





Article

An Empirical Approach for Extreme Behavior Identification through Tweets Using Machine Learning

Waqas Sharif ¹, Shahzad Mumtaz ¹, Zubair Shafiq ², Omer Riaz ¹, Tenvir Ali ¹,
Mujtaba Husnain ¹ and Gyu Sang Choi ^{3,*}

¹ Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan

² Department of Media Studies, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan

³ Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea

* Correspondence: castchoi@ynu.ac.kr

Received: 24 July 2019; Accepted: 3 September 2019; Published: 6 September 2019



Abstract: The rise of social media has led to an increasing online cyber-war via hate and violent comments or speeches, and even slick videos that lead to the promotion of extremism and radicalization. An analysis to sense cyber-extreme content from microblogging sites, specifically Twitter, is a challenging, and an evolving research area since it poses several challenges owing short, noisy, context-dependent, and dynamic nature content. The related tweets were crawled using query words and then carefully labelled into two classes: Extreme (having two sub-classes: pro-Afghanistan government and pro-Taliban) and Neutral. An Exploratory Data Analysis (EDA) using Principal Component Analysis (PCA), was performed for tweets data (having Term Frequency—Inverse Document Frequency (TF-IDF) features) to reduce a high-dimensional data space into a low-dimensional (usually 2-D or 3-D) space. PCA-based visualization has shown better cluster separation between two classes (extreme and neutral), whereas cluster separation, within sub-classes of extreme class, was not clear. The paper also discusses the pros and cons of applying PCA as an EDA in the context of textual data that is usually represented by a high-dimensional feature set. Furthermore, the classification algorithms like naïve Bayes', K Nearest Neighbors (KNN), random forest, Support Vector Machine (SVM) and ensemble classification methods (with bagging and boosting), etc., were applied with PCA-based reduced features and with a complete set of features (TF-IDF features extracted from *n*-gram terms in the tweets). The analysis has shown that an SVM demonstrated an average accuracy of 84% compared with other classification models. It is pertinent to mention that this is the novel reported research work in the context of Afghanistan war zone for Twitter content analysis using machine learning methods.

Keywords: cyber-extreme; Twitter; exploratory data analysis; principal component analysis; term frequency—inverse document frequency; support vector machine; ensemble classification

1. Introduction

In recent years, the internet has become a global platform for communication and dissemination of information. Today, popular social media sites have a huge global reach and audience, with Facebook having more than 2.27 billion monthly active users [1], while YouTube boosting almost 1 billion users every month [2]. Similarly, Twitter has an average of 335 million monthly active users [3] and it increased with 14% per day over the last few years [4]. Some other popular social media are Instagram, LinkedIn, Tumblr, WeChat and WhatsApp. The registered number of total users on these social sites are in the billions [5].

Nowadays, due to the extensive use of social media, a large amount of data is generated on a daily basis. The data generated from these sites is in an unstructured form that creates opportunities, as well as challenges for processing and analysis to make a useful understanding of the underlying hidden patterns.

The data gathered from the social sites are used for various purposes such as;

- Marketing & Promotions
- Recommendation system
- Sentiment analysis
- Review system
- Fraud detection
- Crime monitoring
- Terrorism/Extremism detection

Cyberspace offers freedom of flows of communication and opinion expressing, since there are some groups that are exploiting the capability of social media to spread distorted belief and negative influence on other people. A number of research studies explored that the current social media is regularly being misused by many hate groups to promote radicalization (also stated as cyber-crime, cyber-extremism, and cyber-hate-propaganda) [6–9]. Many people use these social media platforms to promote harmful ideology by spreading extreme contents among their audience. These radical groups post offensive and violent messages, comments, and hateful speeches focusing on their objectives. They even communicate with other such existing virtual communities to extend their network on social media based on sharing a similar agenda. Social media has now become the easiest way for these extremist groups to recruit new members in their groups by reaching a world-wide audience and then these groups gradually tend to influence newly recruited members in spreading violence and extremism. A study presented in [10,11] shows that in 2015, more than 125,000 accounts were detected from Twitter through human judgments that were associated with terrorist activities. As a consequence of growing extremism content on Twitter, the company developed a team of professionals in various countries (United States of America (USA), Ireland, etc.) in order to observe suspected accounts and to block them when identified.

The increasing trend of social media for expressing views requires special attention of language processing and machine learning research community to devise techniques that can help government agencies to automatically identify users who have extreme views. Such efforts will surely help the agencies to control crime as the presence of such content on social media in massive volume is one of the biggest concerns nowadays. The issues discussed above make an investigation of online extreme contents, as an important area of research.

Researchers from various disciplines such as social science, psychology, and computer science are working hard to develop new tools and techniques to counter and combat the problems of identification of online extremism. Specifically, the need for computer science researcher's attention has risen to a great extent to develop an automated system(s) that help to identify users/groups who are posting extreme views on social media platforms. In order to accomplish this task, most of the researchers used approaches like term frequency and lexicon-based dictionaries such as *SentiStrength* and *SentiWordNet* [12–14]. However, the outcome of such techniques is not reliable because human written language is more than word frequencies, since language also has semantic orientations. The traditional lexicon approaches (also stated as bag-of-words) have some good outcomes on general sentiment analysis such as topic modeling and product or movie review system. The lexicon-based approaches are not considered efficient in sensing extremism due to the limitation of ignoring the semantic orientation, because these approaches perceive the meaning of a textual content using the term frequency only. Therefore, these approaches do not seem to be successful in distinguishing between the tweets having either extreme or neutral views.

Unlike dictionary-based approaches, machine learning can also be used to predict social media extreme content. As a common practice, this problem is handled in machine learning using supervised approaches that require pre-defined labels for training a classification model. This limitation is often resolved using sentiment-based dictionaries in order to label each sample. We observed that quite often such dictionary-based approaches to label the training data give false labels due to the limitation of dictionaries to sense the context of sentences. We applied dictionary-based approaches to label tweets, either extreme or neutral, for our Twitter dataset. After carefully observing each tweet and its label, we observed that around 70% of dictionary-based assigned labels, for a sub sample data of 1000 tweets, were incorrect. This has raised some serious questions on earlier work that was often based on a dataset with labels assigned using dictionary-based approaches before applying any machine learning-based classifier [10,14,15]. Therefore, for generating training data, we used a dataset where we carefully read each tweet and then assigned a suitable label to it. The need to improve the automatic labelling procedure requires more attention of researchers than ever before.

Over the last few decades, machine learning has been widely used for various problems like movie and product review systems, authorship attribution, proteomic and genomic studies, sentiment analysis, etc. However, analysis of the extreme or radicalized social media content studies are very few. The earlier published work of identifying extremism on social media platforms is mostly related to a terrorist group called the Islamic State of Iraq and Syria (ISIS). Very little work has been done in the context of Taliban using social media analysis to identify radicalized content (see Table 1). This study involves analysis of the microblogging site Twitter data (crawled using a set of keywords) using machine learning approach to understand that if a tweet or set of tweets have extreme/radicalized content. This study takes Afghanistan as a region of interest for the purpose of analysis.

Table 1. Previous work on radicalization detection.

Source	Extremists Type	Datasets	Source	Methodology
[10]	ISIS	own	Twitter	machine learning
[14]	extremist groups	Own	web forum	lexicon-based
[15]	extremist groups	Own	web forum	lexicon based
[16]	extremist groups	Own	web forum	machine learning
[17]	ISIS	Own	web forum	machine learning
[18]	Hate-speeches	Available	Twitter	bag-of-words
[19]	ISIS	Own	Twitter	Clustering
[20]	ISIS	Own	Twitter	Network
[21]	ISIS	Own	Twitter	machine learning
[22]	radicalized users	Own	YouTube	lexicon-based
[23]	ISIS	Available	Twitter	machine learning
[24]	radicalized users	Own	Twitter	lexicon-based
[proposed]	Taliban	Own	Twitter	machine learning

Among all social media platforms, Twitter is chosen since it is widely used by all communities ranging from common man to celebrities with the short form of text content in every tweet. The maximum size of an earlier tweet message was 140 characters per tweet which had been extended to 250 characters since November 2017. The short length of tweets, however, makes the problem more challenging because users often tweet on a topic giving very limited contextual information making it hard to perform sentiment analysis when compared with other social media platform contents [25].

The main contribution of the proposed methodology are as follows:

- **Importance of Data Labeling:** To build a sentiment classification model, a labelled data set is needed to train and test the machine learning model. Commonly, as mentioned earlier, researchers use a dictionary-based method (i.e., SentiWordNet, SentiStrength) for this purpose. We reported that in aspect-based sentiment analysis the current labeling method has a lack of correctness and reliability.

- **Exploratory Data Analysis (EDA):** An EDA is a technique used for in-depth statistical analysis in order to make a better understanding of the dataset and this is usually achieved through data visualization techniques. Since a tweet-based content is in an unstructured form that requires converting into a structured form before performing a statistical analysis. The resulting structured form of textual data often has high dimensions that are not practically viable to visualize using conventional visualization approaches. In order to resolve this issue, an unsupervised approach called Principal Component Analysis (PCA) [26] is adopted to project a high dimensional data onto a low dimensional space to get an insight by visualizing it onto a 2-D or 3-D plot. The 2-D or 3-D visualization help users to know how each tweet is related to every other tweet in terms of neighborhood and in terms of assigned group labels.
- **Classification Model Building for Sentiment Analysis:** Supervised machine learning algorithms are used for classifying tweets having extreme contents. Several classification methods including very commonly known ones like naïve Bayes', decision tree, random forest, K Nearest Neighbors (KNN), Support Vector Machine (SVM) and lesser-known ones like ensemble classification methods with boosting and bagging approaches were applied for predictive analytics purpose. Furthermore, to boost the accuracy of the classification models, feature sets such as n -grams and TF-IDF are empirically evaluated.

The remainder of the paper is presented as follows: In Section 2, detailed reviews of detecting radicalized groups from the web using sentiment analysis are discussed; In Section 3, the proposed methodology for detecting the extreme content is presented. The experimental setup for evaluating the proposed framework, and the results are demonstrated in Section 4; Section 5 concludes the proposed framework and its effectiveness for detecting extreme contents in Twitter data in the context of Afghanistan.

2. Related Work

In the context of extremism and terrorism-related sentiment analysis, most of the previous work is performed on right-wing extremism or on a terrorist group ISIS. A report on terrorism published by USA Center for Cyber and Homeland Security describes that the social media is a key tool to promote radicalization and the study also reveals that this is a widely adopted way of recruiting new members, in extremist groups, across the world particularly focusing on users from Europe and USA [27]. Another study [28] reported that approximately 89% of organized terrorism on the internet take place through social media networks. In response to a terrorist attack in London on 3 June 2017, British Prime Minister blamed social media site in her dialogue: "We cannot allow this ideology the safe space it needs to breed—yet that is precisely what the internet and the big companies that provide Internet-based services provide" [29]. Therefore, predicting terrorism from social media becomes an essential job because the terrorism poses a serious threat to society and security of citizens. In order to limit this cyber-terrorism, several researchers show their efforts by predicting radicalization content from the social media platform (see Figure 1).

The area of Natural Language Processing (NLP) and sentiment analysis evolve gradually year by year since the last few decades. Figure 1 depicts the machine learning techniques used over a timeline in the existing literature to inspect extreme contents from the web. KNN, naïve Bayes', EDA, data clustering, decision tree, Gradient Boosted Decision Tree (GBDT), and Deep Neural Network (DNN) are the most adopted techniques for radicalization detection on social media site(s) [30–36].

In [12], the authors present a model for detecting radical content from web forums. The model was built using *SentiWordNet*, *WordNet*, and python based Natural Language ToolKit (NLTK) to predict polarity and intensity of radicalization on a web forum. Two different Arabic web forums *Mantada* and *Qawem* were used to perform experiments. A dataset containing 500 sentences was selected from each forum and translated manually into the English language. Each sentence was broken into tokens and then stored in a bag-of-words form, after preprocessing POS tagging was used to assign a tag to each

word. To assign a score to each word, they used a lexicon approach: *WordNet* and *SentiWordNet*. The sentence score was then calculated by taking an average of word scores in a sentence.

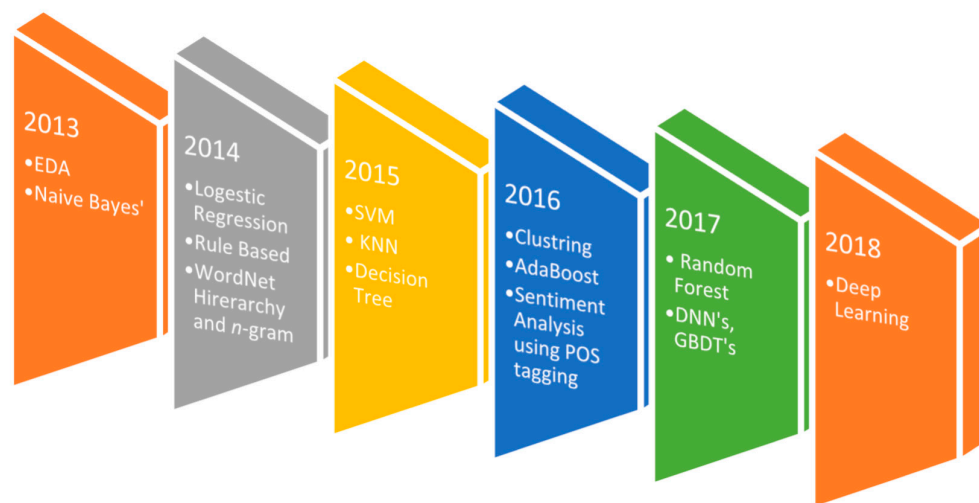


Figure 1. Machine learning techniques used for online radicalization and terrorist detection.

In [14], the authors illustrated a sentiment analysis approach to classify radical text including radical right and radical Islamic on the web. The work presented in the paper is based on a machine learning approach for classifying websites as pro-extremist and anti-extremist. They used a custom-written web crawler called Terrorism and Extremism Network Extractor (TENE) to collect data from 102 online websites. POS tags were used to find high occurrence nouns in the targeted webpages. For assigning sentiment to each page, they used *SentiStrength* dictionary to evaluate sentiment based on the specific keywords. After labeling with *SentiStrength*, decision tree algorithm was used for classification task. The accuracy of the two-class classification model (i.e., two categories of extremists) was observed to be higher compared with three-class (i.e., three categories of extremists), and four class (i.e., four categories of extremists) classification tasks respectively.

In [15], the authors identified the most radical users from the dark web using lexicon-based approach. Four popular web forums (i.e., Gawaher, Islamic Network, Islamic Awakening and Turn to Islam) are investigated that allow discussion on jihadists and terrorism. A list of 400 keywords from these web forums was prepared using POS tagging and then this list was used to get sentiment score of each post of a user. Sentiment analysis on user's online posts and comments considering the POS-based generated dictionary into account was carried out using statistical analysis and named this tool as Sentiment-based Identification of Radical Authors (SIRA). The SIRA predicts the level of extremism in users by analyzing the average sentiment score of posts, volume, severity and duration of any negative posts.

In [22], the authors proposed a computational framework that combines social network analysis with sentiment analysis tools to analyze radical groups on YouTube. The data (i.e., comments, profile) were crawled from a group of 700 suspected YouTuber accounts. The authors investigated these users on different topics with respect to polarity in order to identify the sign of intolerance and extremism. A lexicon-based module was used to determine the target topic, and then the sentiment analysis technique was applied to get the opinion of users towards these topics. Two different results for males and females are drawn to represent the most positive and negative topics in both categories. The results demonstrate that females are more positive toward *al-Qaida* and negative toward *Judaism*. Whereas, for the males, the results demonstrate higher positivity on *Islam*.

In [23], Brookings Institution conducted a study to understand the relationship between social media and terrorism. A combined approach of machine learning and manual processing was adopted to analyze population of ISIS supporters on Twitter. The work presented in this paper investigates

20,000 suspected Twitter accounts and claims that 93% of all examining accounts are supporting ISIS. Furthermore, a minimum threshold of 46,000 and max bound of 90,000 was estimated for existing ISIS-supporting accounts from October to November 2014. Where each account had an average of 1000 followers, considerably higher than any ordinary Twitter user.

A hybrid approach of sentiment analysis was proposed in [10] to predict ISIS supporter accounts on Twitter. Herein, real-time tweets are collected using specific query words (i.e., ISIS, bomb, terrorist) with *Twitter Streaming API*. After collecting data, a lexicon-based approach was used for labelling purpose, where *SentiWordNet* dictionary was used to calculate the score of each tweet. For classification task, naïve Bayes' algorithm was used. In this work a two-fold approach was applied where, after classifying every user, more tweets were collected from the same user account and verification of assigned sentiment score was performed using tweet history.

Another approach was presented in [24] where tweets were grouped into various radical groups based on the presence of special keywords such as "Al-Qaida", "Jihad", "Terrorist Operations" and "Extremism" through a lexical-based approach. A dictionary of semantically related terms (categories) was created by observing hashtags in tweets. For classification task, tweets were vectored based on dictionary related words, such as if any tweets contain dictionary related words, the score of tweets will be 1 otherwise 0. A set of rules was defined for each category and a vector representing each tweet was then compared with all rules to assign the most appropriate category to a tweet having radical content.

Most of the work in the literature regarding extremism detection through social media content analysis using machine learning techniques was done related to ISIS. In the context of Afghanistan, much of the region has been in wars in the last fifty years and hence the literacy rate is not much high and a small community use Twitter as a medium of communication. Our research focuses on analyzing Twitter-based data, generated by the people of Afghanistan, in order to identify users having extreme views or neutral views that highly affects the mode of society. The purpose of this analysis is to help organizations to better understand the extremism in the society and to define strategies to reduce hate views in order to build a better peaceful society.

3. Materials and Methods

This study aims to build a state-of-the-art framework to assess the efficacy of technological advancement in the context of text-based content analysis. In this study we collected data from Twitter using *Twitter Streaming API* and apply standard preprocessing techniques of NLP that ranges from tokenization, stemming, lemmatization, computing TF-IDF features, etc., to generate a dataset. The main objective of this study is to develop a model that can predict a class of a given tweet either as neutral or extreme.

We proposed a framework that involves a two-step process of using machine learning methods in order to get better understanding of the data and predictive ability. The first step is to perform an Exploratory Data Analysis (EDA) [37,38], where the objective is to have a better understanding of underlying hidden patterns in the data. As part of an EDA, we use to apply Principal Component Analysis (PCA) to reduce dimensions of our dataset in order to visually observe hidden patterns in the data. As the second step of our analysis, we apply various machine learning classification models such as SVM, naïve Bayes', decision tree, KNN and ensemble classification methods with boosting and bagging approaches. The evaluation of classification models is a key step. We compute evaluation measures such as accuracy, precision, recall and F-score in order to evaluate predictive performance of the algorithms applied. In our experiments, we also demonstrate the impact of data size on the performance of classification models.

In subsequent sub-sections, the key stages of our analysis are explained. These stages are data collection, data pruning, preprocessing, feature extraction, and exploratory data analysis, predictive model building and its evaluation. Figure 2 represents the schematic representation of the proposed framework of classifying tweets either as extreme or as neutral.

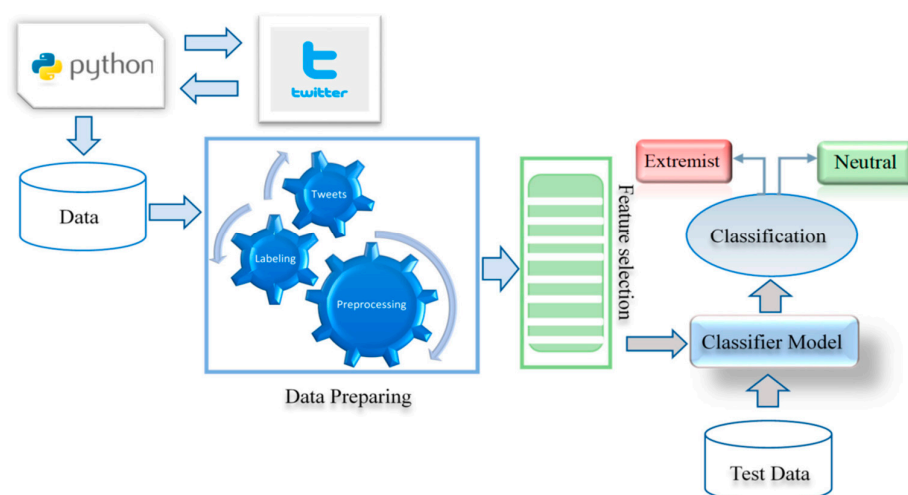


Figure 2. Proposed framework.

3.1. Data Collection and Preparation

Many Twitter-based datasets are available over the internet either freely or commercially for understanding public sentiment on social or political issues [39,40]. However, there is no dataset available in the context of Afghanistan war zone for assessing public sentiment on the war situation in Afghanistan. Therefore, we collected new and relevant data from Twitter associated with our problem. Twitter provides APIs that allow researchers to extract real-time tweets using different parameter settings for text analytics. These APIs extract tweets either based on given query terms or tweets from a profile of a specified user or based on given geo-location or language constraints or combination of any of these. The request to API not only returns tweet text but it also other information that includes username, user location, tweet text, user mentions in tweets, etc. The API supports the JavaScript Object Notation (JSON), Extensible Markup Language (XML) and Really Simply Syndication (RSS) formats.

$$Data [] = Search(term, username, longitude, latitude, lang)$$

It is important to mention that Twitter generates more than 500 million tweets per day on different topics and events [41,42]. To fetch the only relevant tweets from all the available data, a query was carefully prepared. Afghanistan was chosen as a targeted geo-location with the conjunction of a query for data extraction. The reason for selecting Afghanistan as our key focus was because of an extreme level of conflict between the public of Afghanistan on the matter of Taliban and Afghan government. Many *Afghani* support Taliban groups and they express openly about it on social media. Similarly, there are also those people in Afghanistan who support Afghan government and express their opinion against the Taliban on the social media platforms very openly. There is also a big community in Afghanistan who discuss issues and have their opinions—and their opinions are quite fair and neutral with a focus on bringing peace in the community. Considering this situation, we formulated a problem definition stating that studying Afghanistan is the most suitable case scenario to identify extreme and neutral content in tweets using the machine learning techniques. We believe that the same can be replicated to other geo-locations in future studies.

For relevant data extraction, we used several Twitter trends that are related to Taliban matters; while from all of them the #kunduz event was the hottest trending topic. The Kunduz trend was the result of an airstrike conducted by Afghan forces on a religious school in the strong hold area of Taliban on 2 April 2018. Afghan government was of the view that Taliban leaders' presence at the Madrassa was the reason of this attack by the Afghan forces.

In order to collect data from Twitter to understand the sentiment of Afghan people after Kunduz Madrassa attack in Afghanistan, the list of query words was prepared based on the trending topics

on Twitter, from the 2 to 8 April 2018. The total query words that we use for extracting tweets were 60. Table 2 provides a list of 32 query words that returned most of the tweets for our dataset. The remaining query words, like ‘Bomb’, ‘suicide’, ‘Force’, ‘Army’, ‘HungerStrike’, ‘MullahOmer’, etc., are the ones that returned small number of tweets. A total of 7500 tweets were downloaded based on the query words in the geo-location of Afghanistan for a span of one week after the Kunduz incident. All the collected data were stored locally in an Excel format for further processing and analysis purposes.

Table 2. Top trends used for data extractions.

[illegible]

3.1.1. Data Labeling

The accuracy of labeling affects the performance of the model in a way that a classifier may fall apart and there is no guarantee that the model can predict or classify correctly on low-quality labelled data. Labelling a large dataset is a critical building block and a key factor in supervised learning. Even for some machine learning task, this is the costly and time-consuming job. Due to this high cost of the labelling process, researchers often adopt semi-supervised approaches for many real-world applications, where a large amount of unlabelled data is given as input with the conjunction of a few labelled samples to perform a classification task.

The two well-known approaches adopted by the research community for data labeling are manual annotation and automated annotations/data programming. In the automated-labeling approach, various dictionary resources and data programming tools are used in a way that they take dataset as input and generate a label for each sample in the dataset. The examples of such tools are *WordNet*, *SentiWordNet* and *SentiStrength* dictionaries that are easily available for data annotations. In [10,14], an automated approach adopted in data labeling process for supervised classifiers. As we mentioned earlier that the limitation of automated labeling is that it relies only on a catalog of words and the class label is assigned based on the predefined polarity of words available in the dictionary for a given observed sample. Whereas in many applications, the context of the sentence is not only dependent on the word occurrence but even order of the words is also significant. In such a scenario, the context of

the overall sentence is beyond the word occurrence, automated labelling does not give accurate results. Therefore, researchers are trying their best to improve automated labelling mechanisms. Because of limitation of existing automated labelling to label content by ignoring semantic, we opted to choose manual labelling process and we labelled our dataset by carefully reading each tweet by not only considering occurrence of each word but also considering the context and semantics as well. Though hand labelling required more time and human resources, it is more accurate and reliable.

3.1.2. Data Pruning

The extracted tweets are in raw form, that contain “noise” or “undesired data” in the form of irrelevant terms, symbols, links, and punctuation marks. These irrelevant terms in the data are not useful for the model and may reduce the performance of classifiers. To remove such undesired data from the corpus, we perform some preprocessing tasks on data that are described in the subsequent sections (see Figure 3).

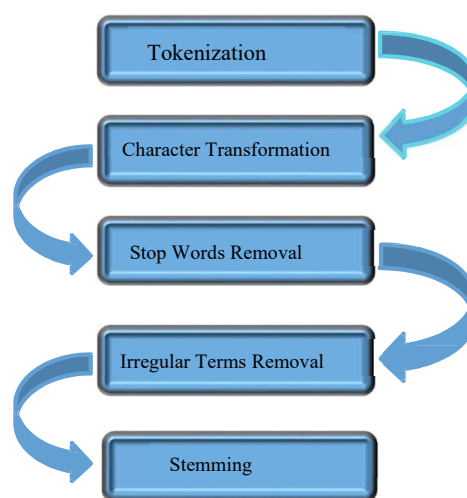


Figure 3. Data preprocessing steps.

Tokenization

In order to eliminate undesired terms, and to construct a word vector for the model, we parsed every tweet into tokens. Tokenization transform the tweet text into words segment such as if we had a tweet like [@pajhwok Former president slams Pakistan’s airstrike in Kunar], it will be segmented with the rule that each word in the tweet is separated by a space or special character in the tokenization phase. The process of tokenization converts the sentence into tokens and generate the following output like [@pajhwo] [Former] [president] [slams] [Pakistan’s] [airstrike] [in] [Kunar].

Character Transformation

It is possible that the same token may appear in multiple tweets with different text case; lower or uppercase. Letter transformation converts each token into a standard format to avoid duplicate feature, herein we converted all the tokens into lowercase.

Stop Words Removal

Tweets contain many unrelated words that lead to an increase in the dimensionality of features. This increase in dimensionality increases computational complexity of classification models. Few terms are entirely useless in tweets for computation algorithm i.e., “the”, “an”, “is”, “am”, “how”, “to”, etc. It is a set of frequent words that carry less important meaning. Such tokens are removed because of non-informative and unnecessary increase in training time and memory overhead.

Removing Irregular Terms

While posting tweets many user tag URL links and images with the combination of text, which also becomes part of tweet message. Here, we only focus on written content; therefore, we removed all the contents which are in uneven form.

Stemming

Due to grammatical reason, people use a different inflected form of words such as kill, killed, and killing. A stemming process is applied for token normalization where the goal is to reduce inflectional form of words to a common base form. Stemming removes affixes from a word and uses the beginning of the word to represent a common base form. For example, the stem of study, studies, and studying is *studi*.

3.1.3. Feature Extraction

Sentiment analysis on text usually requires hand written feature derived from word-level (e.g., airstrike, terrorist), word-level n -gram (e.g., doing_good, good_job), character level n -gram (e.g., b, be, beh, av, ave beha, behave), POS tags (e.g., noun, adjective, verb), word cluster (e.g., maybe, probably, prob collapse to the same cluster), hashtag (e.g., #Afghanistan, #Trump), emoticon (e.g., ☺, ☹), user tags (e.g., @Trump), abbreviations (e.g., WTH, ASAP, ROFL) and elongated words (e.g., yummy, hurrah). Machine learning algorithm needs a numerical picture in the form of a feature vector that enables the model to perform mathematical and statistical investigation. In feature building phase, text data requires to be converted into a manageable representation that is understandable by the algorithm. For this purpose, a feature vector is constructed (i.e., Term Frequency-Inverse Document Frequency) where weighting scheme is applied in order to calculate the score of each token in the corpus. It is defined as:

$$\text{Term Frequency (TF)} = \frac{\text{frequency of term } t \text{ in a tweet}}{\text{total terms in tweets}} \quad (1)$$

$$\text{Inverse Document Frequency (IDF)} = \frac{\text{total tweets}}{\text{number of tweets having a term } t} \quad (2)$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, d) \quad (3)$$

Here, t represents a term in a tweet and d represents a tweet (usually termed as document in text documents).

3.2. Exploratory Data Analysis (EDA)

EDA is known as a way of exploring dataset characteristics by computing basic statistical summary. The statistical summary about the dataset(s) is often used in combination with data visualization techniques in order to have better understanding of the dataset. EDA helps users to think beyond applying classical classification modelling algorithms. By applying EDA, we can have more detailed insights into the data with meaningful information that can also help to define/refine a hypothesis.

In our analysis, we propose to apply Principal Component Analysis (PCA) as an EDA with an objective to transform a high-dimensional data space onto a low-dimensional space (usually 2-D or 3-D). Transformation of 2-D or 3-D through PCA will be useful to get a visualization plot that helps to improve our understanding of the dataset through visual observation in order to find natural groupings in the data. The problem with a tweet-based dataset and other natural text datasets is that it has several features (i.e., terms) and visualizing such high-dimensional features is not viable through classical information visualization tools, so we opted to use PCA to get transformed low-dimensional space in order to plot data on a scatter plot.

Quite often PCA based dimensionality reduction methods are also used to get a transformed space of more than three-dimensions for datasets having dimensions from a few hundred to a few

thousand. In such cases, dimensionality is reduced before applying classification model to reduce complexity of classification model by giving a small set of transformed features.

3.3. Classification Algorithm

Tweet classification is the task of assigning a tweet (T) to one of a set (R) pre-defined class where $R = \{R_1, R_2, R_3, \dots, R_N\}$. Classification tasks are performed by supervised machine learning, where a supervised learning algorithm is used to train the classifier. Normally a set of N training sample is provided to learning algorithm $\{(T_i, R_i) : i = 1, 2, \dots, N\}$ from which it crops a function $F : T \rightarrow R$ that map tweets to classes. Here T_i represents the i th training tweet and R_i is the corresponding class label of T_i .

Choosing a good classifier is an important task to build a robust state-of-the-art predictive ability. In this work various machine learning algorithms; Support Vector Machine (SVM) [43], naïve Bayes' [44], decision tree [45], random forest [46], KNN [47] and ensemble classification methods (with bagging and boosting) [48], etc. are considered. The effectiveness of all these algorithms is demonstrated in the next section for the analysis of our Twitter-based dataset.

3.4. Performance Evaluation

Some of the metrics that are used to evaluate the effectiveness of the classifiers are discussed in the subsequent sub-sections.

- Accuracy

Accuracy is the proportion of the correct prediction made by the model. In simple term, it is the percentage of all the input samples in the dataset that are correctly classified such as, if the model correctly predicts 45 data samples out of 50, the accuracy of the model would be 90%.

$$\text{Accuracy} = \frac{\text{correctly predict sample}}{\text{total samples}} \times 100 \quad (4)$$

- Precision, Recall, and F-score

Precision can be defined as the ratio of true positive predicted and total positive in the data set and it measures the performances with respect to correct prediction. The greater precision means the less miss predicted hits. The recall is the ability of the classifier to predict as many true positive out of the total expected. The performance will be better as precisions and recall have greater values. F-score is the harmonic mean of recall and precision. F-score describes that how precise is the model by taking a mean of precision and recall.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (5)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (6)$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

4. Experiments

With the aim to validate our hypothesis, we made a comprehensive analysis of the collected dataset. The implementation of the framework done in three folds that includes (a) related data collection and labeling (b) preprocessing and Exploratory Data Analysis (c) classification model development and evaluation on the tweet dataset.

4.1. Data Collection and Preprocessing

A Twitter Streaming API was used to collect tweets from the geo-location of Afghanistan from 2nd April 2018 to 8th April 2018 after Kunduz Madrassa attack by Afghan forces. A list of 60 specific query terms, like *Taliban*, *airstrike*, *peace*, *Afghan*, etc. were used to extract tweets from the profile of users (see Section 3.1.1 for query words details). A total of 7500 tweets were downloaded. Many tweets were repeated in the corpus because of retweet feature of Twitter, and we only considered unique tweets in our dataset. We also filtered out tweets that are not relevant to the context of our problem and they appear in a dataset because of some query words in them but their text is not related to our problem. Finally, the resulting dataset has 3380 tweets for the analysis. For the classification task, the labels were required to be given using most appropriate and accurate approach. As mentioned earlier in Section 3.1.1., automated data annotation is often observed to be not reliable with around 70% error rate when compared with humanly assigned classes for our dataset (see Table 3 for sample labels of tweets using *WordNet*, *SentiWordNet* and manual labelling by humans). The dictionaries assign labels in numerical form where range of scores is between {-ve, 0, +ve}. The -ve and +ve score represent the negative and positive sentiment correspondingly and zero denote to a neutral sentiment. Based on these scores it is insufficient to distinguish between anti or pro entities because both writers use similar words. Therefore, we adopted an approach where five different persons from various walks of life (i.e., a journalist, an academic at the Islamia University of Bahawalpur (IUB) from political science department, an academic from computer science department of the IUB and first two authors of this paper) were asked to label tweets. More than 95% of the tweets were given class labels based on the consensus of all five persons. For the remaining 5% of the tweets, the three or more persons agreed to assign a one class label whereas the remaining one or two persons had different views of class label assignments, so in such cases the class label was assigned based on the majority. The main disadvantage of the involvement of humans to label each tweet is a time taking process for preparing a training set for machine learning models. To fully automate the process of detecting extreme contents, the research community needs to focus on developing methods to automatically label each tweet that not only involves the occurrence of the words in a tweet content but also needs to consider the semantics in naturally written text.

Table 3. Example of data labelling comparison.

Tweet	Class Assignment Using WordNet	Class Assignment Using SentiWordNet	Our Approach
Taliban vows ‘serious revenge’ over Afghan airstrike.	<i>pro-Afghan</i>	<i>pro-Taliban</i>	<i>pro-Taliban</i>
UN Human Rights team probing Kunduz airstrike https://t.co/2oSJWlgeb1	<i>Neutral</i>	<i>pro-Afghan</i>	<i>Neutral</i>
Taliban Opened Fire on Civilians after Airstrike https://t.co/MDr3FiNOKG	<i>pro-Afghan</i>	<i>pro-Afghan</i>	<i>pro-Afghan</i>
American trophy hunters have shot and killed more than 78,000 mountain lions over the last three decades #OpFunKill	<i>pro-Afghan</i>	<i>Neutral</i>	<i>Irrelevant</i>
Anti-Taliban movement is speeding up throughout border areas between Afghanistan and Pakistan.	<i>pro-Taliban</i>	<i>pro-Taliban</i>	<i>pro-Afghan</i>

After the labeling process, our dataset was divided into four categories (i.e., pro-Afghan, pro-Taliban, Neutral, Irrelevant) based on the sentiment that it contains. The irrelevant class consists of a few random tweets, which are not related to our specified topic. They are in the corpus due to some query-words, but their tweet text is immaterial with respect to the context. The irrelevant tweets are filtered out from the corpus in further experiments. The description of each class and sub-class in a dataset is presented in Table 4.

Table 4. Description of assigned classes.

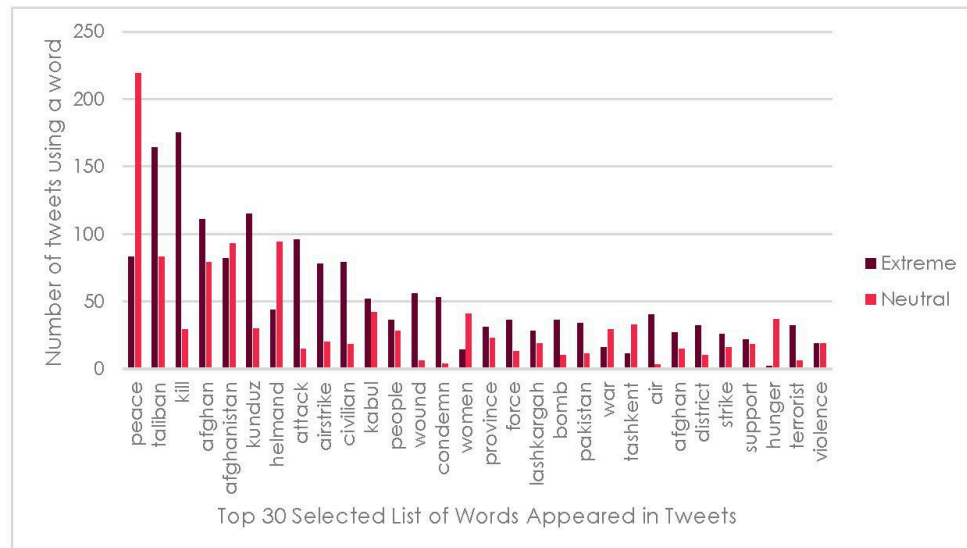
Class	Sub-Class	Description
Extreme	pro-Afghan pro-Taliban	Afghan government supporting accounts or users against Taliban s Taliban supporting accounts or users against the Afghan government
Neutral		a user having a neutral statement about an incident (i.e., user views are not in favor of any party (i.e., Afghan government or Taliban).

4.1.1. Twitter Data Pruning

Data pruning is an essential step, after data crawling, to improve accuracy by making data samples unique and with less noise. Many tweets occurred multiple times in the dataset due to the retweet feature of Twitter. All such tweets are removed by using unique tweet ID assigned to each newly generated tweet. Next, few tweets that are not in English language were also removed. Furthermore, on tweet level all the terms that cannot derive any sentiment are discarded. The task of preprocessing includes removal of duplicate tweets, character case standardization, removal of user mentions and URLs, stop words removal and stemming.

4.1.2. Corpus Analysis

Corpus analysis is performed to understand how often a word is used across tweets. A selected list of words is presented here in Figure 4. The order of the words from left to right is given based on the frequency of their occurrence in tweets with respect to our dataset. The words like [‘Taliban’, ‘kill’, ‘afghan’ and ‘kunduz’, etc] appeared often in the tweets labelled as Extreme whereas the words like [‘peace’, ‘helmand’, ‘Afghanistan’, ‘women’, ‘hunger’, etc.] appeared often in the tweets labelled as neutral.

**Figure 4.** Words distribution.

4.2. Applying PCA

The PCA is considered as a linear method for projecting a high-dimensional dataset onto a low-dimensional space. PCA can also be defined as an orthogonal projection of high-dimensional data to a low-dimensional space in such a way that there will be a maximum variance in the projected data [49–51].

We apply PCA, to our dataset consisting of TF-IDF features extracted from the tweets, to project a high-dimensional data (i.e., TF-IDF features) onto a two-dimensional space. The low-dimensional space is usually considered as 2-D or 3-D with the objective of visualizing it onto a scatter plot to observe

similarity between observed samples (tweets in our case). Figure 5 represents such a visualization plot where two sub-graphs are presented. These sub-figures are similar in terms of data projection but are different in the context of labelling (i.e., colors): The Figure 5a represents a visualization with three classes assigned to observed samples (tweets) whereas Figure 5b represents a graph with two classes. As observed in Figure 5a, a class having *neutral* views (appeared as blue dots) is spread across the whole plot whereas observed samples of classes such as *pro-Taliban* (appeared as red dots) and *pro-Afghan government* (appeared as green dots) are all around the center of the plot with some samples of *Neutral* class in overlapping positions. The result here suggests that tweets having radical/extreme views all share similar characteristics not only in high-dimensional space but also onto a low-dimensional space. This also suggest that using a simple bag-of-words approach, it is difficult to get clear separation in clusters of *pro-Taliban* and *pro-Afghan government* views from tweets. However, it is obvious that tweets having similar keywords but with neutral views are often differentiable than from radical/extreme views. Concluding from the results presented in Figure 5a, we combined *pro-Taliban* and *pro-Afghan government* observed tweet samples as one group (appeared as red dots) in Figure 5b. This Exploratory Data Analysis using PCA explains that it will be easier to classify between neutral and radical/extreme views sharing similar keywords in the tweets. We, therefore, decided to focus in our work to identify tweets either as neutral or as radical/extreme.

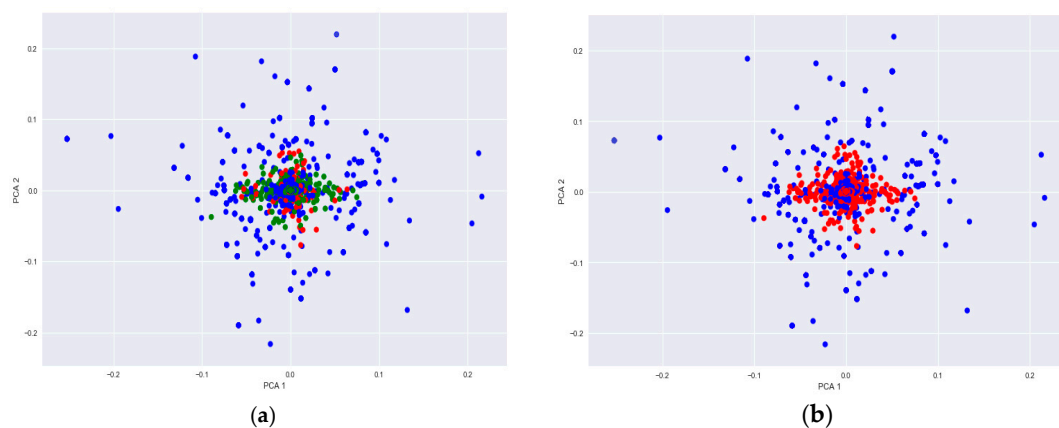


Figure 5. Principal Component Analysis (PCA) Visualization (a) Three class visualization results where blue dots represent Neutral, red dots represent pro-Taliban and green dots represent pro-Afghan government (i.e., anti-Taliban). (b) Two class visualization results where blue dots represent *Neutral* and red dots represent tweets having extreme views (i.e., combining classes pro-Taliban and pro-Afghan government).

PCA is known to be a linear data projection model and attempts to preserve maximum variance in the lower-dimensional space. The variance preservation through PCA is dependent on the type of data in a dataset. It is usually observed that for the dataset having thousands of features, the first few principal components preserve very small variance [26]. Analysis of PCA results for our Twitter data having more than 3000 dimensions (in terms of TF-IDF features), it was observed that projection using first two principal components preserved less than 10% variability suggesting that PCA is not much trustworthy when working with high-dimensional data. Results for our dataset through PCA also suggests that cluster separation between two classes (neutral and extreme) is better than the three classes (i.e., neutral and two subclasses of extreme: pro-Afghan and pro-Taliban). One objective of applying PCA was to get visualization on a 2-D or 3-D scatter plot. We do not present a 3-D plot here, because results were not much different in terms of separation between classes and sub-classes. The other objective of applying PCA was to use it as a preprocess for classification models. The classification model complexity is often directly proportional to the dimensions in the dataset and the PCA-based reduced feature set helps in reducing complexity of classification models. We set a threshold of variance preservation of around 90% and 95% resulting 537 and 678 features respectively.

The classification results presented in the next section are based on a reduced feature set through PCA and on a complete set of TF-IDF features related to tweets.

4.3. Classification Model Building

We opted to apply various prominent classification algorithms for training purposes that best suit text classification. The dataset consists of TF-IDF based features computed from n -gram terms in the tweets and all the tweets have pre-assigned class labels. Seven different classifiers are considered; SVM, naïve Bayes', decision tree, random forest, KNN, and ensemble methods (with bagging and boosting). Each classification model is trained in a supervised fashion using tweets that are labeled either extreme or neutral.

The ensemble classification techniques (with bagging and boosting) are applied to build a strong model by combining a set of weak learners to achieve better performance. Where, individual learners are trained on the same dataset and the prediction task is performed using a combination approach. In order to get better predictive ability, individual members of the classifiers should be accurate and diverse. We, therefore, used different learning algorithms in our ensemble model that were SVM, decision tree, and naïve Bayes', etc. All experiments were performed using complete set of features extracted from our Twitter dataset that consists of TF-IDF feature set and PCA-based transformed reduced features obtained through the original TF-IDF features in order to find the most effective algorithms in terms of classification accuracy. For the evaluation of the classification model, the data is usually split into training and test set. The training data is used to train model and the test set is used to evaluate the performance of the model. This approach of using one part of data as a training set and the other part of data as a test set does not give reliable accuracy as the accuracy with another test set may vary. We, therefore adopted, to use a K -fold cross-validation method [52,53] where we split data into K parts. In each fold, $K-1$ parts are used for training and the remaining one part is used for testing purpose and this process is repeated K times with varying training and test sets. This helps in getting generalization performance of the classification model for the dataset and more reliable results on an unseen test set. All the classification results presented in this paper are based on 10-fold ($K = 10$) cross-validation though experiments with $K = 5$ were also performed, however, predictive accuracy was not much different.

4.4. Results

In this work, our objective was to classify tweets having extreme or non-extreme (neutral) views. We adopted to apply machine learning approaches to demonstrate the efficacy of classifying tweets having extreme views or not. Firstly, it is important to mention that an automated labelling procedure of text data require improvement in order to get more accurate predictive ability using machine learning methods. This was the reason for us to use human assigned class labels, particularly in the context-based sentiment analysis, as a better viable solution (see Section 4.1).

In terms of analysis, we present results in a two-step approach. The first approach in our analysis is applying PCA, as an EDA tool, which is a key to better understand data and in making/refining hypotheses (see Section 4.2). Applying classification models straightaway without understanding natural patterns in the data seems not to be a sensible way of analyzing data. Though we observed that PCA faced a challenge in terms of capturing variance in the first few dimensions for a very high-dimensional dataset. We believe that this has given us better clue of the hidden patterns in the data. The visualization through PCA for our dataset has demonstrated a better separation between clusters of neutral and extreme class. But cluster separation between sub-classes of extremes, as pro-Afghan government and pro-Taliban, were not obvious. These visualization results helped us in making hypothesis that using simplest features like TF-IDF and n -gram for tweets dataset can only help to differentiate between tweets either neutral or extreme. Whereas, discrimination between sub-classes of extreme may require some semantic knowledge in order to get discrimination between them.

The second approach in our analysis was to evaluate our hypothesis of identifying tweets having extreme or neutral views. For this purpose, we opted to apply various classifiers like naïve Bayes, SVM, decision tree, random forest, KNN and ensemble classification models (see Section 4.3). The classification performance is presented in the form of evaluation measures like predictive accuracy, precision, recall, F-score. Classification models often face complexity in terms of training process when working with high-dimensional data. Therefore, we apply PCA to reduce the dimensions considering the assumption that the reduced set of features preserves maximum structure (in terms of variance preservation) of the data and this reduced set of features will be used by the classification models. The classification results, in term of evaluation measures, with various set of features (i.e., unigram, bigram, PCA-based reduced feature set) are presented in Table 5. The results illustrate that the proposed framework has encouraging outcomes for radical/extreme content identification and aspect-based sentiment analysis of tweets.

Table 5. Average score on unigram, bigram feature set and Principal Component Analysis (PCA)-based reduced feature set.

Classification Algorithms	Unigram				Bigram				Reduced Feature Set (Using PCA)			
	Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score
Naïve Bayes'	83.16	82.81	80.2	81.48	81.05	79.73	80.84	80.28	68.06	67.87	67.70	67.78
SVM	80.73	80.65	77.21	78.89	84.71	85.18	83.15	84.15	74.10	75.00	73.21	74.09
Decision Tree	76.06	79.56	69.28	74.06	77.43	80.22	71.34	75.52	70.70	76.17	68.77	72.28
Random Forest	70.7	74.04	70.04	71.98	69.56	84.70	50.87	63.56	76.00	78.77	74.69	76.68
KNN	74.78	73.05	72.82	72.93	73.87	70.34	66.29	68.25	72.55	73.85	73.33	73.59
Bagging	83.12	82.21	81.01	81.61	83.66	81.89	80.01	80.94	70.17	73.91	68.44	71.07
Boosting	81.81	80.30	78.56	79.42	84.43	86.29	82.76	84.49	72.52	73.39	71.43	72.40

It can be observed from the results that naïve Bayes' has shown a better result with 83.16% accuracy on unigram features. The SVM has demonstrated an improvement in the predictive accuracy (with 84.71% accuracy) with bigram features in comparison with unigram features. The ensemble techniques (i.e., boosting and bagging) are also comparable due to the high and stable accuracy over n -grams features. All other individual models (decision tree, KNN, random forest) have maintained less than 80% accuracy. Similarly, all the classifiers are also tested on a reduced feature set obtained through PCA. We present here the results obtained using a reduced feature set preserving 95% variance in the dataset. The experiments were also performed with a reduced feature set ensuring 90% variance in the dataset, but the classification predictive accuracy was not much different. PCA helped to reduce the high dimension features into low dimensions to minimize the overall analysis time required by the classification model training process. The outcome of classifiers on these reduced features is also depicted in Table 5 reporting the random forest by achieving highest accuracy (76%). The results indicate that PCA-based reduced feature set has shown compromised performance of classifiers compared with complete feature set.

Overall, the results of our experiments demonstrate that SVM and ensemble techniques are more stable and accurate to identify extremism from the content of the Twitter dataset. From a feature sets like TF-IDF from uni-gram and bi-gram features and PCA based reduced feature set, the TF-IDF computed from bi-gram features was observed to be more effective with an average accuracy of 84%.

4.4.1. n -Gram Comparison

It is evident from Figure 6 that bigram provides a good stable performance compared with a unigram and trigram regarding the ability to capture the sentiment expressed. On the other hand, the bigram signifies an optimal result to carry out the sentiment analysis of tweets concerning extreme/radical identification. It is obvious from Figure 6 that naïve Bayes' and decision tree vary with the value of ' n ' as compared to SVM, bagging, boosting and KNN, etc.

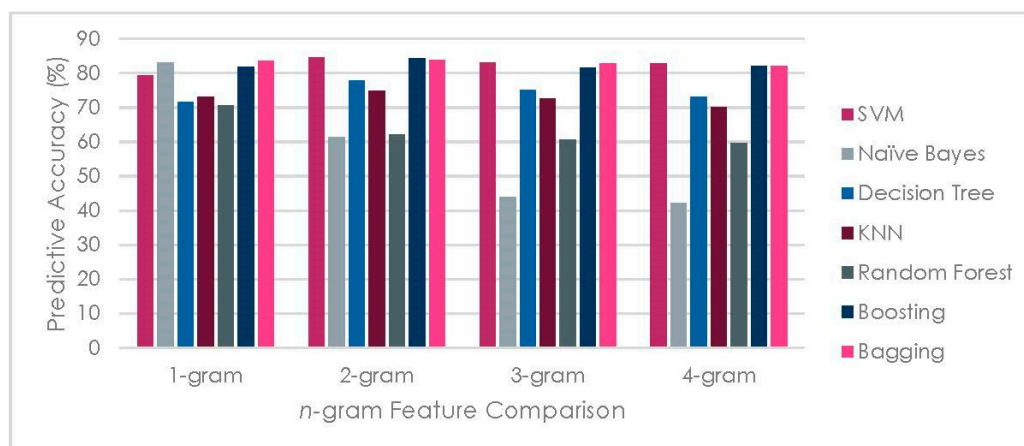


Figure 6. *n*-Gram impact on classification model accuracy.

4.4.2. Data Size Impact on Classification Model Performance

The effect of varying data size performance in terms of accuracy is depicted in Figure 7. Three subsets of the dataset (having 400, 600 and 1000 tweet samples in each subset respectively) were generated to measure the performance of classifiers. It was observed that an increase in tweet observations resulted in gradual improvement in predictive ability of classifiers.

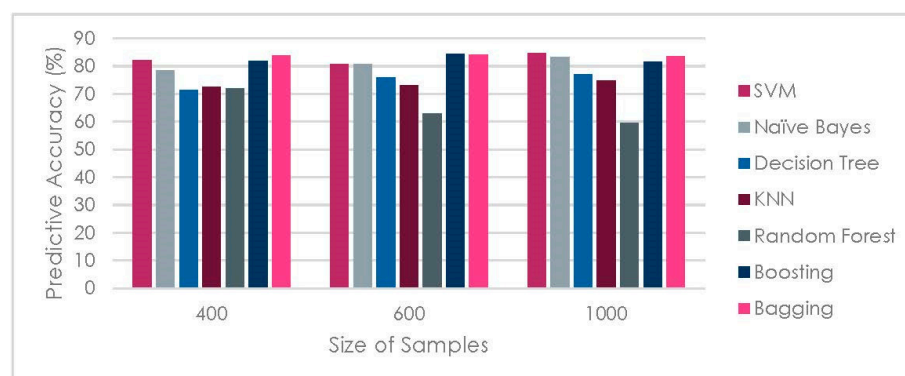


Figure 7. Impact of Sample size on classification model accuracy.

5. Conclusions and Future Work

The wide-spread use of social media has heavily impacted the lives of people in communities. Nowadays, extremist organizations often use social media to disseminate their viewpoints to a larger community; either to generate sympathy for their cause or to recruit people. In this work, our objective is to use social media website Twitter-based tweets data in combination with machine learning approaches to automatically identify user tweets having extreme contents. Many such approaches have been proposed in the literature and they often were able to achieve predictive accuracy of around 80% [10,14–18]. Most of the earlier reported work is related to ISIS and to our knowledge, this is the novel reported research work in the context of the Afghanistan war zone to predict tweets having extreme and neutral contents. The analysis involves processing tweets data to generate TF-IDF features extracted from 1-g, 2-g, 3-g, etc., and PCA-based reduced features. A two-step analysis was performed: Exploratory Data Analysis (EDA) and Classification Modelling. In terms of EDA, we highlighted the importance of an exploratory data analysis in defining a hypothesis that can be useful in getting better predictive ability of the underlying predictive classification problem. We also demonstrate that simplest extracted features from natural language processing domain can help to identify between extreme and neutral tweet content but are not good enough to differentiate between sub-classes of

extreme (i.e., pro-Afghan government or pro-Taliban). In a classification modelling process, various classification models were applied and SVM classification model has shown a predictive accuracy 84% using TF-IDF features extracted from bi-gram features of tweets.

The analysis suggests that in order to get better predictive ability in terms of extreme sub-groups, the tweets content semantic knowledge is required in the analysis. We plan to include semantic knowledge in our future work for this purpose. This is important because many tweets contain similar words but have a different context or even opposite opinions. In aspect-based sentiment analysis, the semantics are required to be included to reduce such limitations.

Author Contributions: Data curation, W.S.; Formal analysis, S.M.; Funding acquisition, G.S.C.; Investigation, S.M. and T.A. and G.S.C.; Methodology, S.M. and W.S.; Project administration, S.M. and G.S.C.; Software, W.S. and O.R.; Supervision, S.M. and G.S.C.; Writing—original draft, W.S. and S.M.; Writing—review & editing, S.M., Z.S., T.A. and M.H.

Funding: This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program. No.10063130, Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1A2C1006159), and MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Promotion), and the 2018 Yeungnam University Research Grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Statista. Number of Monthly Active Facebook Users Worldwide. The Statistics Portal. 2015. Available online: <http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/> (accessed on 10 May 2019).
2. Dogtiev, A. YouTube Statistics. 2018. Available online: <http://www.businessofapps.com/data/youtube-statist> (accessed on 3 April 2019).
3. Clement, J. Twitter Monthly Active User Worldwide 2010 to 2019. The Statistics Portal. 2019. Available online: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (accessed on 3 April 2019).
4. Kats, R. Twitter Gains More Followers. 2017. Available online: <https://www.emarketer.com/Article/Twitter-Gains-More-Followers/1015764> (accessed on 4 April 2019).
5. Clement, J. Number of Social Network Users WorldWide from 2010 to 2021 (in billions). 2019. Available online: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> (accessed on 3 April 2019).
6. Scrivens, R.; Davies, G.; Frank, R. Searching for Signs of Extremism on the Web: An Introduction to Sentiment-Based Identification of Radical Authors. *Behav. Sci. Terror. Political Aggress.* **2018**, *10*, 39–59. [CrossRef]
7. Bowman-Grieve, L. Exploring Stormfront: A Virtual Community of the Radical Right. *Stud. Confl. Terror.* **2009**, *32*, 989–1007. [CrossRef]
8. Sageman, M. *Leaderless Jihad: Terror Networks in the Twenty-first Century*; University of Pennsylvania Press: Philadelphia, PA, USA, 2011.
9. Seib, P.; Janbek, D.M. *Global Terrorism and New Media: The Post-Al Qaeda Generation*, 1st ed.; Routledge: London, UK, 2010.
10. Azizan, S.A.; Aziz, I.A. Terrorism Detection Based on Sentiment Analysis Using Machine Learning. *J. Eng. Appl. Sci.* **2017**, *12*, 691–698.
11. Yadron, D. Twitter Deletes 125,000 ISIS Accounts and expands anti-Terror Teams. Available online: <https://www.theguardian.com/technology/2016/feb/05/twitter-deletes-isis-accounts-terrorism-online> (accessed on 15 May 2019).
12. Chalothorn, T.; Ellman, J. Using SentiWordNet and Sentiment Analysis for Detecting Radical Content on Web Forums. 2012. Available online: http://nrl.northumbria.ac.uk/13075/1/1____Chalothorn_Ellman_SKIMA_2012.pdf (accessed on 10 May 2019).

13. Choi, D.; Ko, B.; Kim, H.; Kim, P. Text Analysis for Detecting Terrorism-Related Articles on the Web. *J. Netw. Comput. Appl.* **2014**, *38*, 16–21. [[CrossRef](#)]
14. Mei, J.; Frank, R. Sentiment Crawling: Extremist Content Collection through a Sentiment Analysis Guided Web-Crawler. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 15), Paris, France, 25–28 August 2015; pp. 1024–1027.
15. Scrivens, R.; Davies, G.; Frank, R.; Mei, J. Sentiment-Based Identification of Radical Authors (SIRA). In Proceedings of the 15th IEEE International Conference on Data Mining Workshops (ICDM Workshops 2015), Atlantic City, NJ, USA, 14–17 November 2015; pp. 979–986.
16. Abbasi, A.; Chen, H. Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intell. Syst.* **2005**, *20*, 67–75. [[CrossRef](#)]
17. Etudo, U.O. Automatically Detecting the Resonance of Terrorist Movement Frames on the Web. Ph.D. Dissertation, Virginia Commonwealth University, Richmond, VA, USA, 2017.
18. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 03–07 April 2017; pp. 759–760.
19. Benigni, M.C. Thesis Proposal: Online Extremist Community Detection, Analysis, and Intervention Ph.D. Dissertation Proposal. 2016. Available online: <https://www.semanticscholar.org/paper/Proposal-%3A-Online-Extremist-Community-Detection-%2C-%2C-Benigni/6461e93348367ee8c5443ece80a46720fe364a7a> (accessed on 10 May 2019).
20. Ferrara, E. Contagion Dynamics of Extremist Propaganda in Social Networks. *Inf. Sci.* **2017**, *418–419*, 1–12. [[CrossRef](#)]
21. Xie, U.; Xu, J.; Lu, T.C. Automated Classification of Extremist Twitter Accounts using Content-Based and Network-Based Features. In Proceedings of the 2016 IEEE International Conference on Big Data, Washington, DC, USA, 5–8 December 2016; pp. 2545–2549.
22. Bermingham, A.; Conway, M.; Mcinerney, L.; Hare, N.O.; Smeaton, A.F. Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. In Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009), Athens, Greece, 20–22 July 2009; pp. 231–236.
23. Berger, J.M.; Morgan, J. The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter. *Brook. Proj. Us Relat. Islamic World* **2015**, *30*, 20.
24. Wadhwa, P.; Bhatia, M.P.S. Classification of Radical Messages on Twitter using Security Associations. In *Case Studies in Secure Computing Achievements and Trends*; Auerbach Publications: Boca Raton, FL, USA, 2014; pp. 273–294.
25. Santos, C.D.; Gatti, M. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 69–78.
26. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
27. Vidino, L.; Hughes, S. *ISIS in America: From Retweets to Raqqa*; Program on Extremism, The George Washington University: Washington, DC, USA, 2015.
28. Weimann, G. Terrorist Groups Recruiting Through Social Media. *CBC News*. 10 January 2012. Available online: <http://www.cbc.ca/news/technology/terrorist-groups-recruiting-through-social-media-1.1131053> (accessed on 30 June 2019).
29. Samuelson, K. British Prime Minister Blaming Social Media on a Terrorist Attack. *Times* (a news website). 2017. Available online: <https://time.com/4804640/london-attack-theresa-may-speech-transcript-full/> (accessed on 10 June 2019).
30. Ting, I.H.; Chi, H.M.; Wu, J.S.; Wang, S.L. An Approach for Hate Groups Detection in Facebook. In *The 3rd International Workshop on Intelligent Data Analysis and Management*; Springer: Dordrecht, The Netherlands, 2013; pp. 101–106.
31. Kwok, I.; Wang, Y. Locate the Hate: Detecting Tweets against Blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*; AAAI Press: Bellevue, WA, USA, 14–18 July 2013; pp. 1621–1622.
32. Scanlon, J.R.; Gerber, M.S. Automatic Detection of Cyber-Recruitment by Violent Extremists. In *Security Informatics*; Springer: Berlin, Germany, 2014; Volume 3, pp. 1–10.

33. Ashcroft, M.; Fisher, A.; Kaati, L.; Omer, E.; Prucha, N. Detecting Jihadist Messages on Twitter. In Proceedings of the IEEE 2015 European Intelligence and Security Informatics Conference (EISIC 2015), Manchester, UK, 7–9 September 2015; pp. 161–164.
34. Agarwal, S.; Sureka, A. Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter. In *International Conference on Distributed Computing and Internet Technology*; Springer: Cham, Switzerland, 2015; pp. 431–442.
35. Devyatkin, D.; Smirnov, I.; Ananyeva, M.; Kobozeva, M.; Chepovskiy, A.; Solovyev, F. Exploring Linguistic Features for Extremist Texts Detection (on the Material of Russian-speaking Illegal Texts). In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 188–190.
36. Kaur, A.; Saini, J.K.; Bansal, D. Detecting Radical Text over Online Media using Deep Learning. *arXiv* **2019**, arXiv:1907.12368.
37. Borcard, D.; Gillet, F.; Legendre, P. Exploratory Data Analysis. In *Numerical Ecology with R*; Springer: Cham, Switzerland, 2018; pp. 11–34.
38. Camacho, J.; Perez-Villegas, A.; Rodríguez-Gómez, R.A.; Jiménez-Mañas, E. Multivariate Exploratory Data Analysis (MEDA) ToolBox for Matlab. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 49–57. [\[CrossRef\]](#)
39. Sommer, S.; Schieber, A.; Hilbert, A.; Heinrich, K. Analyzing Customer sentiments in Microblogs—A Topic-Model-based Approach for Twitter Datasets. In Proceedings of the Americas conference on information systems (AMCIS), Detroit, MI, USA, 4–8 August 2011.
40. Saif, H.; Fernandez, M.; He, Y.; Alani, H. Evaluation Datasets for Twitter Sentiment Analysis: A Survey and A New Dataset, the STS Gold. In Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media, Turin, Italy, 3 December 2013; pp. 9–21.
41. Lundberg, J.; Nordqvist, J.; Matosevic, A. On-the-fly Detection of Autogenerated Tweets. *arXiv* **2018**, arXiv:1802.01197.
42. Ran, C.; Shen, W.; Wang, J. An Attention Factor Graph Model for Tweet Entity Linking. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1135–1144.
43. Guenther, N.; Schonlau, M. Support Vector Machines. *Stata J. Promot. Commun. Stat. Stata* **2016**, *16*, 917–937. [\[CrossRef\]](#)
44. Al-Aidaroos, K.M.; Bakar, A.A.; Othman, Z. Naive Bayes Variants in Classification Learning. In Proceedings of the IEEE International Conference on Information Retrieval & Knowledge Management, Shah Alam, Malaysia, 17–18 March 2010; pp. 276–281.
45. Kingsford, C.; Salzberg, S.L. What are Decision Trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
47. Kramer, O. K-Nearest Neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*; Springer: Berlin, Germany, 2013; pp. 13–23.
48. Rokach, L. *Pattern Classification using Ensemble Methods*; World Scientific Publishing Co., Inc.: River Edge, NJ, USA, 2010.
49. Mumtaz, S. Visualisation of Bioinformatics Datasets. Ph.D. Dissertation, Aston University, Birmingham, UK, 2015.
50. Jolliffe, I.T.; Cadmía, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Ait-Sahalia, Y.; Xiu, D. Principal Component Analysis of High Frequency Data. *J. Am. Stat. Assoc.* **2019**, *114*, 287–303. [\[CrossRef\]](#)
52. Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2010**, *4*, 40–79. [\[CrossRef\]](#)
53. Wong, T.T.; Yeh, P.Y. Reliable Accuracy Estimates from k-fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2019**. [\[CrossRef\]](#)

