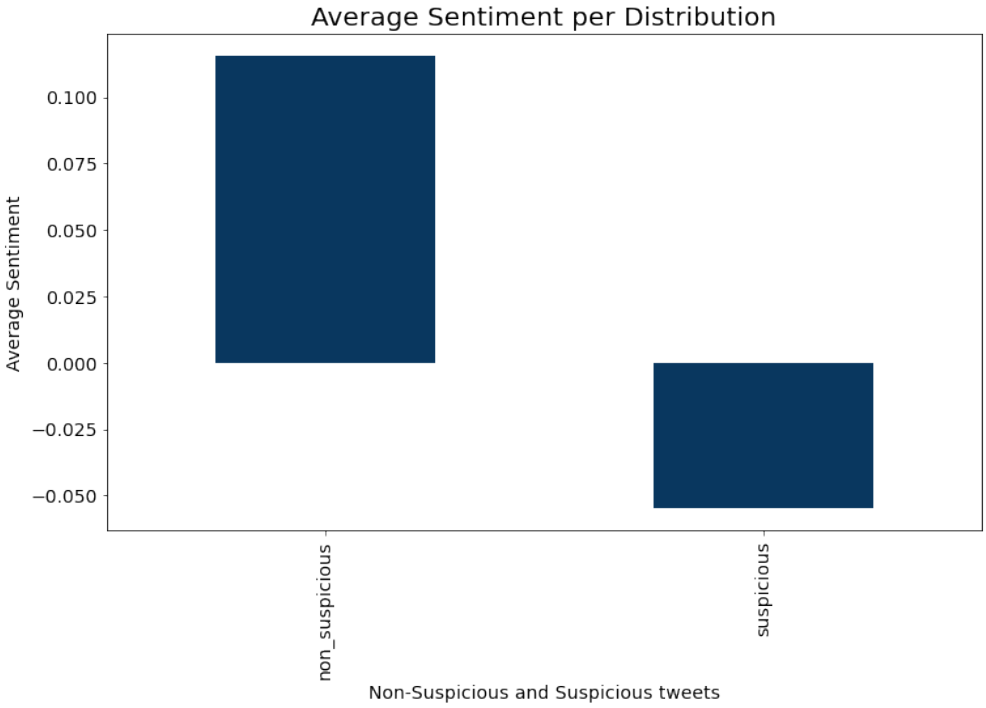

Figure 9: Bar chart representing the average sentiment of the suspicious and non suspicious category.

Fewer documents have extreme negative or positive sentiments. In general, it is observed that most of the documents have mixed sentiments that are slightly divergent from the minimum value. On the contrary, documents with higher values for positive sentiments are greater than those with lower values for positive sentiments.

Figure 9 represents the average sentiment of each category in the corpus. For the non-suspicious category, the average sentiment is positive for the documents. This shows that the non-suspicious tweets are mostly positive in nature. On the other hand, the polarity of suspicious tweets is the opposite. Suspicious tweets

have some level of negative polarity compared to non-suspicious tweets. One conclusion that can be drawn from this observation is that a suspicious tweet is more likely to be of negative sentiment compared to positive sentiment, and a non-suspicious tweet is more likely to be of a positive sentiment. The sentiment is directly proportional to the suspicious and non-suspicious tweets; that is, non-suspicious is positive and suspicious is negative. Compared to the sentiment analysis in each of the categories of the #arsonemergency data set presented in [94], the initial expectation was that supporters of the arson theory would have a negative sentiment, which proved to be wrong, and all categories had a positive sentiment on average. This led to the conclusion that polarized groups in a discussion do not need to have opposite sentiments.

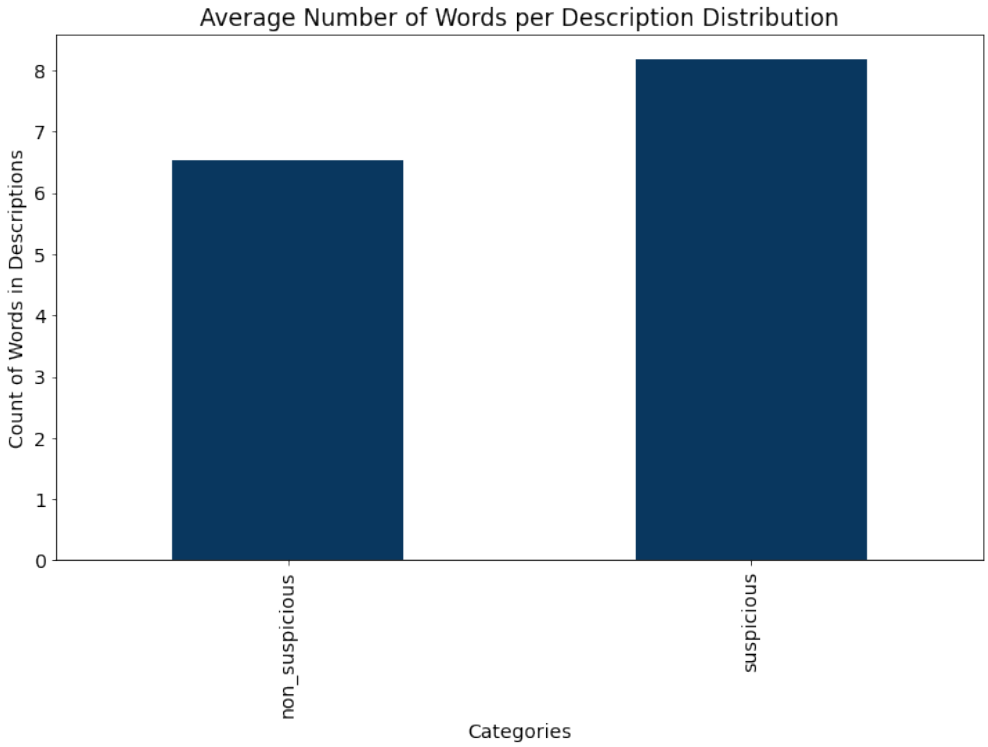

Figure 10: Average word counts in each category of the Suspicious tweets data set.

Figure 10 shows the average word count for each document present in the corpus. This bar chart that describes the word count can be used to see the length of sentences used by different people in various scenarios. The people who tweet suspicious tweets are more descriptive than non-suspicious tweets. Therefore, it can be inferred that suspicious tweets are generally longer than non-suspicious tweets. Although this feature cannot be used alone to distinguish suspicious from non-suspicious tweets, this information can be used as a building block of a larger system.

The top frequency words chart presented in Figure 11 describes the words most commonly used in each category of the corpus. In the non-suspicious category, the commonly used words to be noticed are words such as “good”, “day”, “work”, and “love”. It shows that people mostly use normal words but with a more positive attitude in the case of non-suspicious tweets. In the case of suspicious tweets, the most common words used are “hate”, “sick”, “work”, and “problem”. These words mostly denote negative feelings. Although some differences can be seen in the words for the suspicious and non-suspicious categories,

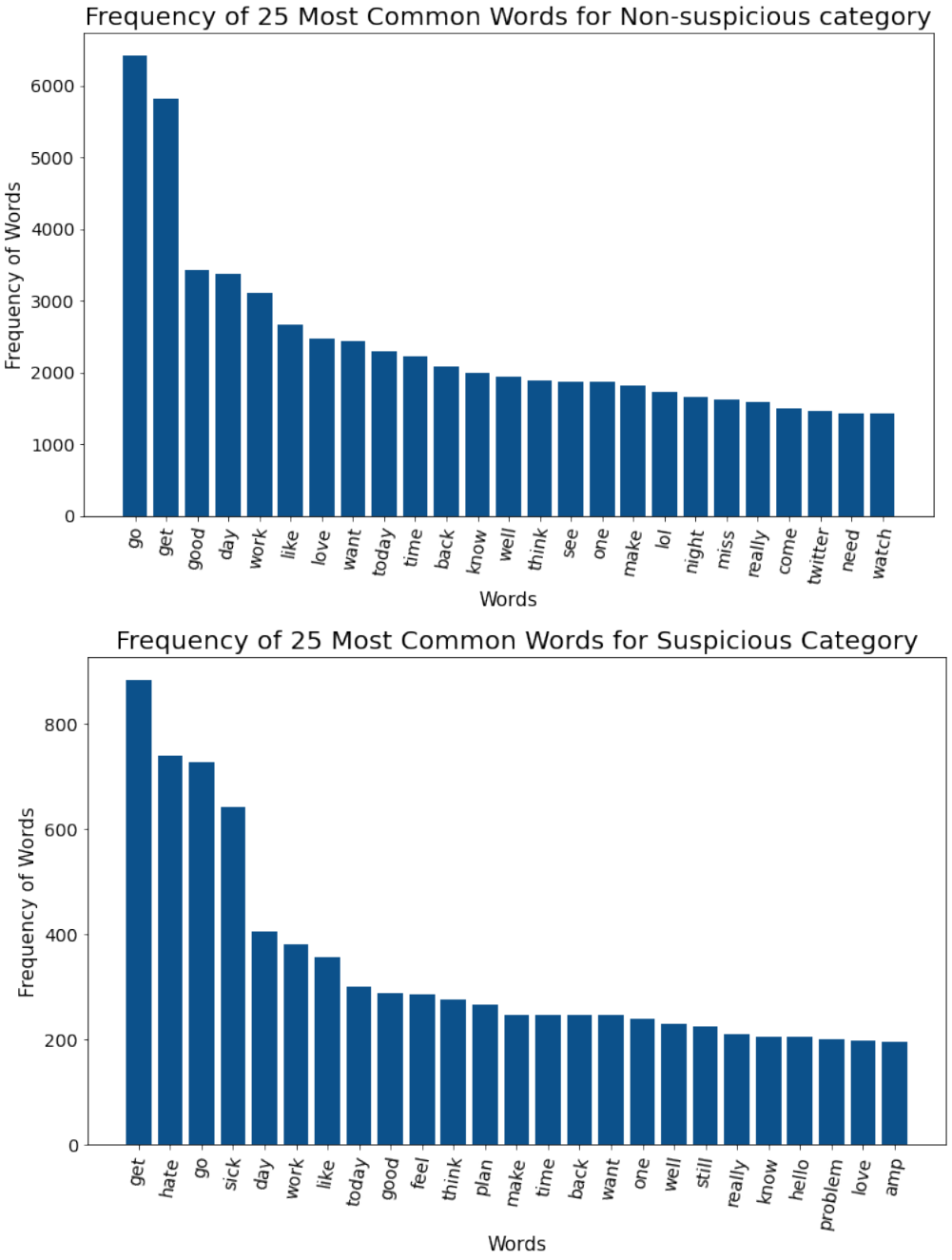

Figure 11: Most Frequently used words in the non-suspicious and suspicious categories. Although a clear distinction of words cannot be seen in the context of both categories, there are unique words and contexts described in each category.

those differences are subtle. One reason for this might be the context of the database on which it was built. Whenever a suspicious tweet is sent, the intention might still be to keep it simple and mundane so that it is not possible to distinguish it from normal tweets. Therefore, the words/unigrams used are common and the only fine-drawn differences that can be seen are in the cases of positive and negative words.

The word cloud presented in figure 12 represents all the common words used in the corpus and their frequency. Words like “good”, “love”, “work”, and “day” are the most frequently used words in the corpus.

Figure 12: Wordcloud representation of the Suspicious tweets data set. The wordcloud generated is based on two categories, i.e. suspicious and non-suspicious tweets. Words with the highest frequency are highlighted in larger fonts.

The absence of unique words related to various domains indicates that discussions or tweets might not be based on a particular topic like in the previous data set. This also means that specifying a particular topic and topic modeling can be difficult in such a scenario.

3.3 The Threat corpus

Hammer et al [43] presents the Threat corpus that contains a large number of documents that are classified as threat or non-threat comments. These sentences are derived from over 10,000 comments and contain more than 30000 documents in the corpus, which are *non-identifiable*. These documents are manually annotated by human observation to represent whether the comment is a violent threat or not. Expressing violent thoughts and influencing other people for the same is a significant problem today[19, 35]. [43] address violent comments as those that are life-threatening and impose security concerns on a particular person or a group of people.

One of the solutions that have been proposed and implemented to contain this type of threat and restrict mishaps is the use of moderators [43]. Various social media platforms use moderators to filter comments or posts and redirect them to original users with the possibility of reformulating these posts. But this process is excessively time-consuming and practically impossible to process all these posts by humans alone or some manual interference. Therefore, the need for an automated system is seen. Scientific research like [2, 3] suggests that the use of various machine learning features is an important part of this kind of filtering and provides effective solutions to these problems. However, in many contemporary scenarios, the problem persists where these data sets are either not present in an adequate amount or are not publicly available. Therefore, this article presents comments on various types of diversity, political and religious [43, 95] . The data set contained a total of 9,845 comments that had 28,643 sentences associated with it. The total number of users sampled was 5,484. Of these total figures, 1,287 comments were classified as consistent

threats and the sentence count for the same category was 1,387. Similarly, 993 users out of 5,484 imposed threat comments according to the collected data.

Words and figures related to the data set are either omitted or masked when used because it might contain cultural or religious inclusions that lie outside the scope of the thesis.

3.3.1 Exploratory Data Analysis on the Threat corpus

The text format of the data obtained was initially managed in the CSV format for further processing. The corpus collected is directly extracted from the comments of YouTube videos, and hence contains various irregularities and inconsistencies. With the help of various libraries and packages, the initial prepossessing was completed in python. This step includes the removal of punctuation, emojis, video tags, user tags, stop words, contractions, and other forms of noise [5].

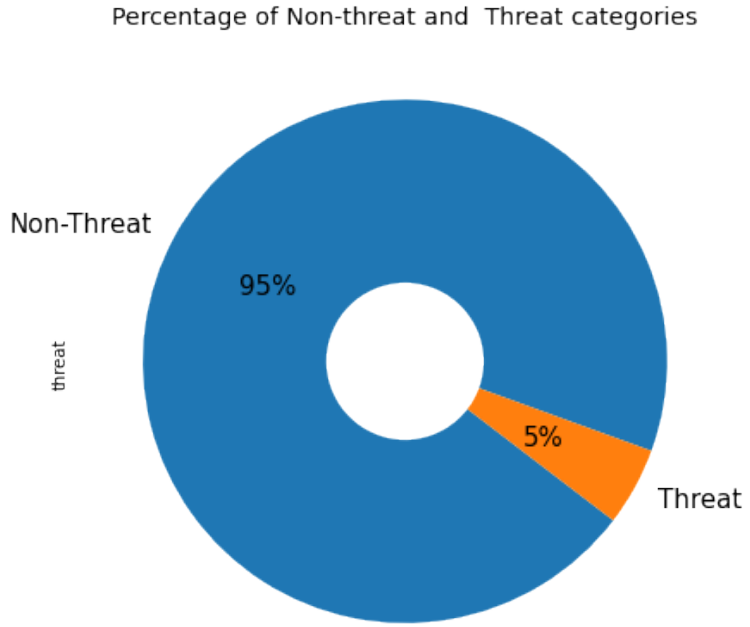

Figure 13: Percentage and Threat and Non-threat comments in the corpus.

The Figure 13 represents that only 5% of the total documents in the corpus are threats. This shows that a small percentage of people pass the threat comments compared to the majority of people. This imbalance of data suggests the need to adapt various methods while building models for classification.

Figure 14 represents the histogram of the sentiment distribution between the documents in the corpus. The sentiments of most tweets range from -0.3 to +0.3. Most tweets are around the 0-range, suggesting a weak resemblance to both positive and negative sentiments. However, in a further evaluation, many documents fall into moderate to extreme levels of both positive and negative sentiments. Extremes in negative sentiments (ranging from 0.75 to more) are slightly more than positive sentiment. In general, negative sentiments seem to dominate according to expectations.

[5] `https://www.nltk.org/`

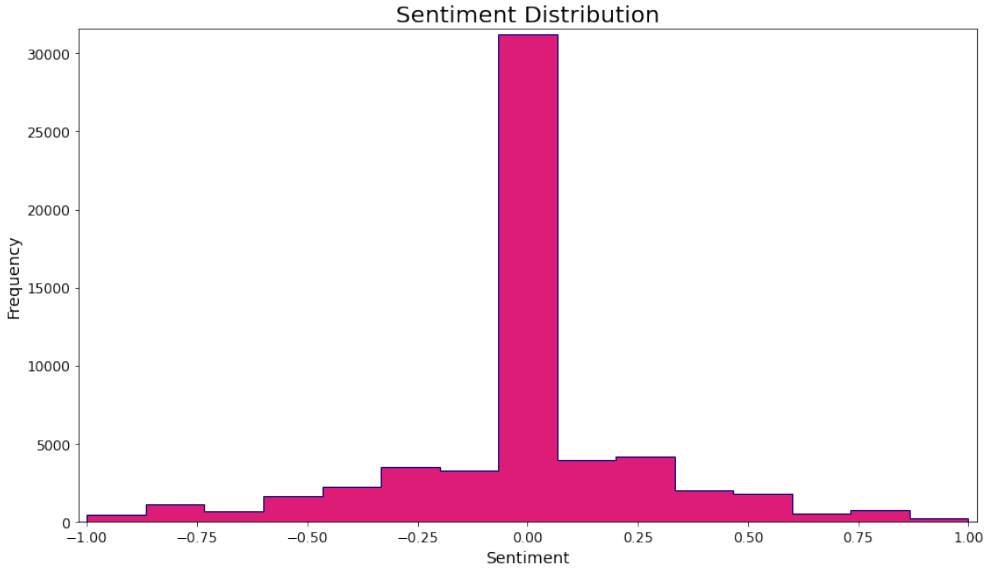

Figure 14: Histogram showing the distribution of sentiment across the corpus.

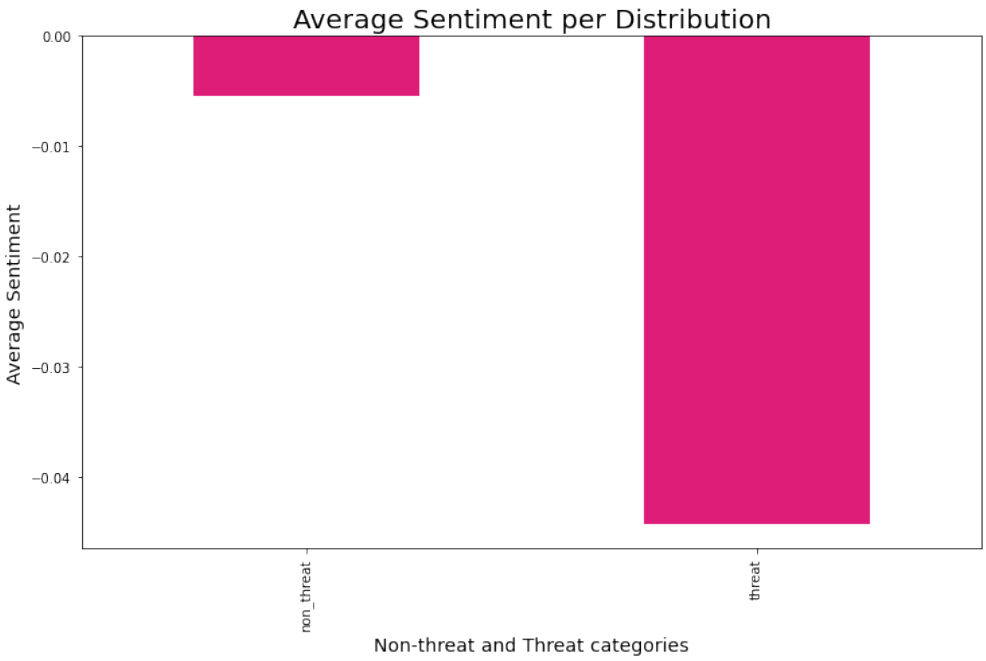

Figure 15: Bar chart depicting the average sentiment of the threat and non-threat categories.

The average sentiment for each threat and non threat category were computed. Although there is a significant difference in average sentiment scores between the two categories, both have negative scores on average. The threat category has a higher negative score, as expected, because the sentences in the category frequently use negative and threatening words. The computation of scores is negative for the non-threat category because even if the sentences are nonthreatening and might not impose security issues, the arguments are presented in a negative context, and there exists a high discrepancy in sensitive matters such as religion, race, and culture.

Average word count in each of the categories of the Threat data set was calculated. The non-threat

category has a lower average word count than the threat category. This represents that on average one uses more words to write a threat than non-threat comments. Compared to the previous suspicious data set, where the average word count for suspicious tweets was higher, it can be made analogous that longer sentences are present to denote arguments that are usually negative.

Frequently used words were computed for both the threat and non-threat categories. Although many of the words observed were common in both categories, the difference is in frequency and use of words. In the threat category, the most commonly used words are those that indicate life-threatening words, swear words, religion, discriminating words, hate words, and various negative sentiments. Among these, the frequency of use of words related to life threats is the highest. The non-threat category emphasizes the use of words related to religion, cultures, and a community of people more compared to threatening words. These words, when used, are intended mainly in a different context.

Wordcloud was also computed for the corpus. The presence of high discrimination, hate, bully, severe threats, targets on a particular group of people was seen from the wordcloud representation. It was clearly evident that the data set is a collection of sentences with debates, hates, and disrespect between various people and groups in different topic scenarios.

4 Topic Modeling

Initially, the Latent Dirichlet Allocation and Linear Discriminant Analysis was performed with the aid of python's *scikit-learn* [6] library. Both of these methods did not produce satisfactory results, and hence the clustering method in combination with PCA was used as described in the literature review. Various reasons for the failure of these methods can be identified. Agrawal et al. [4] mention that even small changes such as shuffling the training data can produce inefficient results. Furthermore, the lack of context, shorter sentences, or even variations in sentences could produce inefficient results for topic modeling. Various supervised methods also fail when the datasets are large and heavy-tailed vocabularies are used [30] which is also the case in the datasets used in the thesis. Upon analyzing various supervised and unsupervised methods for topic modeling as presented in different papers such as [4, 30, 55, 59] it was concluded that topic modeling is highly content-specific and a fixed method that fits the whole scenario is not yet devised. It was also evident that various techniques were still continuously considered and explored. Given the promising results of the K-means clustering in combination with other topic modeling methods like PCA, this thesis is based on the same technique.

4.1 Cluster Analysis

Clustering helps visualize similar groups with a similar motive or intention together and also to find outliers present in the data [65]. The main goal is to obtain a group of data that is ideally distinguishable from another group [70, 65, 53] where the elements of a group share similar properties, but the different groups have their own unique characteristics. Clustering is an unsupervised technique and is therefore considered challenging compared to simple classification techniques. Various aspects of deep learning are used for this purpose in different areas. This is done primarily for pattern recognition in the case of image processing, data analysis, and many other reasons [53]. Although various techniques have been explored for the extraction of knowledge and have produced effective results for formal languages in the case of NLP, the scarcity of formal procedures in these sectors remains [29]. As discussed in the literature review section, apart from regular topic modeling techniques such as linear discriminant analysis, latent Dirichlet allocation, and principal component analysis (PCA), clustering methods in combination with these techniques provide competitive results in topic modeling. A combination of PCA and K means clustering was applied to the present data sets.

The clustering scenario of NLP is as follows. Initially, the input data is an amalgamation of different entities that are present in the corpus in our context. As proposed in the EDA section, all of these corpus and data are subjected to their respective cleaning and prepossessing. Furthermore, a clustering algorithm is applied that provides the respective clusters as output. Clustering is a technique that consists of several variations and algorithms according to the context for which it is applied. The two main clustering metrics are usually dominant, namely distance metrics and similarity metrics, and the use of a specific type of

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html`
`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html`

technique is defined by the application domain and the type of data set [65]. Oyelade et al. [70] present in their paper that with the increase in the volume of information in all sectors in the form of text, audio, and visuals, clustering is important not only for the classification of these words but also for understanding the perception of the data and its underlying meanings, and it can also be used to manage and summarize the data.

K-Means clustering is a partition clustering technique. To iteratively improve the quality of partitions, the objective function is optimized [70]. The optimization of the objective function is given as [65]:

$$E = \sum_{1}^{k} \; | \; |X_i - m_i| \; |^2 \tag{4}$$

In cluster C, X_i is one of the points present in the cluster. For a particular cluster K_i, m_i is the mean of the cluster. The objective is to reduce the value of E in the clusters. If we assume that the number K of clusters is to be obtained from a certain corpus, E in the expression above (equation 4) represents the sum of squared error. This error comprehends all the data because it is the sum of the average Euclidean distance of the centroid and each data point. We initialize with K=1 and compute E until the value added by the user is reached.

Clustering of K means was performed according to Algorithms 1 and 2 (as discussed in the Literature review and given below) for all data sets.

Algorithm 1 :- K means clustering.

1. Initially, select the number of K clusters and initialize the centroid.
2. Clusters of K numbers are formed by adjusting all data points to the centroid that are the closest.
3. Recalculate the centroid for each of the formed clusters.
4. Repeat steps 2 and 3 until the centroid does not change.

The main goal is to optimize the number of clusters supported by the elbow method whose algorithm is given below.

Algorithm 2 :- Determining the elbow points.

1. Initialize the value K = 1.
2. Start.
3. Increase the value of K by 1 in each iteration.
4. Calculate the sum of squared errors according to Equation E mentioned in 4. This is also known as the cost of the optimal solution.
5. Observe a point where the cost or error changes drastically.
6. This is the point that shows the true value of K.
7. End.

The methodology was followed as described in Figure 16. The documents were initially cleaned and pre-processed as described in the exploratory data analysis section. In addition to processes such as document noise removal and tokenization, term weighting and principal component analysis (PCA) were used to improve performance [13]. After completion of the preprocessing, the documents are represented in N-dimensional vector spaces. This is done using the term Frequency-Inverse Document Frequency (TF-IDF) term. This is used because this vectorizer considers the relevance and importance of the terms used in the documents. In the N-dimensional vector space, N refers to the number of terms or words. TF-IDF is calculated as:

$$TF - IDF = TF * IDF$$

TF is given by the total number of words present in the document d by the total number of words in the document d. IDF is given by the total number of documents present in the data set by the total number of documents that contain a particular word [81].

PCA is a dimensionality reduction technique that takes data from several ratios and reduces them into smaller indexes. These indices ensure the originality of all initial ratios [80, 48]. The elbow method is then used according to the previous methodology, which helps to determine the optimum number of clusters. This iterative method is followed according to Figure 16.

The cost of solutions for all data sets is presented in Table 2,. As per our objective, we intend to find the point on the elbow where the error is minimum. This helps to find the optimal number of K for each data set. For the #arsonemergency data set, presented in Table 2 it is observable that there is a drastic decrease in cost from K = 2 to K = 3 and the cost gradually decreases later. Therefore, according to the assumptions made previously, n = 3 is the elbow point where the optimal number of groups can be achieved. In the case of the Suspicious tweet data set, the scenario is similar as well. From K = 2 to K = 3 the drop in cost is large and steady after the point. Therefore, the elbow point is also at K = 3 for this data set. Finally, a similar procedure was also followed for the Threat data set. From the data in Table 2 it can be seen that the cost decreases from 649.4 to 327.4 from N = 2 to 3 which is a significant decrease in value. The cost decreases from 327.4 to 241.2, 174.9, and 141.7, respectively, which are constant smaller changes

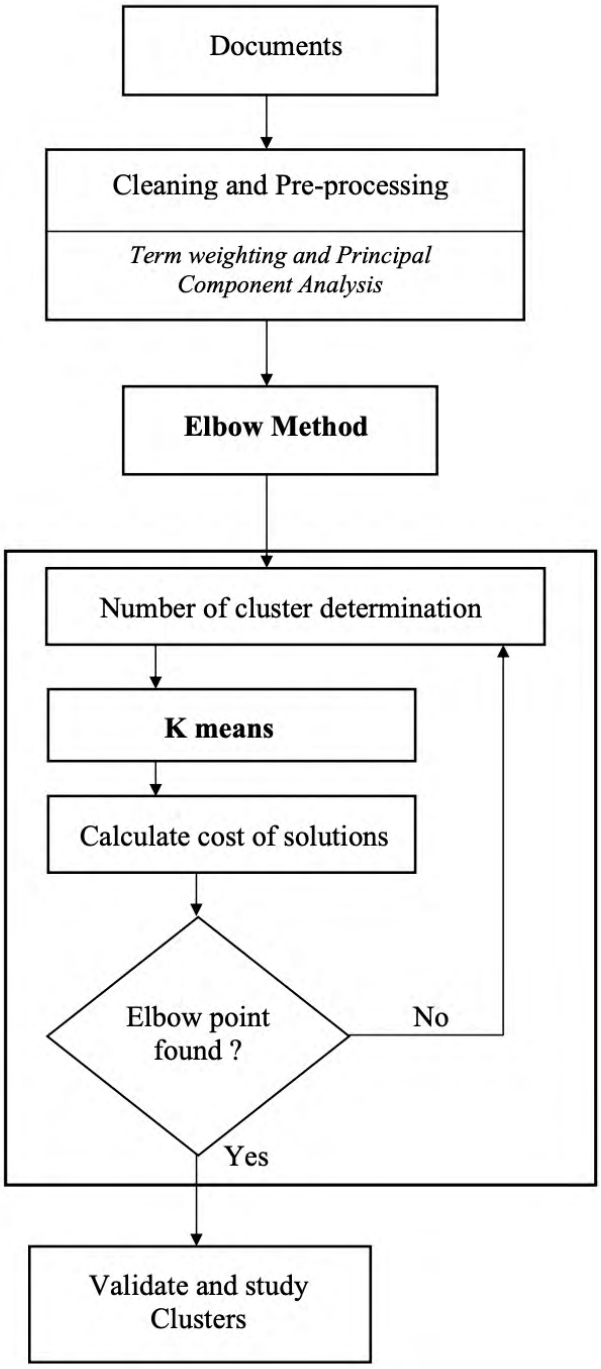

Figure 16: Flow chart of the followed methodology for K means clustering.

based on increasing K values. Therefore, the optimal value of K is also 3. In conclusion, the optimal value of K is 3 for all data sets.

Cost of solutions			
Number of Clusters(K)	#arsonemergency	Suspicious tweets	Threat
1	539.8	1056.2	1118.3
2	288.0	639.6	649.4
3	**119.1**	**307.5**	**327.4**
4	88.3	233.9	241.2
5	58.6	169.3	174.9
6	44.2	137.5	141.7

Table 2: The cost of solutions for each value of K is presented in the above table for all the data sets mentioned in the thesis. The cost of solutions is different for each data set and value of K from which the 'elbow' point is to be determined.

Number of elements			
Cluster Number(K)	#arsonemergency	Suspicious tweets	Threat
1	9,867	52,789	49,790
2	466	3,080	4,352
3	225	4,131	3,142

Table 3: The number of elements present in each cluster for all data sets.

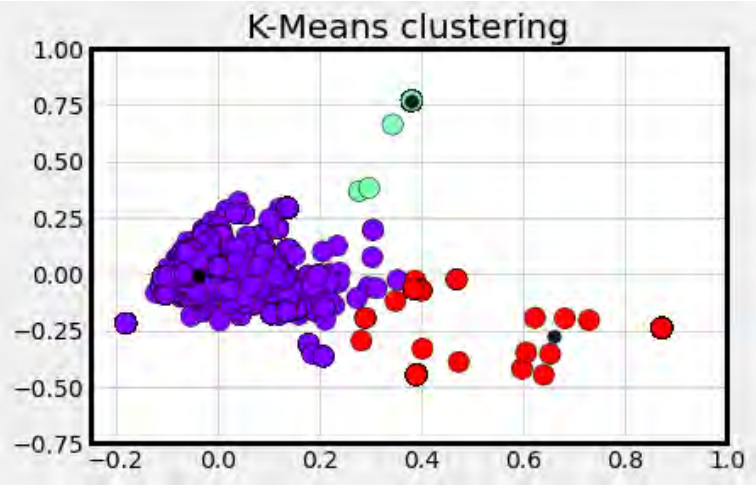

Figure 17: Clusters of data obtained from the #arsonemergency data set.

4.1.1 K-means Clustering on #arsonemergency data set

Figure 17 represents the 3 different clusters that are formed according to the K means clustering algorithm. Table 3 represents the data obtained from the clustering, which shows that most of the elements are in the first group, that is, 9,867. The second and third clusters contain smaller sets of datapoints, that is, 466 and 255. It can be seen that the data points in the first cluster are tightly spaced around the centroid (represented in black), indicating that these points have similar themes and representations. The second

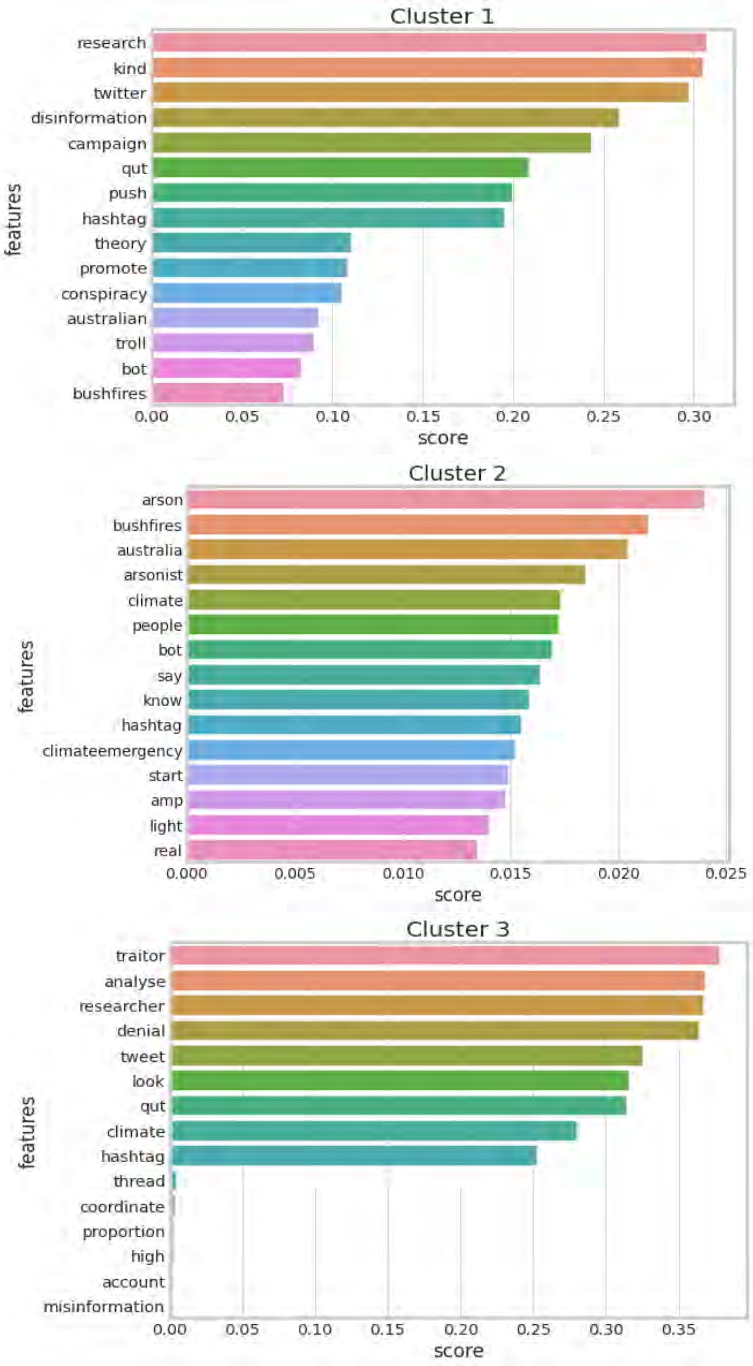

Figure 18: Figures representing the most important features in each cluster as measured by TF-IDF for the #arsonemergency data set. The figures from left to right represent Clusters 0, 1 and 2 respectively.

and third clusters are scattered and contain relatively low data points.

Once the clusters have formed, the most significant or dominant words were extracted from each group, which can be visualized in the Figure 18. The top 15 words are presented along with their scores, and each group carries a different theme. In the analysis, it is observable that each of the clusters contains unique words. Upon analyzing each of the clusters and their dominant words, we can draw some inferences. The first group talks about *disinformation, conspiracy, trolling, bots, promotions, campaigns, and bushfires*. This

indicates that this cluster somehow represents the opposer category, i.e. the views of the opposers of arson theory. However, the second group talks about *people, arson, arsonists, lighting fires*. This mainly supports the view of the supporter category for arson theory. Although a specific topic for the third cluster cannot be exactly determined, it is still observable that it talks about traits, denial, research, and analysis, which denotes that there is a knowledge gap about the topic that needs to be investigated. The main advantage and goal of forming clusters is that these clusters represent data that have similar characteristics. This also helps to dive deeper into one specific cluster and explore more of its features.

4.1.2 K-means Clustering on Suspicious tweet data set

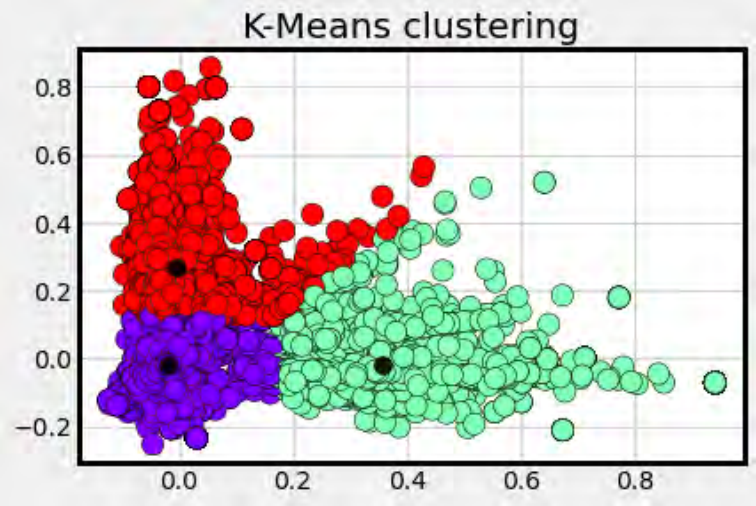

Figure 19: Clusters of data obtained from the Suspicious tweet data set.

Figure 19 represents the clustering of K means for the Suspicious tweet data set. As determined by *elbow method*, there are three different groups represented in different colors. Referencing to Table 3, the first, second, and third groups contain 52789, 3080, and 4131 data points, respectively. The first group is significantly larger in number compared to the second and third groups. It is evident that the clusters are distinct and are mostly tightly spaced around their centroids, but the boundaries are close. Outliers can be observed in the figures that determine the K-means clusters. This is because the mean value is sensitive to outliers and can easily affect the overall value.

Similarly to the previous data set, the top words are also calculated for each cluster. The dominant top words and their scores are represented in figure 20. In the analysis of each of the figures, from the first cluster the dominant words present are *good, feel, look, luck, time, day, morning, and night*. It can be inferred that the theme of this group is mostly positive. Conversations revolve around good feelings and positive vibes. The second group presents words that are relevant to *work, such as hour, work, and week*. Although the group did not have higher scores for each of these words, the collective analysis of these words represents the theme related to the work. The third group represents words that are relevant to love and time, but a specific theme cannot be determined. However, the sentiment is generally positive in the group.

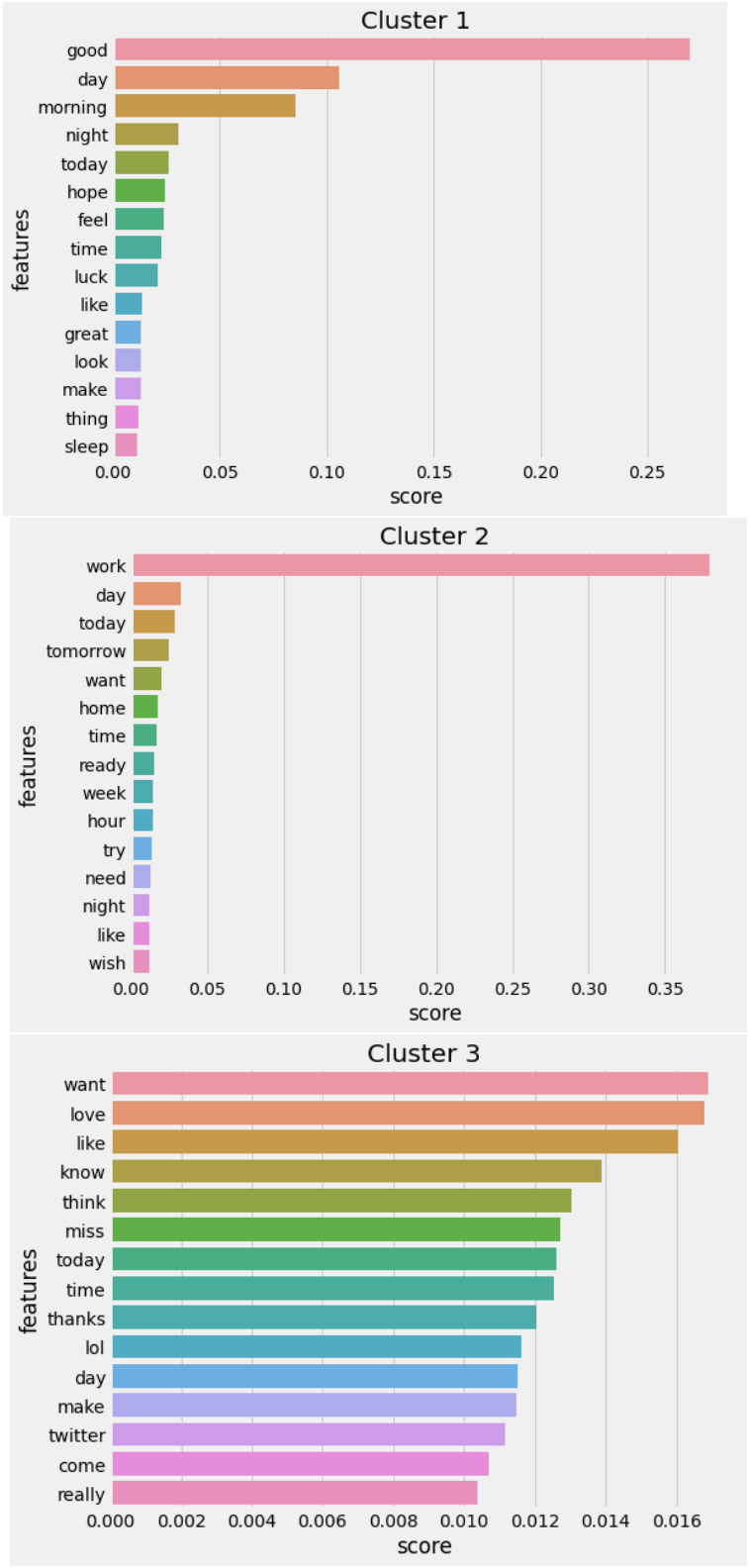

Figure 20: Figures representing the most important features in each cluster as measured by TF-IDF for the Suspicious data set. The figures from left to right represent Clusters 0, 1 and 2 respectively.

4.1.3 K-means Clustering on Threat data set

K means that the clustering for the Threat data set is present in Figure 21. After the calculation of the loss and the analysis of the data in Table 2 the optimal number of groups was found to be 3. Therefore,

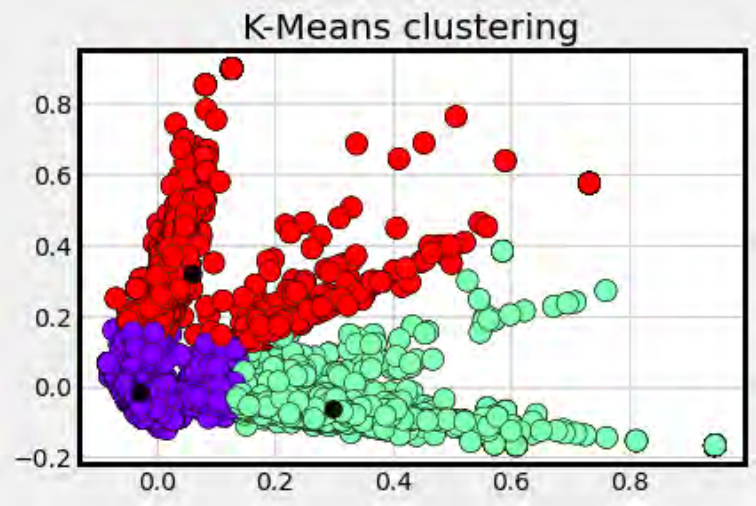

Figure 21: Clusters of data obtained from the Threat data set.

the value used for K was 3. These 3 clusters are represented in different colors. Table 3 shows that there are 49790, 4352, and 3142 data points in groups 1, 2, and 3, respectively. From the sentiment analysis section, we noticed that the data set had highly negative sentiments for both threat and non-threat cases. Although there are threat and non-threat comments in the corpus, the debate includes feuds, conflicts, and arguments relating to various sensitive topics like *religion, group of people, country, gender, race.* Therefore, these groups are expected to have negative feelings.

Dominant words of three groups formed by applying the grouping of the K means were obtained. When analyzing each of the clusters, it can was seen that the first cluster talks about things like *religion, god, country.* The second group also talks about religions and countries but promotes more hate compared to the first group. Words such as terrorists are used to promote violence and negative emotions. The third group contains the highest use of the word violent threat. This indicates that this group consists of documents that contain a large violent threat. One common property that all these groups share is negative sentiment and hate speech. But the difference lies in the fact that some clusters are violent and contain a threat.

5 Moral Valence of Tweets

A high-level understanding of the tweets and the values it promotes are given by the five categories of moral foundations described in Table 1. To quantify the framing bias in the news, the backbone is the Moral Foundation Theory similar to [61, 72]. *Bias and Intensity* is calculated for every dictionary in the corpus towards the moral foundation axis. The relevance of a document to the moral foundation category is given by the term *bias*. If a document has a negative bias value, then it more depicts vice dimensions, whereas the virtue dimension is shown if the bias value is positive. The relevance of the tweet to each moral foundation is identified by the term *intensity* [41]. The average activation of the moral foundation was taken for each of the data sets to observe their influence in each of the five domains. A specific pattern is followed for the three data sets considered in this thesis. First, the mean was computed for the whole document, i.e. considering all the categories that are present in it. The document was then classified into its unique labels, and then the mean activation scores for them were found separately. For example, in the case of the Suspicious tweet data set, the MDF score average of all the documents were derived first, and then the average scores for each of the categories, namely suspicious and non-suspicious, were computed later. This was done because the contribution of each of the categories could be seen and analyzed.

5.1 Method applied

To map the moral valence of documents in the corpus, the FrameAxis [51] method is used, as discussed in the literature review (Section 2.3 and subsections). Furthermore, to map scores the extended form of the MFD [37] is used. The following section describes the figures and results obtained.

5.2 Results analysis from the FrameAxis framework.

The dictionary-based approach provides mappings of moral valence that are easier to understand and interpret. The following section presents the bar graphs that present the mean scores for each of the data sets and its individual categories. In addition, box plots are presented that represent each of the individual categories in the datasets. These boxes represent the figures of the model coefficients with 0.95 confidence interval. The bar plots represent the vice and virtue domain of each of the moral foundations whereas the boxplots represents the *bias and intensity* for the same.

Figure 22 shows the average moral scores for the #arsonemergency data set. This shows that on average all the moral foundations care, authority, fairness, loyalty, and sanctity have a significant contribution in the corpus. The vice domain is dominant, referring to the words present that do not represent high moral standards in conversations. The reason for this could be the discrepancy in the thinking of the opposing and supporting community. It can also be observed that people do not necessarily make use of or do not consider the obligations to use moral words while convincing or imposing their thought on someone. Although the virtue domain score for the care group is higher than others, it is still significantly behind compared to the vice score in the same group. Words such as damage, destroy, disease, lie, trait, disagree, and many other vice words are commonly used in the corpus, which justifies these scores. In the individual analysis

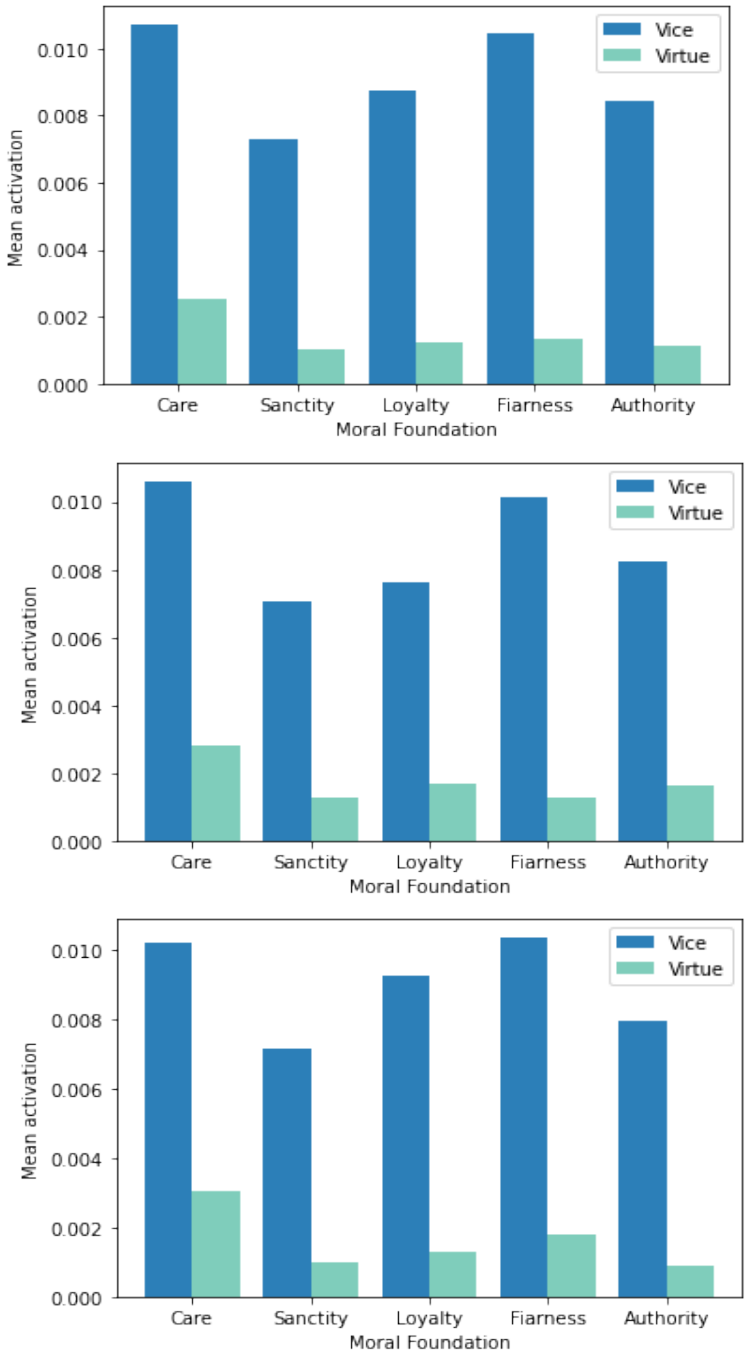

Figure 22: Mean activation scores for the #arsonemergency data set. The figures from top to bottom denote the scores for 1) Overall Documents. 2) The supporter category 3) The opposer category respectively.

of each of the categories, the vice domain is dominant in both cases. Both categories express care in the virtue domain, but the scores are relatively low. This suggests that these tweets and conversations tend to promote a negative attitude, regardless of whether the group is in favor or against the theory of arson. Bias and intensity scores are presented in Figure 23. Negative bias values support the outcome of the bar graphs where the vice domain is prominent. The values of intensity are comparable for each of supporter and opposer categories, determining that both of the parties use moral words but there is insufficient evidence

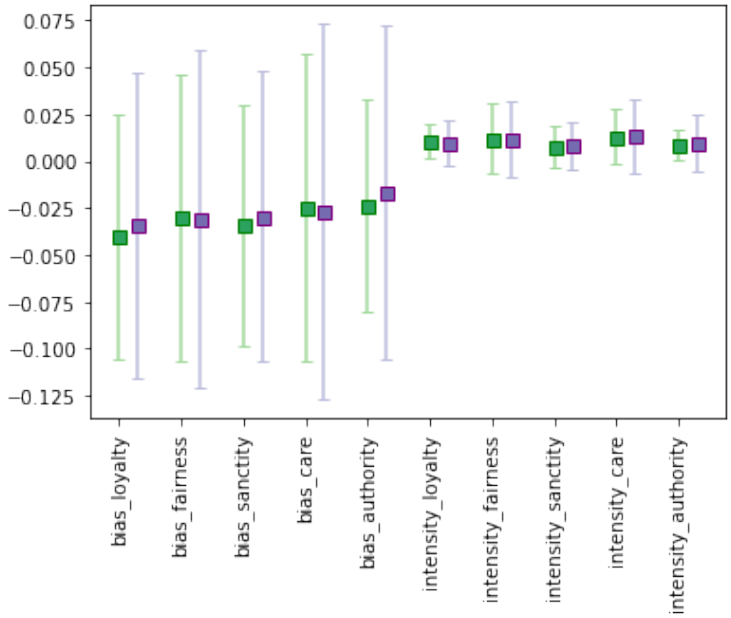

Figure 23: Bias and Intensity calculations for the #arsonemergency data set. The *green plots determine the bias and intensity for the opposer category* whereas *the purple plot determine the bias and intensity for the supporter category.*

to distinguish the two parties.

For the Suspicious tweet data set, as presented in Figure 24, the most active moral foundations are the loyalty and care sectors in the virtue domain. This corpus is not based around a specific domain or topic, and hence the scores for moral foundations are diverse. Although the virtue domain precedes all other categories of moral foundations, the fairness category is ruled by the vice domain. The vice domain is in charge of the corpus in general because the use of words like *good, love, like, care, and family* is commonly used in the corpus. For fairness, it can be predicted that the corpus contains a collection of words with low moral standards such as *bias, dishonest, unfair, and unjust.* On individual analysis of each of the non-suspicious and suspicious categories, the difference in vice and virtue domains can be clearly seen. The tweets of the non-suspicious categories are of virtuous nature mostly whereas suspicious tweets are dominated by the vice domain. However, it is not rigid that only one domain precedes in both the suspicious and non-suspicious categories. For the suspicious category, words such as *kill, hate, illegal* are more commonly used words, and therefore the vice domain is dominant for purity, fairness and authority. However, the presence of virtuousness is also significant (especially in care and loyalty) because positive words such as *peace, empathy, family, community.* are frequently used. For the non-suspicious category, the virtue domain leads because of the use of words with higher moral strength.The first figure is analogous to the non-suspicious category because of a significant imbalance in the data where the suspicious tweets are less populated compared to the non-suspicious tweets in the corpus. In general, it can be concluded that the corpus consists of a mixture of positive and negative contexts. Upon analysis of the preset box plots in Figure 27, it can be inferred that the non-suspicious tweets advocate the virtue side of loyalty and care. Fairness and authority are dominated by the vice domain that aligns with the respective bar graphs, as

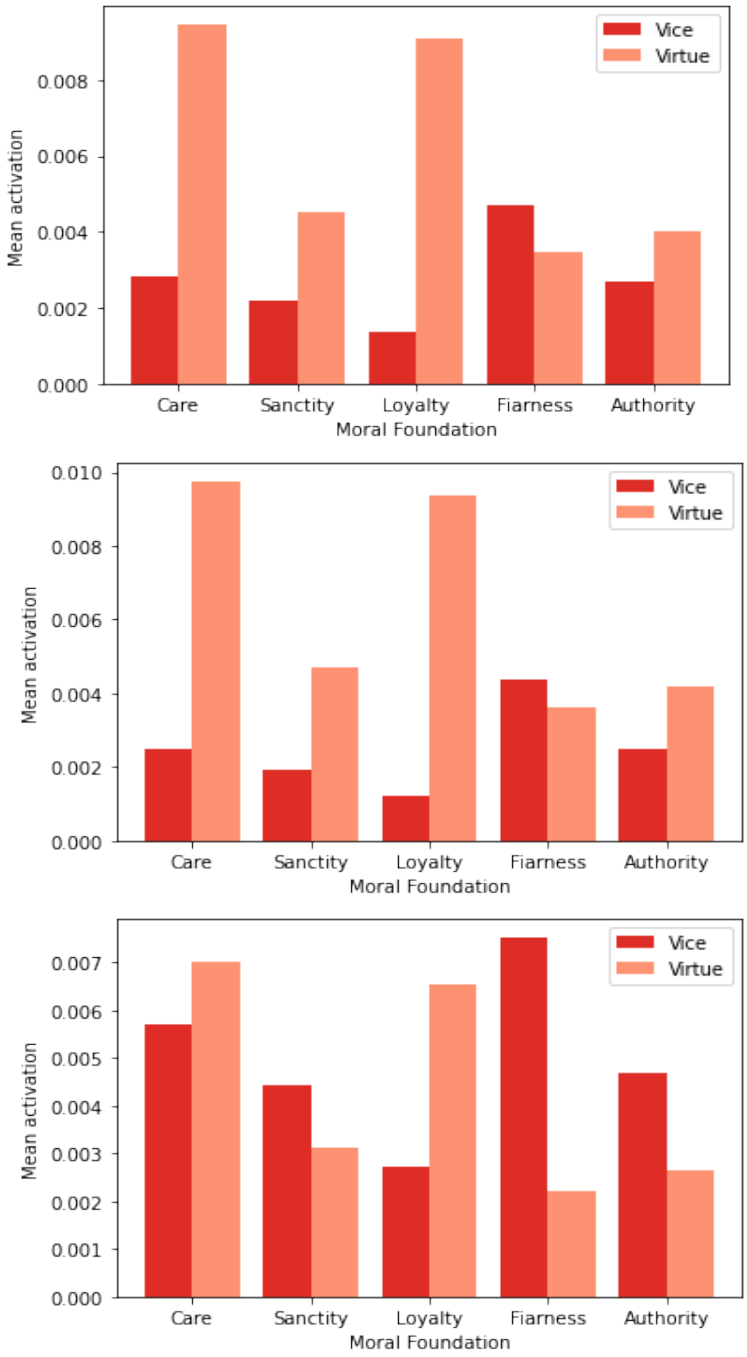

Figure 24: Mean activation scores for the Suspicious tweet data set. The figures from top to bottom indicate the scores for the 1) All documents in the corpus (suspicious and non suspicious tweets), 2) Non-suspicious category 3) Suspicious category respectively.

shown in Figure 24. The intensity scores on the other hand are comparable for each of the categories and range near 0 referring to the fact that the use of moral words is not strong in each of the categories.

Figure 26 contains 3 bar charts that depict the moral foundation scores for the Threat data set. Initially, by analyzing the average sentiment scores and the frequently used words in, it can be predicted that in either of the threat and non-threat categories, words with low morality and discipline have been frequently used.

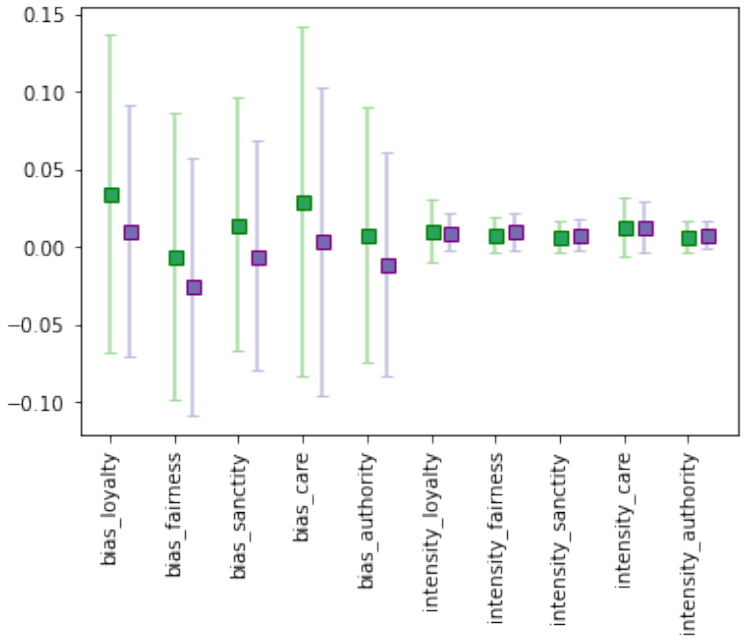

Figure 25: Bias and Intensity calculations for the Suspicious tweet data set. The *green plots determine the bias and intensity for the non-suspicious tweets* whereas the *purple plot determine the bias and intensity for the suspicious tweets.*

This leads to the initial assumption that the scores obtained will demonstrate a vice nature rather than a virtuous one. Similarly to the Suspicious tweet data set, the data imbalance is significant for the two different groups. The first figure denotes the overall moral foundation scores where the vice domain is clearly in command. However, the presence of virtuousness is also significant in the overall scenario. In individual analysis, the second figure is highly dominated and has higher scores in the vice domain. As mentioned above, this is the result of the use of several hate words, disrespectful speech, discrimination based on country, gender, and culture, and many other immoral words. In the context of the classification of these texts, it can be clearly concluded that most threats carry a high degree of immoral behavior and negative feelings. The difference between the moral scores of the threat and non-threat categories is that although the vice domain is prevalent in the non-threat category similar to the threat category, the presence of virtuous nature is significant compared to the threat category where the virtuous domain is lacking. This is a fine line that can be seen between the two categories. The results are justifiable because, in the non-threat category, the tweets do not only express negative sentiments, but also the feeling that they should maintain the peace and harmony, and hence defy hate and violence for a better living. Words such as *god, faith, love, give, hope, life.* are commonly used, which is a symbol of passivity. Overall, it can be concluded that the Threat corpus is overruled by the vice domain in the threat category, and although it is also true for the non-threat category, the presence of virtuous domain is significant as well. The box plots representing the relevant coefficient values for the threat and non-threat categories are represented in figure 27. It can be seen that negative values are dominant for the calculation of the bias of each portion, which aligns with the dominant vice scores obtained from the bar graph and is justified previously. One thing to be noted

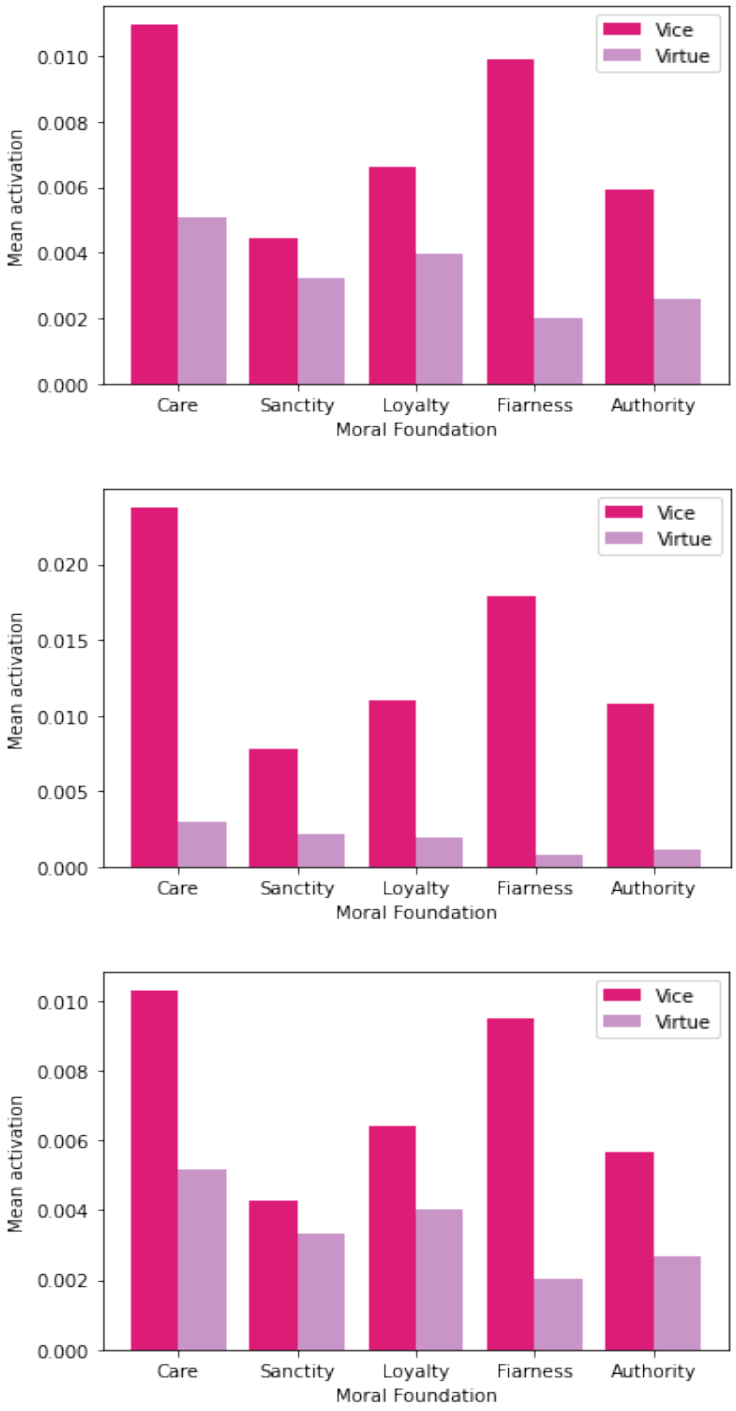

Figure 26: Mean activation scores for the Threat data set. The figures from top to bottom indicate the scores for the 1) All documents in the corpus (threat and non-threat tweets), 2) Threat tweets 3) Non-threat tweets.

is that the bias scores for the threat category is strongly negative compared to the non-threat ones which indicate that the threat comments are highly immoral with respect to each of the five foundation axis. The intensity of care and fairness in the threat category determines that this group supports and spreads more hate and promotes inequity.

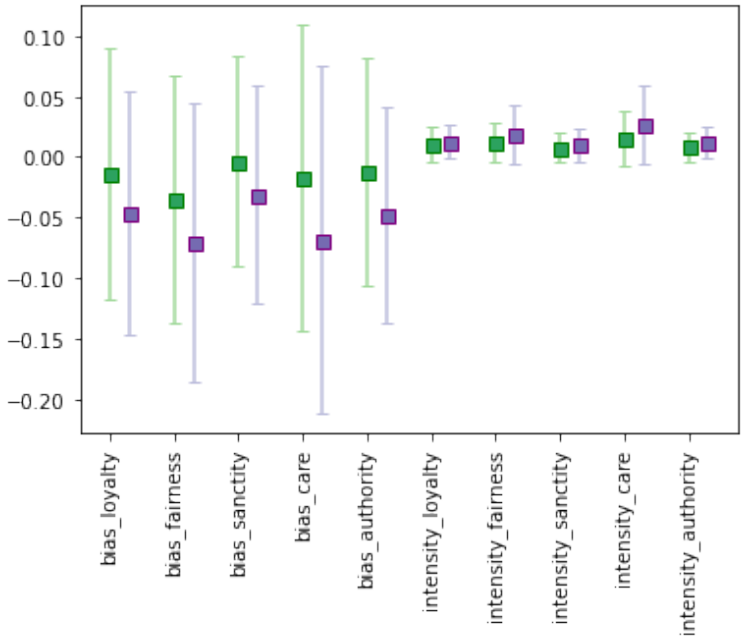

Figure 27: Bias and Intensity calculations for the Threat data set. The *green plots determine the bias and intensity for the non-threat tweets* whereas the *purple plot determine the bias and intensity for the threat tweets.*

data set	authority.vice	authority.virtue	care.vice	care.virtue	fairness.vice	fairness.virtue	loyalty.vice	loyalty.virtue	sanctity.vice	sanctity.virtue
#arsonemergency	0.008449	0.001126	0.010724	0.002536	0.010452	0.001346	0.008751	0.001229	0.007299	0.001027
suspicious tweets	0.002708	0.004028	0.002841	0.009474	0.004696	0.003484	0.001361	0.002175	0.004521	0.004696
Threat	0.005918	0.002602	0.010967	0.005050	0.009904	0.001982	0.006634	0.003937	0.004456	0.003245
Priniski et al., [72]	0.027	0.0125	0.023	0.018	0.019	0.011	0.008	0.018	0.005	0.01

Table 4: Vice and Virtue scores for all Moral Groups defined in [41]

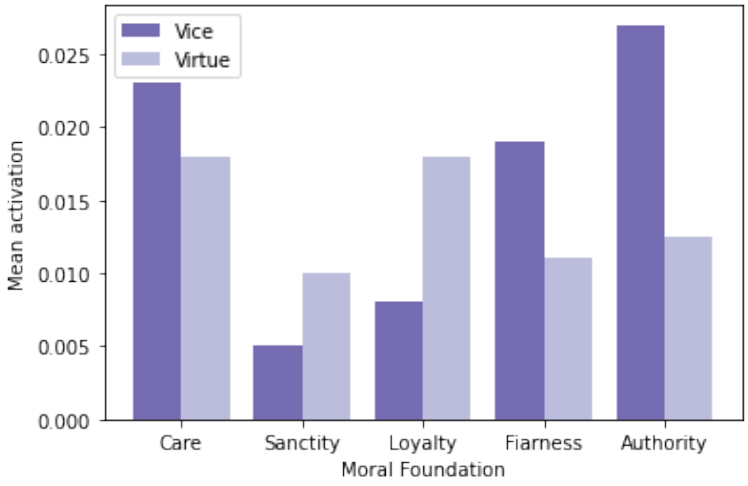

Figure 28: Mean activation scores obtained in paper [72] for the tweets following the killing of George Floyd.

Priniski et al. [72] use the same method to map the moral valence of tweets using FrameAxis [51] as presented in this thesis. The data set used in the article is the collection of the tweet wave that followed

the murder of George Floyd on 25^{th}, May 2020. This incident was followed by a large protest that is considered one of the largest social movements in American history [16], in which more than 26 million people participated. Police brutality was a major topic of discussion, where trends like "BlackLivesMatter" were prevalent on the Internet. This not only vocalized the topic of equality but also promoted it. Moral viewpoints in arguments and discussions. Moral outrage is expected to be high in this kind of scenario among the people, and the connectivity among or in group among the people increases because they collectively stand against a sensitive matter. When analyzing the data from Table 4 and Figure 28, it can be seen that the vice domain of authority, care, and fairness is presiding. However, in the domain of sanctity and loyalty, virtuosity precedes.

In general analysis of the three data sets used in the article and comparing them with [72], it is found that in the cases where people express a firm opinion or debate about a topic such as supporting and opposing the theory of arsonemergency in the case of #arsonemergency data set [94], opinion and feud between various groups of people regarding culture and religion in the Threat data set [43] or protesting against injustice in the killing of George Floyd [72], the vice or negative perspective is dominant in most cases. In the analysis of the connection of sentiment analysis and moral valence scores, a fixed connection cannot be established in this case. Although sentiment analysis resulted in positive numbers in the case of the #arsonemergency data set, the moral valence scores assigned were dominated by the vice domain. However, in cases where negative sentiment was prevalent, the mapped moral valences were dominated by the vice domain as well. For example, the general sentiment of the suspicious group was negative and the non-suspicious group was positive, as represented in Figure 9, the moral valence mapping of the correspondent was vice dominant, as shown in Figure 24.

6 Classification

Machine learning is a subfield of artificial intelligence in which different problems are solved by machines without the need to explicitly command them [64]. New insights are learned from the training data and machine learning algorithms which are used to solve further problems. The advancement in the field has also greatly benefited NLP. Various machine learning techniques such as Random Forest, Support Vector Machine (SVM), and deep learning techniques such as Convolution and Recurrent Neural Networks have significantly contributed to the field of Natural Language Processing.

6.1 BERT

BERT is a transformer-based model whose working is defined in [92, 28]. Compared to LSTM, they are faster and truly bidirectional, which helps to capture the context in general. BERT works in 2 different phases, the first being the prepossessing and the second one the fine-tuning. In the first phase, the major concern is to know the representation of the words. In short, this can also be considered as "what is a language?" The other phase is the fine-tuning phase, which is used to solve the problem in the domain context.

6.1.1 Pre-training phase

The prepossessing is the first step. To understand the context of the language, there are two unsupervised tasks that BERT utilizes simultaneously, masked language modeling and next-sentence prediction. Masked language modeling is a technique in which a certain fraction of words in a sentence are masked and these masked tokens must be predicted by the model in the context. This helps BERT to understand the bidirectional context within a sentence. Sentence prediction is then used used to understand the relevance of two sentences and the context in different sentences. Note that the pre-training phase has already trained the BERT model to a certain extent with proper context representation. The next step is to fit the model in the context of the domain of the problem to be solved.

Diving deep into the pre-training stage, here, the simultaneous occurrence of both the Masked Language Modeling and Next Sentence Prediction takes place. Two sentences are fed as input, where some of the words are masked. Words are represented as tokens and embedded using pre-trained embedding. The output of the preprocessing model is threefold. The prediction of the next sentence is represented by C in Figure 29 . This is a binary value that represents 1 if the sentence follows the previous sentence in context and 0 in the other case. Each of the T is the word vector that corresponds to the output of the problem of the masked language model problem and E represents the embedding [28].

The embedding takes place in three stages. The first one is the Token Embedding (embedding like word2vec), the second one is the segment embedding which is the embedding of sentences (sentence number), and the third is the positional embedding. This is the position of the sentence within the sentence encoded within the vector. When these vectors are added, an input embedding is created for the BERT. Segment and positional vectors are used to preserve ordering, because that is a necessity of language models [28, 92].

Figure 29: Pre training and fine tuning of BERT taken from [28]. Both the pre traing and fine tuning uses the same architecture and the same pre trained architecture is used across various NLP tasks. Only the fine tuning step changes according to the context where it is used.

However, the scope of this thesis does not lie in the pre-training of BERT. This is because papers like [28] pre-trained BERT on a large corpus called the BookCorpus [100], which contains around 800 M words, and the English Wikipedia that contains around 2500M words. These trained models are present and can be used form libraries like TensorFlow[7]. This helps to gain accuracy and includes sophisticated structures, which is why these pre-trained models are used.

6.1.2 Fine tuning phase

The fine-tuning step is the predecessor of the prepossessing step. This is the phase in which BERT is used to find solutions to NLP tasks. The model is trained to perform a specific task for this purpose. The fully connected output layers should now be connected to a fresh set of output layers so that a specific task can be performed. For example, in the question answering task, the output is the answer sets, or in any classification task, the output is the set of classes. The training time of BERT is shorter compared to other models because only the output parameters must be learned from scratch and only the input parameters must be fine-tuned [28].

1. **Methodology used**

   The fine tuning has been followed as advised in the paper Devlin et al [28]. The fine tuning follows the following series of steps. The initial preprocessing and the model setup steps are identical for all the data sets and only the different training phases are applied according to the scenario.

   (a) Setup of the data set: The setup of the data set is completed for all data sets. In each of the data sets, the tables concerned in the data frame are only the text description and the corresponding labels. Hence, these data frames are initially taken. Each of the data sets is then set according to the need. Initially, for the #arsonemergency data set, the "unaffiliated" group is removed. Then

[7] `https://www.tensorflow.org/text/tutorials/bert_glue`

the "supporter" and "opposer" groups are labeled as 1 and 0 respectively. For the Suspicious tweet data set, the same labeling method is applied where 1 is the "non-suspicious" category and 0 is the "suspicious" category of tweets. This is also followed along the Threat corpus where the "non-threat" category is represented as 1 and the "threat" category is represented as 0. The text and label values are then respectively abstracted.

(b) Preprocessing the data set: The preprocessing starts from the tokenization of the texts in the data set. Tokenization is important because the models cannot accept the texts directly and must be converted into numbers. Various methods can be used for tokenization, and in the case of BERT, the BERT tokenizer was used. Devlin et al [28] propose models with two different sizes [8] [9]. The BERT_{BASE} is a smaller model that contains 12 self attention heads (denoted as A), 12 layers of transformer blocks (denoted as L), and 768 hidden layers (denoted as H). Furthermore, the trainable parameters in this model are 110M. The figures for BERT_{LARGE} are 16 for A, 24 for L, 1024 for H, and contain 340M training parameters. In this thesis, the smaller model is used, that is, the BERT_{BASE} model. Sentences are initially taken and split into word-level tokens. A sample representation of the token representation is shown in Figure 30. The figure represents a sample that gives the words and the token IDs of the words. Each word and symbol have their own unique representation. It is important to note that the initial preprocessing step does not contain any kind of data cleaning because BERT is bidirectional and the removal of words and other steps might cause the context to be lost.

The embedding of sentences is an important part of preprocessing and is performed in a few steps, as suggested in [28].

i. Initially, special tokens are added at the beginning and end of sentences. The *CLS(ID 101) and SEP(ID 102)* tokens are used at the beginning and end of each sentence.

ii. The next step was to make sentences of the same length, which is achieved by means of padding. Padding is the process of adding values so that the lengths of shorter sentences match a standard length. Padding is done with the padding token *PAD* that has token ID 0.

iii. Finally, an attention mask is created. It is the series of tokens that is fed into the model. The attention mask is represented by 0/1 where 1 represents the tokens that are to be considered for training purposes in the model and to learn the contextual representation from the text.

(c) Splitting Data Sets: The 80% and 20% divisions are made in the training and validation data set for each corpus.

(d) Fine-tuning task: For fine-tuning, Devin et al. [28] suggest various configurations for batch size, learning rate, and number of epochs. All of these configurations were tested for different data sets, and hence various results were obtained. The fine-tuning was done on the base case of BERT, which contained around 110M parameters to train. Training was carried out in the environment

[8] `https://huggingface.co/bert-base-uncased`

[9] `https://github.com/google-research/bert`

Tokens	Token IDs
rt	19387
@	1030
she	2016
##el	2884
##f	2546
_	1035
1	1015
:	1024
why	2339
has	2038
twitter	10474
removed	3718
the	1996
#	1001
arson	24912

Figure 30: Token and token ID representation as results obtained from the BERT tokenizer. This Table is an example of a sentence from the #arsonemergency data set. This process is applied to each of the documents in the corpus at the word level.

provided by Google Colab in the presence of GPU (K80 model, 32 GB RAM). Training time and improvement in training error were different for all data sets.

(e) Classification: Classification is carried out using a simple softmax function, which converts the number vector to probability. To represent the whole sentence, BERT takes in **h** which is the final hidden state of the initial token, which is the [CLS] token. The classification probability of label c is given as [86]:

$$p(c|\mathbf{h}) = softmax(Wh)$$

here, W is a parametric metric that is task specific.

2. **Validation measures.**

The fine-tuning of BERT is used in different papers, such as [28, 52, 98], where different validation measures were used. But collectively, the most common methods used were the precision, F1, precision, and recall scores, which are calculated with the help of a confusion matrix. The four different terms are *True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN).* The number

of samples that are correctly classified into positive and negative classes is given by *TP and TN*, respectively. The samples that are incorrectly classified into positive and negative classes are given by the terms *FN and FP* respectively. The formulas for the respective figures are given as follows [93]:

$$Accuracy = \frac{TP + TN}{Total\ Samples}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

6.1.3 Experiments and results.

Figure 31: Change in Training loss value as the number of epoch increases. Ideally, the training loss is supposed to decrease to a minimum point possible but it is not always the case. Problems like over fitting can affect the performance measures of these values.

Figures 31 and 32 represent metrics related to data set training. The batch configuration was 16 and the learning rate was set to 5e-5. Hyper parameter such as the learning rate was chosen because careful selection of these parameter was needed to prevent the catastrophic forgetting problem [86]. Initially, we look at the training loss during training cycles or epochs. Training loss is a metric that determines how well training data fit the deep learning model. After each epoch, this value is calculated for the data sets that are fed into the model. These values are then noted and plotted against the epochs. For the #arsonemergency data

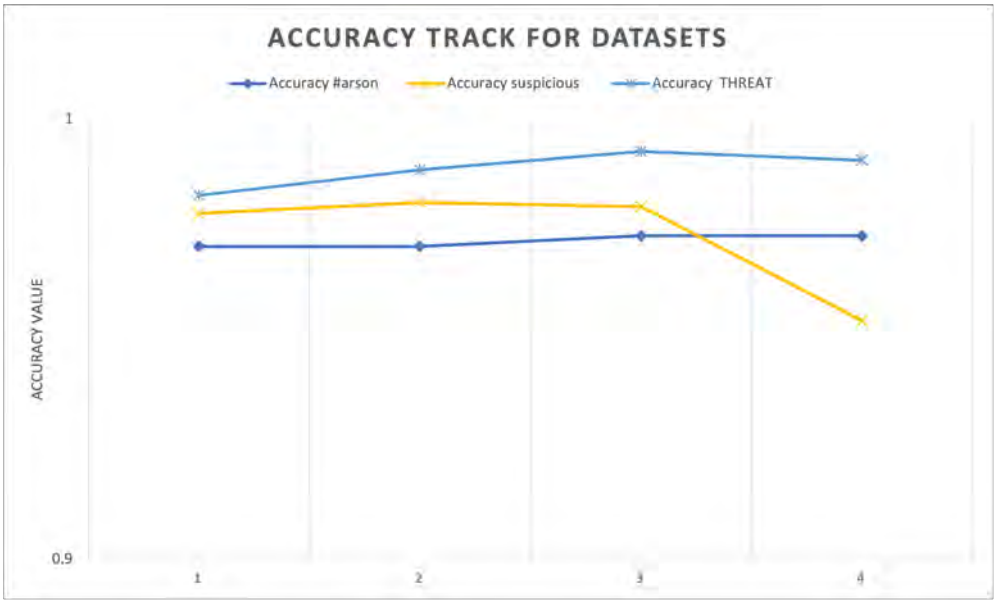

Figure 32: Change in accuracy of the model as the number of epochs increase in the context of all three data sets. Colors are individually assigned and labelled in the figure. The accuracy during the training increases until a certain epoch and decreases. The model with the best training values are saved.

set the orange line in Figure 31 is continuously decreasing. But it cannot be concluded that the smallest value is the correct value and the corresponding accuracy of the epoch must be investigated. From further evaluation of the code base, it was found that after the fourth epoch the accuracy decreased, and hence the fourth epoch was the best.

The gray line denotes the training loss values of the Suspicious tweet data set. The value decreases until the third epoch and increases in the fourth epoch. On further investigation about the same data set for the accuracy track, the accuracy decreases slightly from the 2nd to the third epoch and drastically from the 3rd to 4th approach. Hence, the second epoch is chosen as the best model. The batch size was 32 and the learning rate was set to 5e-5.

Finally, the green lines denote the training loss for the Threat data set. Training loss is consistently decreasing from the second epoch to the fourth epoch. When observing the accuracy graph for the same data set, the accuracy scores decrease after the third epoch. So, even though the training loss was the smallest in the fourth epoch, the accuracy scores suggest saving the third epoch as the best scores. The batch size was 32 and the learning rate was 5e-5.

Dataset	Accuracy	Precision	Recall	F1
#arsonemergency	.97	.97	.98	.95
Suspicious Tweets	.98	.98	.99	.86
THREAT	.98	.98	.99	.74

Table 5: Results obtained from classification. The fine tuning of a popular transformer based architecture BERT is used which provides higher accuracy in all the data sets when different configurations are applied.

The results presented in Table 5 are validating figures for all data sets used. In general, the results are mainly around or above the 95% precision, representing a higher performance of the model. The reason why BERT demonstrates high results can be seen in multiple folds. First, the BERT is based on a transformer-based architecture that uses the attention-based model [92], because the context of a word is learned with respect to other words in the corpus. This produces better representations of the text that is used for training purposes. The bidirectional technique of understanding the words and its techniques increased the accuracy of knowledge of the results. Furthermore, the 2 phases of BERT (preprocessing and fine-tuning) as described previously provide additional advantage. The pre-training was based on a powerful corpus in which the context of language is already learned. The fine-tuning is very specific and focuses only on one task. Also, the parameters considered for training are abundant (around 110M in our case), which ensures that multiple parameters are to be tuned for efficient results.

7 Discussion

This work provides a significant contribution to the growing area of social cyber security. This work has thoroughly compared and contrasted three types of datasets related to malicious activities. Insights gained from this work can serve as a foundation for more sophisticated work on intent classification.

Various experiments were performed in the thesis to obtain different results that were meaningful in numerous ways. Initially, on performing the exploratory data analysis the sentiment scores were found to be a credible measure to extract the emotional valance of the documents in each of the datasets. Furthermore, word frequency graphs were a measure to indicate the category of vocabulary used by different groups of people in different data sets. This helped to distinguish two different categories of users and their choice of words for a particular scenario in all the data sets that were used for analysis. Some of the initial assumptions that were made, such as in the case of the bushfires dataset, that the supporters and opposers category, would have the opposite polarity were proven wrong. This could only be validated through the exploratory data analysis, given we were dealing with a large corpus. EDA was followed by topic modeling where we identified various underlying topics that were through the use of K means clustering and Principal Component Analysis, (after poor results obtained from methods like the Linear Discriminant Analysis and Latent Dirichlet Allocation). Although the results obtained through the approach used in the thesis had some flaws, the results obtained had various topics that were sensible. This approach helped understand the hidden context that was more detailed compared to the annotations that were initially present in the datasets.

The FrameAxis method used in the thesis used two terms *Bias and Intensity* to calculate the moral value of the documents and thus provided both the information on the inclination of the documents towards the moral foundations and the extent of moral words used in each of the corpus. Fluctuation in the moral foundation scores was observed across the datasets, hence it is plausible to establish the connection that the morality of users' opinions varied widely depending upon the context of the topic they were participating in. However, it was presumed that texts with a more negative sentiment distribution would receive a low morality score, which was mostly true, but some results partly undermining the assumptions were also received. Finally, a state-of-the-art model was used for classification of malicious text, where experiments were carried out by repeatedly fine tuning various parameters which resulted in a high degree of classification-accuracy. This was obtained due to two reasons. Firstly, the BERT was pretrained in a very large corpus which generally means more accuracy in the context of machine learning. Secondly, BERT is able to account for the words context through its bidirectional learning approach, which gives precise representation of the context.

The use of data sets related to malicious content played an important role in the thesis, where the focus was shifted to various ill-nature activities. The arsonemergency data set helped to understand the behavior of tweets that spread misinformation on the theory of arson and also helped to explore the intentions of the opposing group. Furthermore, in the suspicious tweet dataset, the interesting thing to notice was that even though the morality scores were dominant on the positive side, the sentiments were opposing, which explores the area for further research. Also, it was initially presumed that the threat corpus would result in low

moral scores with a dominant vice nature, but the virtuous nature was significant in the corpus. However, it was evident that all groups that represented malicious behavior had very low or negative sentiment scores, vice dominant morality, and abundant use of negative words. The extremities of these results were reached in cases where sensitive topic of discussions like religion, cultures, and various groups of people were involved. Although topic modeling provided distinct results in case of #arsonemergency and threat data set where word analysis drew clear boundaries between topics, the difference was subtle for the suspicious tweet dataset. This behavior can also be considered as a basis for further research.

Various challenges were faced during the investigation. Initially, the lack of proper and publicly available data sets in the field of detecting malicious content was a major hurdle. This bounded us to use newer datasets in the field. This caused difficulty in bencmarking and contrasting the results obtained against other methods. However, this also provided the opportunity to explore, analyze, and understand the behavior of these datasets, which helped to fulfill our research objective of analyzing the behavior of languages.

8 Conclusion

Despite the usefulness of social networks for large audiences to freely present, support or oppose opinions, they also provide space for malicious activities. This can be supported by the content of the datasets presented in the thesis, where one of them contains the spread of misinformation, another contains suspicious tweets, and the other contains the spread and imposition of threat and violence in the community. It is crucial to contain this kind of behavior as much as possible in order to create a safe and peaceful environment for social media users. Hence, this thesis studies the malicious activities related datasets using NLP techniques in order to feed this research into developing more nuanced methods for detecting and combating online misinformation and social cyber threats. These NLP methods were used to detect and analyze various patterns related to the use of words, feelings, topics of discussions, and moral behaviors. These results were presented, analyzed, and different conclusions were drawn according to the context of the datasets. Although some very clear patterns were discovered during the analysis that mark a clear distinction between the different groups of interest, the opportunity to extend the investigation and analysis of the domain still remains.

References

[1] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] Swati Agarwal and Ashish Sureka. "Using common-sense knowledge-base for detecting word obfuscation in adversarial communication". In: *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*. IEEE. 2015, pp. 1–6.

[3] Swati Agarwal and Ashish Sureka. "Using knn and svm based one-class classifier for detecting online radicalization on twitter". In: *International Conference on Distributed Computing and Internet Technology*. Springer. 2015, pp. 431–442.

[4] Amritanshu Agrawal, Wei Fu, and Tim Menzies. "What is wrong with topic modeling? And how to fix it using search-based software engineering". In: *Information and Software Technology* 98 (2018), pp. 74–88.

[5] Theodore T Allen, Zhenhuan Sui, and Kaveh Akbari. "Exploratory text data analysis for quality hypothesis generation". In: *Quality Engineering* 30.4 (2018), pp. 701–712.

[6] Theodore T Allen and Hui Xiong. "Pareto charting using multifield freestyle text data applied to Toyota Camry user reviews". In: *Applied Stochastic Models in Business and Industry* 28.2 (2012), pp. 152–163.

[7] TT Allen, N Parker, and Z Sui. *Using innovative text analytics on a military specific corpus. 84th Military Operations Research Society (MORS) Symposium, Quantico, VA, June, 170*. 2016.

[8] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. "Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment". In: *arXiv preprint arXiv:1806.05521* (2018).

[9] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction". In: *Knowledge-based systems* 191 (2020), p. 105184.

[10] Claus Boye Asmussen and Charles Møller. "Smart literature review: a practical topic modelling approach to exploratory literature review". In: *Journal of Big Data* 6.1 (2019), pp. 1–18.

[11] Luigi Asprino et al. "Uncovering Values: Detecting Latent Moral Content from Natural Language with Explainable and Non-Trained Methods". In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 2022, pp. 33–41.

[12] David Bamman and Noah A Smith. "Open extraction of fine-grained political statements". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 76–85.

[13] Purnima Bholowalia and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN". In: *International Journal of Computer Applications* 105.9 (2014).

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[15] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* 29 (2016).

[16] Larry Buchanan, Quoctrung Bui, and Jugal K Patel. "Black Lives Matter may be the largest movement in US history". In: *The New York Times* 3 (2020).

[17] Katherine S Button et al. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature reviews neuroscience* 14.5 (2013), pp. 365–376.

[18] Yong Chen et al. "Experimental explorations on short text topic mining between LDA and NMF based Schemes". In: *Knowledge-Based Systems* 163 (2019), pp. 1–13.

[19] Naganna Chetty and Sreejith Alathur. "Hate speech review in the context of online social networks". In: *Aggression and violent behavior* 40 (2018), pp. 108–118.

[20] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[21] Dennis Chong and James N Druckman. "Framing theory". In: *Annual review of political science* 10.1 (2007), pp. 103–126.

[22] Wei Yen Chong, Bhawani Selvaretnam, and Lay-Ki Soon. "Natural language processing for sentiment analysis: an exploratory analysis on tweets". In: *2014 4th international conference on artificial intelligence with applications in engineering and technology*. IEEE. 2014, pp. 212–217.

[23] KR1442 Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[24] Scott Clifford and Jennifer Jerit. "How words do the work of politics: Moral foundations theory and the debate over stem cell research". In: *The Journal of Politics* 75.3 (2013), pp. 659–671.

[25] Stephan A Curiskis et al. "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit". In: *Information Processing & Management* 57.2 (2020), p. 102034.

[26] Andrew M Dai and Quoc V Le. "Semi-supervised sequence learning". In: *Advances in neural information processing systems* 28 (2015).

[27] Jeroen De Mast and Albert Trip. "Exploratory data analysis in quality-improvement projects". In: *Journal of Quality Technology* 39.4 (2007), pp. 301–311.

[28] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[29] Gaurav Dhiman et al. "MOSOA: A new multi-objective seagull optimization algorithm". In: *Expert Systems with Applications* 167 (2021), p. 114150.

[30] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453.

[31] Matthias Eickhoff and Nicole Neuss. "Topic modelling methodology: its use in information systems and other managerial disciplines". In: (2017).

[32] Robert M Entman. "Framing: Towards clarification of a fractured paradigm". In: *McQuail's reader in mass communication theory* 390 (1993), p. 397.

[33] Dustin A Fife and Joseph Lee Rodgers. "Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis"." In: *American Psychologist* 77.3 (2022), p. 453.

[34] Dean Fulgoni et al. "An empirical exploration of moral foundations theory in partisan news sources". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 3730–3736.

[35] Iginio Gagliardone et al. *Countering online hate speech*. Unesco Publishing, 2015.

[36] Kristel M Gallagher and John A Updegraff. "Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review". In: *Annals of behavioral medicine* 43.1 (2012), pp. 101–116.

[37] Justin Garten et al. "Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis". In: *Behavior research methods* 50.1 (2018), pp. 344–361.

[38] Justin Garten et al. "Morality between the lines: Detecting moral sentiment in text". In: *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*. 2016.

[39] Jesse Graham et al. "Mapping the moral domain." In: *Journal of personality and social psychology* 101.2 (2011), p. 366.

[40] Dhruv Grewal, Jerry Gotlieb, and Howard Marmorstein. "The moderating effects of message framing and source credibility on the price-perceived risk relationship". In: *Journal of consumer research* 21.1 (1994), pp. 145–153.

[41] Jonathan Haidt and Jesse Graham. "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize". In: *Social Justice Research* 20.1 (2007), pp. 98–116.

[42] Jonathan Haidt and Craig Joseph. "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues". In: *Daedalus* 133.4 (2004), pp. 55–66.

[43] Hugo L Hammer et al. "Threat: A large annotated corpus for detection of violent threats". In: *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2019, pp. 1–5.

[44] Frederic R Hopp et al. "The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text". In: *Behavior research methods* 53.1 (2021), pp. 232–246.

[45] Jeremy Howard and Sebastian Ruder. "Universal language model fine-tuning for text classification". In: *arXiv preprint arXiv:1801.06146* (2018).

[46] Albert H Huang et al. "A study of factors that contribute to online review helpfulness". In: *Computers in Human Behavior* 48 (2015), pp. 17–27.

[47] Ioana Hulpus et al. "Knowledge graphs meet moral values". In: Association for Computational Linguistics. 2020.

[48] Ahmad Izzuddin. "Optimasi Cluster pada Algoritma K-Means dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen". In: *Energy-Jurnal Ilmiah Ilmu-Ilmu Teknik* 5.2 (2015), pp. 41–46.

[49] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. "Quantitative analysis of large amounts of journalistic texts using topic modelling". In: *Digital journalism* 4.1 (2016), pp. 89–106.

[50] Andrew T Jebb, Scott Parrigon, and Sang Eun Woo. "Exploratory data analysis as a foundation of inductive research". In: *Human Resource Management Review* 27.2 (2017), pp. 265–276.

[51] Haewoon Kwak et al. "FrameAxis: characterizing microframe bias and intensity with word embedding". In: *PeerJ Computer Science* 7 (2021), e644.

[52] Jieh-Sheng Lee and Jieh Hsiang. "Patent classification by fine-tuning BERT language model". In: *World Patent Information* 61 (2020), p. 101965.

[53] Guang Li et al. "Research on the natural language recognition method based on cluster analysis using neural network". In: *Mathematical Problems in Engineering* 2021 (2021).

[54] Yang Li and Tao Yang. "Word embedding for understanding natural language: a survey". In: *Guide to big data applications*. Springer, 2018, pp. 83–104.

[55] Kar Wai Lim and Wray Buntine. "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon". In: *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 2014, pp. 1319–1328.

[56] Durairaj Maheswaran and Joan Meyers-Levy. "The influence of message framing and issue involvement". In: *Journal of Marketing research* 27.3 (1990), pp. 361–367.

[57] Madhumita Guha Majumder, Sangita Dutta Gupta, and Justin Paul. "Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis". In: *Journal of Business Research* 150 (2022), pp. 147–164.

[58] E Matthew. "Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. Deep contextualized word representations." In: *Proc. of NAACL*. 2018.

[59] Rishabh Mehrotra et al. "Improving lda topic models for microblogs via tweet pooling and automatic labeling". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 2013, pp. 889–892.

[60] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[61] Negar Mokhberian et al. "Moral framing and ideological bias of news". In: *International Conference on Social Informatics*. Springer. 2020, pp. 206–219.

[62] Stephan Morgenthaler. "Exploratory data analysis". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009), pp. 33–44.

[63] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. "Natural language processing: an introduction". In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[64] Tatwadarshi P Nagarhalli, Vinod Vaze, and NK Rana. "Impact of machine learning in natural language processing: A review". In: *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*. IEEE. 2021, pp. 1529–1534.

[65] Arpita Nagpal, Arnan Jatain, and Deepti Gaur. "Review based on data clustering algorithms". In: *2013 IEEE conference on information & communication technologies*. IEEE. 2013, pp. 298–303.

[66] Elhadji Ille Gado Nassara, Edith Grall-Maës, and Malika Kharouf. "Linear discriminant analysis for large-scale data: Application on text and image data". In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2016, pp. 961–964.

[67] Thomas E Nelson, Zoe M Oxley, and Rosalee A Clawson. "Toward a psychology of framing effects". In: *Political behavior* 19.3 (1997), pp. 221–246.

[68] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. "Topic modelling for qualitative studies". In: *Journal of Information Science* 43.1 (2017), pp. 88–102.

[69] Peter Organisciak, Benjamin M Schmidt, and J Stephen Downie. "Giving shape to large digital libraries through exploratory data analysis". In: *Journal of the Association for Information Science and Technology* 73.2 (2022), pp. 317–332.

[70] Jelili Oyelade et al. "Data Clustering: Algorithms and Its Applications". In: *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*. 2019, pp. 71–81. DOI: 10.1109/ICCSA.2019.000-1.

[71] Manu Panwar, Amit Wadhwa, and Sanjeev Pippal. "Recommendation System with Exploratory Data Analytics using Machine Learning". In: *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE. 2021, pp. 84–87.

[72] J Hunter Priniski et al. "Mapping moral valence of tweets following the killing of George Floyd". In: *arXiv preprint arXiv:2104.09578* (2021).

[73] Jipeng Qiang et al. "Short text topic modeling techniques, applications, and performance: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* 34.3 (2020), pp. 1427–1445.

[74] Junaid Rashid, Syed Muhammad Adnan Shah, and Aun Irtaza. "An efficient topic modeling approach for text mining and information retrieval through K-means clustering". In: *Mehran University Research Journal of Engineering & Technology* 39.1 (2020), pp. 213–222.

[75] Shamik Roy and Dan Goldwasser. "Analysis of Nuanced Stances and Sentiment Towards Entities of US Politicians through the Lens of Moral Foundation Theory". In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, June 2021, pp. 1–13. DOI: 10.18653/v1/2021.socialnlp-1.1. URL: https://aclanthology.org/2021.socialnlp-1.1.

[76] Zoë Wilkinson Saldaña. "Sentiment Analysis for Exploratory Data Analysis". In: *Programming Historian* (2018).

[77] Ma Shiela C Sapul, Than Htike Aung, and Rachsuda Jiamthapthaksin. "Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms". In: *2017 14th international joint conference on computer science and software engineering (JCSSE)*. IEEE. 2017, pp. 1–6.

[78] Dietram A Scheufele. "Framing as a theory of media effects". In: *Journal of communication* 49.1 (1999), pp. 103–122.

[79] Alok Sharma and Kuldip K Paliwal. "Linear discriminant analysis for the small sample size problem: an overview". In: *International Journal of Machine Learning and Cybernetics* 6.3 (2015), pp. 443–454.

[80] Sughash Sharma. "Applied multivariate techniques". In: (1996).

[81] Aditi Anand Shetkar and S Fernandes. "Text categorization of documents using K-means and K-means++ clustering algorithm". In: *Int J Recent Innov Tren Comput Commun* 4.6 (2016), pp. 485–489.

[82] Yanchuan Sim et al. "Measuring ideological proportions in political speeches". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 91–101.

[83] Paul M Sniderman and Sean M Theriault. "The structure of political argument and the logic of issue framing". In: *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change* (2004), pp. 133–65.

[84] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. "The general inquirer: A computer approach to content analysis." In: (1966).

[85] Zhenhuan Sui. "Hierarchical text topic modeling with applications in social media-enabled cyber maintenance decision analysis and quality hypothesis generation". PhD thesis. The Ohio State University, 2017.

[86] Chi Sun et al. "How to fine-tune bert for text classification?" In: *China national conference on Chinese computational linguistics*. Springer. 2019, pp. 194–206.

[87] MA Syakur et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster". In: *IOP conference series: materials science and engineering*. Vol. 336. 1. IOP Publishing. 2018, p. 012017.

[88] Livia Teernstra et al. "The morality machine: Tracking moral values in tweets". In: *International Symposium on Intelligent Data Analysis*. Springer. 2016, pp. 26–37.

[89] M Thangaraj and M Sivakami. "Text classification techniques: A literature review". In: *Interdisciplinary Journal of Information, Knowledge, and Management* 13 (2018), p. 117.

[90] John W Tukey et al. *Exploratory data analysis*. Vol. 2. Reading, MA, 1977.

[91] Joseph Turian, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. 2010, pp. 384–394.

[92] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[93] Ni Wayan Surya Wardhani et al. "Cross-validation metrics for evaluating classification performance on imbalanced data". In: *2019 international conference on computer, control, informatics and its applications (IC3INA)*. IEEE. 2019, pp. 14–18.

[94] Derek Weber et al. "# ArsonEmergency and Australia's "Black Summer": Polarisation and Misinformation on Social Media". In: *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer. 2020, pp. 159–173.

[95] Aksel Wester et al. "Threat detection in online discussions". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2016, pp. 66–71.

[96] Pengtao Xie and Eric P Xing. "Integrating document clustering and topic modeling". In: *arXiv preprint arXiv:1309.6874* (2013).

[97] Qi Yang and Lin Shang. "Multi-task learning with bidirectional language models for text classification". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[98] Shehel Yoosuf and Yin Yang. "Fine-grained propaganda detection with fine-tuned BERT". In: *Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda*. 2019, pp. 87–91.

[99] Weizhong Zhao et al. "A heuristic approach to determine an appropriate number of topics in topic modeling". In: *BMC bioinformatics*. Vol. 16. 13. Springer. 2015, pp. 1–10.

[100] Yukun Zhu et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.