Abusive Tweets Detection Using Supervised Learning With Contextual Features

By

Kamal Hussain

Department of Computer Science
Quaid-i-Azam University
Islamabad, Pakistan
January, 2020

Dedicated to

My grandmother, father and elder brother Azhar and my supervisor Dr. Rabeeh for their tremendous sacrifices in making my future bright

Declaration

I hereby declare that this dissertation is the presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly with due reference to the literature and acknowledgment of collaborative research and discussions.

This work was done under the guidance of Dr. Rabeeh Ayaz Abbasi, Department of Computer Sciences, Quaid-i-Azam University, Islamabad.

Date: 30th Jan, 2020

Student Name

Abstract

Social media are one of the most drastically growing platforms on the web which share and generate content in real-time. Among a diverse set of social media platforms, Twitter is the most widely used microblogging platform used to share millions of statuses every minute. Twitter enables people to publicly share short messages called tweets. Occasionally, some people use abuse in their tweets to offend others. It has severe consequences for public in general and targeted victims in particular. Many approaches have been proposed to detect abuse on Twitter which use content and lexicon based features. In this thesis, we propose an approach which uses contextual features including time window and sliding window. We use these features in various supervised machine learning algorithms and perform a diverse set of experiments, including various combinations of features used with a variety of supervised machine learning algorithms. We compare the results of the proposed approach with state-of-the-art methods. The results show that the proposed approach outperforms the state-of-art methods.

Acknowledgment

I courteously express my profound gratitude to Allah Almighty for His countless bounties on me. I also thankful to Holy Prophet Hazrat Muhammad (PBUM), for guiding the righteous and pious life. At first, I express my gratitude to my supervisor Dr. Rabeeh Ayaz Abbasi for giving me an opportunity to work under his kind supervision.

Secondly, I extremely grateful to all my respected teachers for their worthy guidance especially Dr. Onaiza Maqbool, Dr. Shuaib Karim, Dr. Ghazanfar Farooq, Dr. Muddassar Azam Sindhu, Dr. Muhammad Usman, Dr. Umer Rashid and Dr. Akmal Saeed Khattak. I am also thankful to the staff members, especially Shabbir Sahib (ChaChu), Mubashir Bhai and Aurangzeb.

I express thanks to my class fellows and friends especially Mehran Yousaf Malik and Sher Ali for their moral support and encouragement. Moreover, I extend my thankfulness to my senior fellows Zafar Saeed, Naveed Tariq, Khawaja Bilal, Nouman Khan and Javad Ali.

I also express my deepest gratitude to my family for their prayers in my success. My father, grandmother, elder brother Azhar Hussain and Uncle Engr. Muhammad Ali deserve special appreciation for their moral and financial support.

Thanks and Regards,

Kamal Hussain

Contents

List of Tables

List of Figures

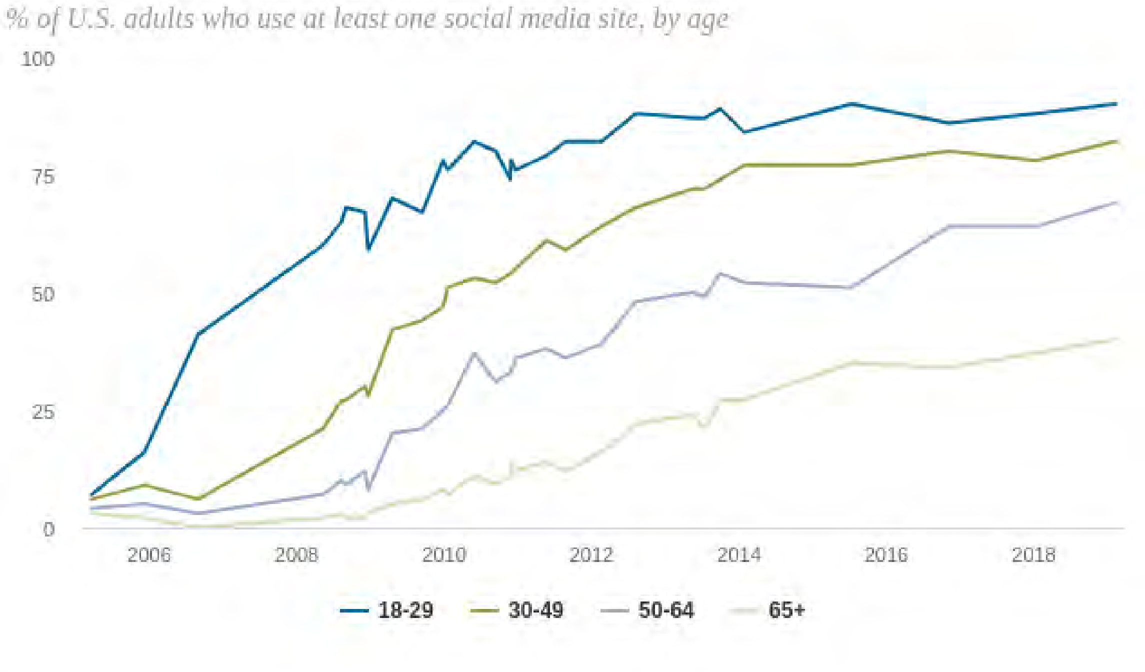
Chapter 1

Introduction

Social media constitute different platforms which are designed to share and generate contents by common people. The contents have various kinds of information, events, and opinions which are shared in real-time. In these platforms, users interact among their friends, family, and other users. These platforms can share content among different users simultaneously. There are various forms of social media platforms including the social networks (Twitter, Facebook and LinkedIn), media sharing networks (Instagram, Snapchat and YouTube), discussion forums (Reddit, Quora and Digg) and so on [Dollarhide, 2019].

Globally, during the last decade, the use of social media widely spread over the world. In every minute, 510,000 and 350,000 comments are posted on Facebook and Twitter respectively [Gaydhani et al., 2018]. According to a PEW research center survey, only 5% adults used social media in 2005. In 2019, 72% of adults used one of the social media platforms in the USA [PEW, 2019]. According to this survey, social media usage is increasing among all age groups as shown in Figure 1.1.

Twitter is one of the most popular social media platform which is used to share small posts (containing upto 240 characters). These posts are called tweets. Twitter provides real-time data through Twitter Application Program Interface (API). Twit-



Source: Surveys conducted 2005-2019.

Figure 1.1: Survey: Social media use by age [PEW, 2019]

ter APIs are available publicly and are easy to get access. Owing to the worldwide use of this platform, it has attracted the attention of researchers to investigate different aspects observed on Twitter.

The excessive use of Twitter generates millions of tweets every minute. Due to the increased use of Twitter, the users are interconnected with different backgrounds and cultures. But, a few users misuse this platform and target other users in the form of abuse, harassment, hate, bullying and trolling, etc., [Duggan, 2017].

Furthermore, the researchers have been investigating these issues. Accordingly, our focuses are on abusive language, which is one of the raising problems on Twitter. Abusive language has been defined as an insult, profanity, vulgarity and violent speech to targets other [Lee et al., 2018b]. Due to the high growth of the users, million of tweets are posted every minute [Gaydhani et al., 2018]. These days, abusive language is frequently growing on Twitter. Thus, some users utilize abusive language on Twitter excessively in the form of individuals or groups. Groups target religion, communities and cultures [Waseem et al., 2017].

The abusive language on Twitter has been detected in varying contexts including the abusive tweets [Nobata et al., 2016, Lee et al., 2018b, Davidson et al., 2017], behavior [Founta et al., 2018a, Chatzakou et al., 2017b] and users [Abozinadah and Jones, 2017, García-Recuero et al., 2018]. The focus of this research is abusive tweets detection. In existing approaches, the researchers have identified the abusive tweets in different aspects, which are based on abusive words [Wiegand et al., 2018, Gitari et al., 2015], abusive contents [Davidson et al., 2017, Lee et al., 2018b] and abusive contexts [Ribeiro et al., 2018b, Fehn Unsvåg and Gambäck, 2018]. First, the abusive word-based approach is the simplest way to detect abusive language. This approach is based on the lexicon, lexicon based on the sentiments and negative polarity expressions. Lexicon based approaches use a fixed set of words which might not work in every scenario, therefore researchers proposed content-based approaches.

Content-based approaches have been used using different features and algorithms to detect the abusive tweets. In this approach, the studies have used different Natural Language Processing (NLP) features which include linguistic features, semantic features, syntactic features and statistical features. These features support the detection of abuse with high accuracy. At last, the contextual features further improve the detection of abusive language. The contextual features are additional information about tweets and users that help the identification. Some of these features show the popularity of users and their tweets. For example, retweets, favorites, followees and followers. Some other tweet features are also used based on tweets content but used as context features. For example, mentions, hashtags, URLs, word length and bad-word. Moreover, a few binary features have been used in contextual features including verified account, is tweet a reply and is tweet a status, etc.

In literature, the detection of abusive language has been evaluated by different techniques including machine learning, deep learning, social network analysis and other methods. However, machine learning and deep learning algorithms are being used more frequently. In machine learning, both supervised and unsupervised learning algorithms have been used for the detection of abusive language. In supervised learning, several classification techniques have been used which include SVM, Random Forest, Naïve Bayes and Logistic Regression, Decision Tree and K-Nearest neighbor, etc. The unsupervised learning algorithms used for the detection of abusive language which include clustering, similarity and rule-based methods. The deep learning techniques have also been used to detect abusive language including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Social network methods have been used to check the relationships between users and other entities like words. The network methods are also used as features that include centrality measures, page rank and graph traversal based features.

1.1 Motivation

Abusive language is a major problem growing on Twitter. Unfortunately, a few users target others based on race, religion, culture and gender. Those users spread abusive language on Twitter using hateful speech, cyberbully and harassment [Waseem et al., 2017]. Statista conducted a survey to identify the entity responsible for abusive behavior on Twitter. In this survey, they have taken the opinion of different users. The results of survey are shown in Figure 1.2. Results show that 48% of responsibilities are on both users and platforms, 41% on users and 8% on social media platforms [Diwanji, 2018]. The results of the survey suggest that it is mainly the users who post abusive contents. Therefore in this thesis, we are interested in developing methods to identify abusive content.

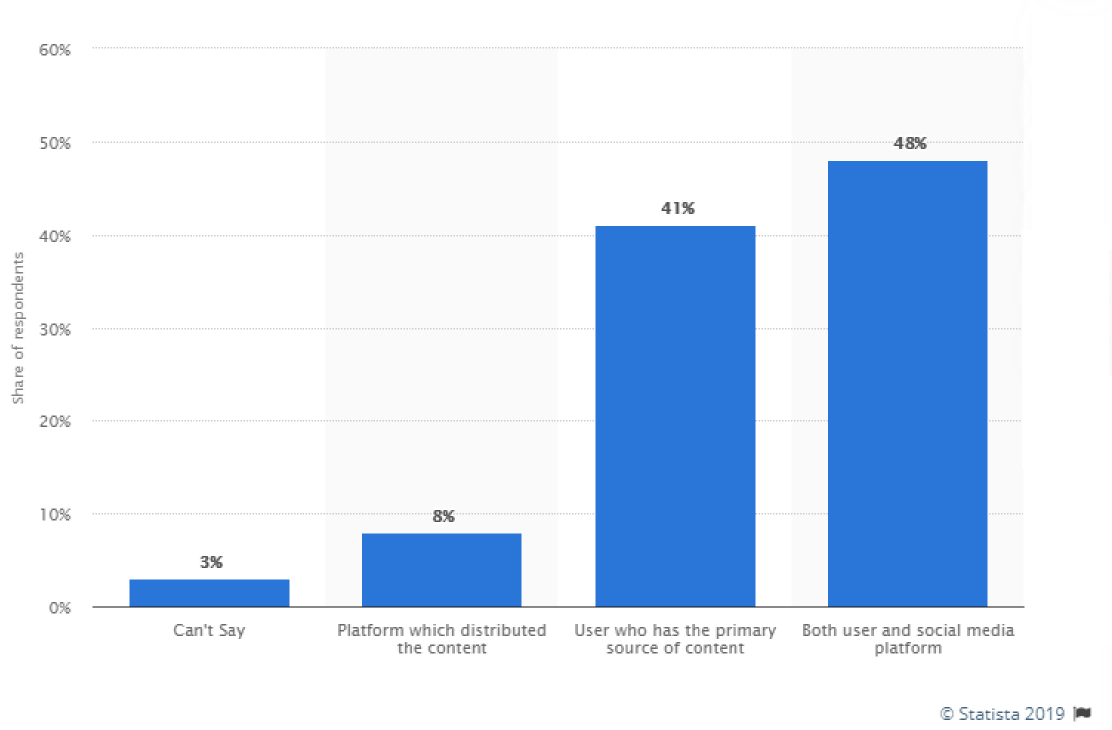


Figure 1.2: Survey: Users and platforms [Diwanji, 2018]

Recently, the researchers have been attracted to detect and investigate the detection of abusive language on Twitter. Limited research employs the context of tweets for detecting abuse, particularly those which use time as a context. Therefore study the temporal features for detecting abusive contents.

We have explored the detection of abusive language based on contextual features by using various features such as tweet features, user features and window-based features. In these features, we focus on the combination of important factors of the tweet- and user-based features.

1.2 Research Questions

As per literature review, we have formulated two questions:

- RQ1: Do contextual features perform better than the content-based features for the detection of abusive tweets?
- RQ2: Which combinations of the features are most significant for accurately detecting of abusive tweets?

1.3 Research Contributions

The main contributions of this research are:

- Using combination of user-based, content-based and window-based features, and finding the best combination of these features.
- Measuring the performance of a variety of classification algorithms and ensemble methods.
- Comparing the results of individual features with each other and the proposed method with the state-of-the-art methods used for detecting abusive tweets.

1.4 Thesis Organizations

Chapter 2 presents the literature review. First, we discuss the methods of abuse detection which include abusive language, abusive user and abusive behavior detection. We focus on the abusive language detection which has been performed under different aspects, including keyword-based, content-based and context-based detection. Afterwards, we also describe the algorithms and methods which include machine learning, deep learning, social network analysis and other methods. We describe the datasets used in literature and also discuss the evaluation measures.

Chapter 3 discusses the dataset creation process. We extend a commonly used existing dataset for abusive language detection. We extend the dataset by collecting additional tweets to expand the contexts of existing tweets. This chapter also describes the data processing and filtering. Lastly, the chapter presents the data statistics.

Chapter 4 categorizes into two parts. First, the proposed using content-based and context-based features is discussed. We propose three context-based features which include tweet features, user features and window-based features. The supervised machine learning algorithms have been used to evaluate the performance of detecting abusive tweets. We use two types of supervised learning algorithms, which include linear classification and ensembles methods. Second, we describe the experiment process including state-of-the-art methods which are used for comparison with our proposed approach. We also describe the evaluation process, we use different evaluation measures including precision, recall, f1-measure and accuracy.

Chapter 5 describes the results of both the proposed methods and the state-of-the-art. The features of the proposed approach are evaluated individually as well as compared with each other. This follows the discussion of the results of combined features. Lastly, we discuss the comparison of the proposed approach with the state-of-the-art.

Chapter 6 concludes and summarize the thesis. It also discusses of the limitation of our proposed approach and the possible future directions.

1.5 Summary

In this chapter, we have described the overview of social media including its benefits, increasing use, problems and platforms like Twitter. Previous techniques of abusive language detection have been discussed. We have also presented the motivation, research questions and research contributions of this thesis.

Chapter 2

Related Work

In this chapter, we discuss the literature related to abusive language. Social media platforms constitute a major part of today's web. In these platforms, the people connect and interact with each other. Also, users can share information with each other [Saeed et al., 2019, Chakrabarty and Gupta, 2018]. However, these platforms have gained wide spread popularity during recent year. Twitter is one of the most popular social media platforms where the users share contents in form of short messages called tweets [Founta et al., 2018a]. As the platform gained its popularity, some people started to share negative tweets. The negativity included abuse, cyberbully, troll, hateful speech and fake news among many others [Chatzakou et al., 2017b, Founta et al., 2018a, Ribeiro et al., 2018b].

The severity of negativity has attracted researchers to identify negative contents and users posting such contents. Our study is focused on the detection of abusive tweets. Accordingly, the abuse identification on Twitter has been performed based on content, user and behavior detection. Detecting abusive tweets mainly include keyword-based, content or tweets-based and context-based approaches [Nobata et al., 2016, Park and Fung, 2017, Lee et al., 2018b, Fehn Unsvåg and Gambäck, 2018]. Recently, researchers have been using context-based approach for the detection of abusive language [Nobata et al., 2016, Fehn Unsvåg and Gambäck, 2018]. Moreover, these features have been evaluated by using Machine Learning, Deep Learning and some other methods.

We have organized this chapter into following sections: Section2.1 discusses different types of abusive language detection techniques which includes contents-based, users-based and behavior-based detection. Section 2.2. describes the data preprocessing for further processing. Section 2.3 presents features extraction Section 2.3.3 discusses the additional features which are based on content features, context features, network features and temporal features. Section 2.4 presents the algorithms and methods which have been used for the detection of abusive language. Section 2.5 presents the data collection techniques from literature. Section 2.6 presents the evaluation methodologies. Finally, Section 2.8 summarizes this chapter.

2.1 Abusive Language Detection

Recently, researchers have been addressing the problem pf detecting abuse on Twitter. Some users target other users by using offensive language or profanity [Zampieri et al., 2019]. Researchers have focused on detecting abusive content [Nobata et al., 2016, Lee et al., 2018b, Davidson et al., 2017], abusive users [Abozinadah and Jones, 2017, García-Recuero et al., 2018] and abusive behavior [Founta et al., 2018a, Chatzakou et al., 2017b]. In the following sections, we discuss these techniques in detail.

2.1.1 Abusive Contents Detection

In content detection, the algorithms identify abusive tweets posted on Twitter. Abusive content detection algorithms mostly use features from natural language processing (NLP) which include the linguistic features [Clarke and Grieve, 2017,Nobata et al., 2016, Wiegand et al., 2018], semantic features [Chatzakou et al., 2017b, Gitari et al., 2015, Watanabe et al., 2018], syntactic features [Clarke and Grieve, 2017, Chatzakou et al., 2017b, Watanabe et al., 2018] and statistical features [Abozinadah and Jones, 2017, Watanabe et al., 2018]. Recently, other features have been used to detect abusive tweets. These features include context [Tahmasbi and Rastegari, 2018, Pitsilis et al., 2018, García-Recuero et al., 2018] and metadata features [Founta et al., 2018a, García-Recuero et al., 2018, Tahmasbi and Rastegari, 2018]. These features are taken from tweets and users information [Founta et al., 2018a, Tahmasbi and Rastegari, 2018, García-Recuero et al., 2018].

2.1.2 Abusive User Detection

Abusive user detection has been performed based on number the tweets of every user. Moreover, the user profile information has also been used to check the connectivity among other users and abusive behavior for detecting the abusive users [Kwon et al., 2018, García-Recuero et al., 2018, Fehn Unsvåg and Gambäck, 2018]. These connections have been computed through social network analysis techniques [Abozinadah and Jones, 2017, García-Recuero et al., 2018, Fehn Unsvåg and Gambäck, 2018] and similarity measures [Kwon et al., 2018, García-Recuero et al., 2018, Zampieri et al., 2019]. Abusive users consistently share abuse in their posts on Twitter. Mostly, their profile information is not well-managed [Abozinadah and Jones, 2017, Kwon et al., 2018] and limited social relations [García-Recuero et al., 2018]. However, such users frequently target others in the form of bullying [Tahmasbi and Rastegari, 2018], abusing [Kwon et al., 2018, Waseem et al., 2017], offensive [Davidson et al., 2017] and aggressive [Chatzakou et al., 2017b]

2.1.3 Abusive Behavior Detection

The detection of abusive behavior is another type of abuse detection technique which recognizes the behavior of users from their tweets and their connections. The abusive behavior is sometimes based on different situations which promote the abusive tweets by using specific content or through specifics hashtags on Twitter [Chatzakou et al., 2017b, García-recuero, 2016]. The abusive users and normal users can be differentiated from their behavior in real life, but the task becomes challenging in case of Twitter [Founta et al., 2018a]. Abusive tweets portray different behaviors including harassment and aggression [Chatzakou et al., 2017b, Founta et al., 2018a]. Accordingly, these types of users can be blocked and deleted from Twitter because of frequent abusive tweets to others users and also by sharing abusive content [Founta et al., 2018b, Chatzakou et al., 2017b]. For example, the Gamer-Gate controversy on Twitter, which was mainly based on video game culture but hashtag #Gamergate was a beginning of the campaign of harassment, bullying, trolling, abusive language and hateful speech on the Twitter. Many accounts got blocked, suspended and deleted by getting involved in #Gamergate controversy from Twitter [Chatzakou et al., 2017b].

2.2 Data Pre-processing

Pre-processing prepares the data for further processing. The researchers have used several methods to pre-precess the text [Gaydhani et al., 2018, Pitsilis et al., 2018].

The pre-process of text may include the removal of unwanted characters, punctuation, symbols, digits, alphanumeric, stop-words, repeated words or sentences, URLs, mentions, hashtags and negative words spam [Chen et al., 2017, Gaydhani et al., 2018, Pitsilis et al., 2018, Chatzakou et al., 2017b]. In literature, a few researchers have used URLs, mentions and hashtags as features, therefore, these features have not been removed during preprocessing [Chatzakou et al., 2017b].

During pro-processing of text may include the replacement of words. In this process, all the text into lowercase and the abbreviations and the misspelled word are replaced into their expanded form [Gaydhani et al., 2018, Pitsilis et al., 2018]. Stemming and lemmatization methods are used to convert the words into base or root word [Founta et al., 2018a]. There are various methods for stemming and lemmatization which include porter stemmer [Chen et al., 2018, Founta et al., 2018a, Gaydhani et al., 2018], snowball [Founta et al., 2018a] and WordNet lemmatizer [Chen et al., 2018, Clarke and Grieve, 2017].

2.3 Features Extraction

Feature extraction converts the input data into such a form which can be used by a machine learning. In literature, different features are used for the detection of abusive language. These features have been acquired from tweets and user information [García-Recuero et al., 2018]. The tweets information contains the content as well as context or metadata features [Chakrabarty and Gupta, 2018, Chen et al., 2017, García-Recuero et al., 2018, Lee et al., 2018b, Tahmasbi and Rastegari, 2018]. The user features have been used as contextual features which address the relationship between users [García-Recuero et al., 2018]. Network or graph-based features have been used to analyze the connectivity and relationship between users [Founta et al., 2018a, Tahmasbi and Rastegari, 2018].

The content-based from text features are extracted using Natural Language processing (NLP). The content-based features may include lexicon [Lee et al., 2018a, Wiegand et al., 2018] or use content [Lee et al., 2018b, Davidson et al., 2017]. Moreover, the context used to identify the abuse include the tweets, users and network features. These features help the detection more accurately [Chen et al., 2017]. In this section, we discuss the keyword-based, content-based and network-based features in detail.

2.3.1 Keywords-Based Features

The keyword-based approach is the simplest way to detect abusive word from tweets on Twitter. It is also known as lexicon. lexicon is external list of related domains, for example, abusive word list [Lee et al., 2018a]. The keywords based approaches compare lexicon with every word in data using sentiments and polarity expressions [Wiegand et al., 2018, Gitari et al., 2015]. Keyword-based approaches have also been developed using different features including: sentiment score of every word [Wiegand et al., 2018]

The sentiment analysis features have been used for the detection of abusive words which computes sentiment score of every word. These sentiment scores are used in polarity expression, the polarity expression decides to the category of abusive word or non-abusive word. Polarity expression distinct the positive form and negative form of a word. Accordingly, the abusive keywords are identified [Wiegand et al., 2018, Gitari et al., 2015].

Lexicon-based features have been used for the detection of abusive keywords. Every word in the tweet is matched with the abusive words in the lexicon. Using lexicon, overall polarity expression decides whether a sentence is abusive or not [Ribeiro et al., 2018b, Wiegand et al., 2018, Gitari et al., 2015, Choudhury and Breslin, 2010].

The researchers have used Natural Language processing (NLP) techniques which include linguistic [Wiegand et al., 2018] and n-gram [Wiegand et al., 2018, Gitari et al., 2015] for the identification of abusive word.

Moreover, the word embedding mostly used for Deep Learning based approaches. Word embedding transforms the words or tokens into vectors, it can be called data vectorization. These include word2vec, continuous-bag of words and skip-ngram [Lee et al., 2018a, Wiegand et al., 2018]. These features have been summarised in Table 2.1.

Features	Reference
Sentiment	[Wiegand et al., 2018, Gitari et al., 2015, Choudhury and Breslin, 2010]
Lexicon	[Wiegand et al., 2018, Gitari et al., 2015, Choudhury and Breslin, 2010]
Polarity expression	[Choudhury and Breslin, 2010, Gitari et al., 2015, Wiegand et al., 2018]
Pattern	[Wiegand et al., 2018]
Ngram	[Wiegand et al., 2018, Lee et al., 2018a]
Word2vec	[Wiegand et al., 2018, Lee et al., 2018a]
Continious-bag of words	[Lee et al., 2018a]
Skip-gram	[Wiegand et al., 2018]

Table 2.1: Summary of keyword-based approaches

2.3.2 Content-Based Features

Abusive content-based approach on Twitter have been performed by incorporating different features. These features have been extracted using Natural Language Processing (NLP) techniques which include linguistic features [Lee et al., 2018b], syntactic features [Chatzakou et al., 2017b], semantic features [Nobata et al., 2016] and statistical features [Davidson et al., 2017, Lee et al., 2018b]. summaries these features in Table 2.2.

The linguistic features contain several sub-categories to use as features which are text length, average number of words, number of punctuations, question marks, repeated words and punctuations, grammar and spelling mistakes [Lee et al., 2018b]. Syntactic features check the relationship of the words [Lee et al., 2018b, Waseem et al., 2017] and part of the speech (POS)tags [Lee et al., 2018b]. These features make tuples of words and capture the dependency relation between words [Chatzakou et al., 2017b]. Semantic features check the similarity between the words by using different methods [Lee et al., 2018a, Waseem et al., 2017], like: Cosine similarity [Lee et al., 2018a] and Edit distance [Waseem et al., 2017]. In this case, statistical features are used to transform the data into vectors because the requirement of classification.

These features include n-gram [Watanabe et al., 2018], Tf-Idf [Lee et al., 2018b] and word embedding [Founta et al., 2018a].

N-gram has been used at word-level [Watanabe et al., 2018,Fehn Unsvåg and Gambäck, 2018] and character level [Sharma et al., 2018, Lee et al., 2018b]. In literature, the word-level has been used as unigram, bigram and tri-gram and even sometime tetra-gram have also been used [Watanabe et al., 2018, Sharma et al., 2018]. The character-level n-grams have been used mostly in range of 2-8 grams [Sharma et al., 2018]. Tf-idf has been also been used with n-grams to normalize the weight of each word [Lee et al., 2018b, Davidson et al., 2017]. Word embedding is another type of data transformation techniques which has mostly used in deep learning and neural network [Founta et al., 2018a, Lee et al., 2018a]. Word [Chen et al., 2017] and pre-trained [Park and Fung, 2017, Lee et al., 2018b] embedding with different dimension. All above features have been summarized in Table 2.2.

Reference	**Features**
[Lee et al., 2018b, Nobata et al., 2016]	Linguistic
[Nobata et al., 2016]	Semantic
[Davidson et al., 2017,Nobata et al., 2016]	Syntactic (relation)
[Nobata et al., 2016]	Words length/ average
[Nobata et al., 2016]	Punctuations, Spelling mistakes
[Nobata et al., 2016,Davidson et al., 2017]	Part of speeches (POS) tag
[Gaydhani et al., 2018,Nobata et al., 2016, Park and Fung, 2017,Chen et al., 2017,Lee et al., 2018b]	N-gram(Word): range(1 to 4)
[Gaydhani et al., 2018,Nobata et al., 2016, Chen et al., 2017, Lee et al., 2018b]	N-gram (Character): range(3 to 8)
[Nobata et al., 2016, Chen et al., 2017, Davidson et al., 2017, Lee et al., 2018b]	TF-Idf
[Gaydhani et al., 2018]	Word2vec
[Park and Fung, 2017]	Comment2vec
[Chen et al., 2017, Lee et al., 2018b]	GLOVE
[Park and Fung, 2017, Chen et al., 2017]	FastText

Table 2.2: Summary of content-based approaches.

2.3.3 Context-Based Features

Context-based features include additional attributes of tweets which contain the relations between the tweets and user profile information attributes. For example, followers, followees, and retweets and use the relationships among attributes as well. The context features have been used to extract the features from tweets and user information. Recently, the researchers have used context features to detect the abusive tweets which include the tweets features [Chatzakou et al., 2017b, Watanabe et al., 2018], users features [Ribeiro et al., 2018b, Fehn Unsvåg and Gambäck, 2018] and networks features [Pitsilis et al., 2018].

Tweet Features

The tweet features contain the characteristics of tweets information. These features have been categorized into different categories including: hashtags, mentions, URLs, retweets, favorite, replies, tweets word or tweets, tweet words average, syllables and sentiment analysis [Chatzakou et al., 2017b, Watanabe et al., 2018, Ribeiro et al., 2018b]. Summarize these features in Table 2.3

In these features, researchers have used count values of attributes. For example, number of retweets and favorites [Tahmasbi and Rastegari, 2018]. In these features, words [Cecillon et al., 2019] and hashtags [Chatzakou et al., 2017a] are also used as features. For example, @name or #Gamergate and bad-words [Chatzakou et al., 2017b, Ribeiro et al., 2018b].

Ref	**Tweet-based**
[García-Recuero et al., 2018]	Hashtags
[García-Recuero et al., 2018]	Mentions
[Abozinadah and Jones, 2017]	URLs
[García-Recuero et al., 2018]	Retweets
[García-Recuero et al., 2018]	Replies (True/False)

Table 2.3: Summary of contextual features.

User Features

The user features constitute characteristics of user profile. In literature, these features have been used with the combination of user profile information such as followers, friends/followees, favorites, listed, account age, Geographical location, profile URL, profile image, user description, account verified(True/False) [Ribeiro et al., 2018b, Chatzakou et al., 2017a, Fehn Unsvåg and Gambäck, 2018] (See in Table 2.4).

The user features have been mostly used by their count as features. For example, counts of features are the number of followers, etc. [Founta et al., 2018a]. A few approaches used the features like name [Cecillon et al., 2019], text [Cecillon et al., 2019], and tags [Chatzakou et al., 2017a]. For example, geographical location, the screen name of the user and users description [Mathur et al., 2018, Chatzakou et al., 2017b]. These features have been summarized in Table 2.4 In user features, the relationship between users have been checked to find the ratio, difference, and similarity between the connected users [García-recuero, 2016].

Ref	**User-based**
[Fehn Unsvåg and Gambäck, 2018]	Profile mages
[Fehn Unsvåg and Gambäck, 2018]	Profile description
[García-Recuero et al., 2018]	Account verification (True/False)
[Chatzakou et al., 2017a]	followees
[Chatzakou et al., 2017a]	Followers
[García-Recuero et al., 2018]	Favorites
[García-Recuero et al., 2018]	Listed
[García-Recuero et al., 2018]	Account age
[Fehn Unsvåg and Gambäck, 2018]	Geo-Location
[Fehn Unsvåg and Gambäck, 2018]	Profile URLs

Table 2.4: Summary of contextual user-based features.

Network Features

A network, In case of social media, describes the connections and relations between between users or other entities. It shows, how the users are connected with each other? These connections are taken as contextual features in several studies [Cecillon et al., 2019]. The network features have been extracted from the tweet and user features. These features check the connectivity and similarity between users by using social network analysis methods. The Social Network Analysis methods are centrality measure [Founta et al., 2018a, Chatzakou et al., 2017b, Tahmasbi and Rastegari, 2018] and graph based. These features have been summarized in Table 2.5. These features check and identify connection between users by using these centrality measurement methods [Founta et al., 2018a, Chatzakou et al., 2017b]. For example, the number of followers and number of followees connections, number retweets user's connection and connection between followers and followees [Founta et al., 2018a, Ribeiro et al., 2018b, García-Recuero et al., 2018, Fehn Unsvåg and Gambäck, 2018].

Reference	**Network-based**
[Cecillon et al., 2019]	Community
[Cecillon et al., 2019, Founta et al., 2018a]	Centrality
[Founta et al., 2018a, Chatzakou et al., 2017a]	Eigenvector Centrality
[Cecillon et al., 2019]	Degree/strength centrality
[Cecillon et al., 2019, Founta et al., 2018a]	Closeness
[Ribeiro et al., 2018b, Chatzakou et al., 2017a]	Betweenness
[Cecillon et al., 2019, Founta et al., 2018a]	Hub/Authority
[Abozinadah and Jones, 2017]	Page-Rank
[García-Recuero et al., 2018]	Bound-Breadth first search
[García-Recuero et al., 2018]	Similarity

Table 2.5: Summary of network-based contextual features

Temporal Features

Temporal features contain the sequence of features which comprise of a specific time period, for example, tweets posted by a user in the past one hour. These features are based on time sequences of multiple tweets generated during an event. The event detection uses temporal data to find the events. Recently, other features have also been also used for the detection of events by using machine learning and social network analysis [Wang et al., 2016, Saeed et al., 2019, Wang and Goutte, 2017].

Event detection is performed by analyzing the tweets are related to an event. For example, the discussion about any sports, upcoming or recent election and any accidents within specific times [Wang and Goutte, 2017, Saeed et al., 2019]. The detection of event can be performed by using a set of sliding time interval. Temporal tweets have been used as features to detect an event [Saeed et al., 2019, Wang et al., 2016].

2.4 Algorithms and Methods

In literature, the methods and algorithms have been used to detect the abusive language including machine learning, deep learning and other methods. Social network analysis and graph-based features have also been used and discussed in Section 2.3.

2.4.1 Machine Learning

In machine learning algorithms, unsupervised, semi-supervised, supervised and regression have been used for the detection of abusive language. In unsupervised machine learning, the detection of abusive tweets has been using several methods which include clustering [Chatzakou et al., 2017b, Lee et al., 2018b], similarity [Lee et al., 2018a, García-Recuero et al., 2018] and neighborhood [Ribeiro et al., 2018b, Chen et al., 2018]. Several supervised machine learning methods are also used including linear classification [Davidson et al., 2017, Lee et al., 2018b] and ensemble methods [Sharma et al., 2018, Davidson et al., 2017]. However, regression methods have also been used to predict the abusive tweets [ElSherief et al., 2018, García-Recuero et al., 2018].

Supervised Learning

Supervised learning recognizes the object based on input data given to a model. The model is trained using labeled data. Labeled data contains examples from different classes for which the model is trained [Gitari et al., 2015]. It has been used to detect the abusive language by using different features of Natural Language Processing (NLP) and with a few additional features [Chatzakou et al., 2017b]. The supervised learning has been used to detect abusive language with their specifications in Table 2.6. The ensembles methods have also been used for the detection of abuse on Twitter. These methods are also summarized in Table 2.6 with their parameters.

Algorithm	**Parameters**	**Reference**
SVM	Kernel = linear and C = 1	[Abozinadah and Jones, 2017, Davidson et al., 2017]
Logistic Regression	Regulation "penalty= L2" , random number generator "random_state = 42", optimization problem "solver = liblinear" and multi_class =auto	[Park and Fung, 2017, Davidson et al., 2017]
Naïve Bayes	MultimonialNB with additive smoothing and alpha =1.0	[Chatzakou et al., 2017a, Founta et al., 2018a]
Decision Tree	Methods: ID3, C4.5, C5.0, CART, J48, LADTree, LMT and NBTree	[García-Recuero et al., 2018, Fehn Unsvåg and Gambäck, 2018]
Random Forest	Number tree in forest "n_estimators =35" and Random number generato "random_state =42"	[Lee et al., 2018b,Sharma et al., 2018]
AdaBoosting	Number tree in forest "n_estimators =100", sub_sample = 1 and learning_rate =1.0	[Pitsilis et al., 2018, Davidson et al., 2017]
GradBoosting	number tree in forest "n_estimators =35", learning_rate =1.0 and Random number generator "random_state =42"	[Lee et al., 2018b]

Table 2.6: Summary of classifications

Unsupervised and Semi-supervised Learning

The unsupervised learning algorithm is used for unlabeled datasets, it recognizes the patterns through measuring similarity objects [Chatzakou et al., 2017b]. The semi-supervised learning have been used when dataset contains only a few labeled objects and most of the data is unlabeled [Watanabe et al., 2018, Ribeiro et al., 2018b].

The unsupervised learning algorithms have been used for abusive language and hate speech detection. The algorithms have summarized in Table 2.7

Method	**Reference**
Clustering	[Chatzakou et al., 2017a]
K-mean Cluster	[Chatzakou et al., 2017a]
Latent Topic Cluster	[Lee et al., 2018b]
KNN	[Chen et al., 2017]
Cosine similarity	[Lee et al., 2018a]
Edit distance	[Lee et al., 2018a]
Levenshtein distance	[Chen et al., 2017]
Euclidean distance	[Chatzakou et al., 2017a]
Hellinger distance	[ElSherief et al., 2018]
Deregard rules	[ElSherief et al., 2018]
Fuzzy rules	[Sharma et al., 2018]
Jaccard index	[García-Recuero et al., 2018]
Kendell rank	[Founta et al., 2018a]
Correlation coefficient	[Founta et al., 2018a]
Pearson and spearman correlation coefficient	[Founta et al., 2018a]

Table 2.7: Summary of unsupervised learning

Regression Methods

Regression represents the relationship against one dependent attributes with independent attributes [Sharma et al., 2018]. Regression methods have been used to predict the targets and victims of hateful users and abusive tweets [ElSherief et al., 2018] which include the Multivariant regression [ElSherief et al., 2018], Linear regression [ElSherief et al., 2018], Linear binomial regression [ElSherief et al., 2018], Quartile regression [ElSherief et al., 2018], Poisson regression [ElSherief et al., 2018], Negative binomial regression [ElSherief et al., 2018] and Vowpal Wabbit's regression [Nobata et al., 2016]

2.4.2 Deep Learning

Deep learning generally consists of multiple layers of neurons in form of neural networks. They are often used in identification and prediction processes. Recently, deep learning has been used to detect the abusive language. Deep learning use methods of Neural Network with many different layers which include Convolutional Neural Network (CNN) and Recurrent Neural Network for the detection of abusive language by incorporating with their specifications and requirements [Lee et al., 2018b, Zimmerman et al., 2018].

Convolutional Neural Network (CNN)

Recently, CNN have been used for the detection of abusive language which use different dimensions and convolutional layers with activation functions, loss functions, pooling and optimization functions [Lee et al., 2018b, Zimmerman et al., 2018, Zhang and Luo, 2018]. The methods of deep learning summarized in Table 2.8

Mostly the 200 or 300 dimensions and different size of convolutional layers of word-level and character levels of text with activation function, dropout learning rate, avoid over-fitting, loss functions, optimizations function, multi-layer perception and full-connected layer with also used hyper-parameter The parameter are listed in Table 2.8).

Recurrent Neural Network (RNN)

RNN checks the connection of the nodes of the directed graph with a sequence of inputs on input states. This method and its types are bi-directional. There are two different types of methods are used: Long Short Time Memory (LSTM) and Gated

Recurrent Unit (GRU). Bi-LSTM is Bi-directional Long Short Term Memory of RNN. The RNN has been used for different specifications which deal with the recurrent sequences layer with activation function, loss function, optimization function, and some other requirements. The LSTM is mostly used for the detection of abusive language and textual data recently [Chen et al., 2017, Founta et al., 2018a, Zhang and Luo, 2018, Madisetty and Sankar Desarkar, 2018, Pitsilis et al., 2018, Wiegand et al., 2018].

The specification and requirement used in RNN are activation function, dropout learning rate, avoid over-fitting, loss functions, optimization function, output. Summarized the RNN methods and their specifications in Table 2.8.

Algorithm	**Parameters**	**Ref**
Convolutional Neural Network (CNN)	Convolutional layers, rectified linear unit (ReLU), stochastic gradient descent, sigmoid, learning rate, L1-regularisation, cross-entropy, Adam optimizer, mean pooling, max pooling and multi-layer perception full-connected layer	[Lee et al., 2018b, Zimmerman et al., 2018]
Recurrent Neural Network (RNN)	Rectified linear unit (ReLU), stochastic gradient descent, sigmoid, tanh, SoftMax, dropout learning rate, L1-regularization's, categorical cross-entropy, binary cross-entropy, Adam optimizer, mini-batch gradient descent and fully-connected layer	[Lee et al., 2018b, Pitsilis et al., 2018]
Long Short Term Memory (LSTM)		[Chen et al., 2017, Pitsilis et al., 2018]
Gated Recurrent Unit GRU		[Chen et al., 2017]

Table 2.8: Summary of deep learning

2.4.3 Other Methods

Representation Learning

Graph-Sage ([Hamilton et al., 2017]) has been used as semi-supervised learning which combines graph-based features and machine learning methods. It transforms the graph or network into vectors that can be used by deep learning or machine learning algorithms through representation learning. The graph is converted into a low-dimensional vector like, for example, using Node2Vec [Ribeiro et al., 2018b].

Transfer Learning

Transfer learning has been also used in existing methods which is used when domain or knowledge transforms into also another domain. It outperformed when used in natural language processing where it helps to translate the text into another domain [Mathur et al., 2018, Kshirsagar et al., 2018, Founta et al., 2018a].

Extraction, Transformation and Loading (ETL)

It is a Data warehouse technique which denotes as the extracts, transforms and loads the data into a data warehouse system. It is used for the data to collect all features as input, because it is used to identify the features [García-Recuero et al., 2018].

Multi-Dimensional Analysis (MDA)

MDA has also been used for the detection of abusive language by using multi-dimension. The feature set is related to the part of speeches and grammatical structures. During this method, the multiple correspondence analysis (MCA) reduces the dimensions. The main function of MCA is the conversion of the high-dimensions into low-dimension [Clarke and Grieve, 2017].

2.5 Dataset Collection

Social media platforms sometimes provide data which can be used for research. Therefore, researchers created the datasets according to their domains and label the data by using labeling mechanisms which are discussed as in Section 2.5.2. Some studies use the existing datasets which already exist (as see Section 2.5.1). The datasets are taken from Social Media platforms such as Twitter, Facebook, Flickr, Yahoo and Blogs websites [Chen et al., 2017, Nobata et al., 2016, Gitari et al., 2015]. However, Twitter provides its data collection API with to access publicly available tweets (as see in Section 3.2). The overview of datasets of some existing is studies shown in Table 2.9

2.5.1 Existing Datasets

In this section, we discuss the existing datasets for the detection of abusive language, hateful speech, cyberbullying and aggression. The datasets which have been used in existing approaches include in Table 2.9.

Some datasets have been provided publicly with label data, but some are not available publicly. Dataset in [Waseem and Hovy, 2016] is commonly used in existing papers for detection of abusive language and hateful speech. The dataset is classified into sexism, racism and normal (see in Table 2.9). The dataset is a Twitter data which contain class sexism with 3,383 tweets with 613 unique users, a class racism contains 1972 tweets with only 9 unique users and the class normal contain 11559 tweets which are neither sexism or racism with 614 unique users [Pitsilis et al., 2018].

Another main important dataset in [Davidson et al., 2017] has been used for detecting hateful speech and abusive language. The dataset is labeled into hate, offensive and neither. The dataset comprises 24802 tweets [Chen et al., 2017, Founta et al., 2018a, Ribeiro et al., 2018a] (see in Table 2.9).

Reference	**Details**	**Size**	**Labels**	**Used By**
[Waseem and Hovy, 2016]	Twitter Data: Abusive language and hate speech Dataset	Tweets: 16914 Users: 1236	Sexism, Racism and Normal	[Founta et al., 2018a, Gaydhani et al., 2018, Park and Fung, 2017,Pitsilis et al., 2018]
[Davidson et al., 2017]	Twitter Data: Abusive language and hate speech Dataset	Tweets: 24802	Hateful, Offensive and Normal	[Gaydhani et al., 2018, Founta et al., 2018a, Park and Fung, 2017]
[Chatzakou et al., 2017a]	Cyberbullying: user behavior	Tweets: 9484 users: 1303	Aggressive, bullying, spam and normal	[Founta et al., 2018a]
[Rajadesingan et al., 2015]	Twitterdata: Sarcasm	Tweets: 61075	Sarcasm and non-sarcasm	[Founta et al., 2018a]
[Jindal and Liu, 2008]	Amazan Review Data (AMZ): product and review and purpose: compute polarity intensity/word embedding	Corpus: 1.2 billion		[Wiegand et al., 2018]
[Baroni et al., 2009]	Web as Corpus (WAC): large web corpus and purpose: compute word embedding	Corpus: 2.3 billion		[Wiegand et al., 2018]
	Rateitall (RIA): review purpose: compute polarity intensity	Corpus: 4.7 million		[Wiegand et al., 2018]
[Nobata et al., 2016]	Primary dataset: Temporal dataset: yahoo dataset about news and finance moderated by yahoo employee	Finance: 438436 news: 726073	Abusive and clean	
[Nobata et al., 2016]	WWW2015: yahoo dataset about news and financ	Comments: 951736	Abusive and clean	
[Nobata et al., 2016]	Primary dataset: Temporal dataset: yahoo dataset about news and finance moderated by yahoo employee	Finance data: 759402 News data: 1390774	Abusive and clean	
[Golbeck et al., 2017]	Harassment Twitter data manual labeled	Tweets:35000	Harassing and non-harassing	[Kshirsagar et al., 2018, Founta et al., 2018a]

Table 2.9: Summary of the existing datasets.

2.5.2 Data Collection and Data Labeling

The datasets have been crawled through API[1] and Scrappy [2]. In Twitter API, there are different data collection methods through different APIs which are Streaming API and Search API. These APIs are publicly accessible. By using search and stream, data can be collected using keywords, ids and user timeline [Founta et al., 2018b, Chatzakou et al., 2017b, García-recuero, 2016]

The datasets have been labeled by incorporating Crowdsourcing: group peoples of organization which label the data, CrowdFlower: the combination of machine learning, artificial intelligence and human expert text annotator, TextBlob python library: sentiment base labeling, sampling likes: snowball sampling, boost sampling, random sampling and boost random sampling, and manual labeling and judgment" [Watanabe et al., 2018, Founta et al., 2018b].

2.6 Evaluation Methodologies

In the section, we discuss the experiments and evaluation process of previous approaches. The first process is data splitting in which the datasets have been split into train and test. Cross-validation has been used for the split technique in which mostly used K-fold cross-validation. K-fold cross-validation is based on Kth-fold values. It has Kth iterations for splitting data into training and testing. Mostly, K-fold cross-validation has been used in existing studies with 5 to 10 K values [Zhang and Luo, 2018, Wiegand et al., 2018, Nobata et al., 2016]. Now, we discuss the evaluation metrics which have been used to evaluate the methods and algorithms. Evaluation metric based on confusion matrix which can be evaluated by using confusion matrix attributes, the attributes are based on predicted and actual class. The evaluation metric include precision, recall, f-measure and accuracy and other methods like

[1] `https://developer.Twitter.com/en/docs/tweets/search/api-reference`

[2] `https://github.com/bisguzar/Twitter-scraper`

Area Under Curve (AUC) and Receiver Operating Characteristics (ROC) [Ribeiro et al., 2018b, Founta et al., 2018a, Nobata et al., 2016, Chatzakou et al., 2017b, García-recuero, 2016, Zhang and Luo, 2018, Watanabe et al., 2018].

2.7 Research Gaps

In literature, we have discussed a variety of abuse detection methods which are based on lexicon, content and context. But, these approaches have a number of limitations. First, the lexicon-based approaches do not perform well for identify the abusive contents when contents include new words which are not in the lexicon [Gitari et al., 2015, Lee et al., 2018a, Wiegand et al., 2018]. This issue can be resolved by using content-based detection instead of using lexicon [Sharma et al., 2018, Ribeiro et al., 2018b]. But, in content-based approaches if ambiguous or new words occur, the detection approach does not perform well [Chen et al., 2017]. To understand the semantics of ambiguous and new words, contextual features have been used. Contextual features also perform better when the context contains patterns to distinct the abusive and non abusive language. Contextual features have to be designed carefully to avoid overfitting, for example, if the hashtags, mentions, or URLs are included in the context, such a context may result in overfitting and the machine learning algorithm may associate some of these features with abuse (or non-abuse) [Chatzakou et al., 2017b], [Watanabe et al., 2018].

2.8 Summary

This chapter covered the literature review in which we have discussed various methods of abusive language detection. The detection methods include keyword-based, content-based and context-based. The abusive language detection has been performed by using natural language processing methods which are linguistic, semantic, syntactic and n-gram features. Additional features have been used for the detection of abusive language. These additional features are extracted from user information and tweets information. Context features and metadata features are also categorized into tweets features, user features and network features. Temporal features include the window-based feature, these features have been mostly used for event detection. Further, we described the algorithms and methods which includes machine learning, deep learning and social network analysis methods. Other methods have been used for detection abusive language which include regression algorithm, ETL (extracting, transforming and loading) and MDA (Multi-dimensional analysis).

we also discussed the challenges presented in previous techniques about keywords. Then discussed the challenges about content and context-based detection. Further, we discussed the previously used datasets and data collection strategies. At last, we discussed about the evaluation measure to evaluate the abusive language.

Chapter 3

Dataset

Twitter provides access to real-time information from the Twitter data repository through the Twitter API. The existing approaches have been used and generated using various kinds of datasets. Some of these datasets are publicly available. The researchers use existing datasets to investigate different domains. We have focused on abusive language detection, which uses algorithms trained and evaluated on existing datasets. We have performed various data collection strategies to extract the data. We have also illustrated how we process the data. Data filtration has been performed to reduce the imbalance and missing values in the dataset. Data statistics have been presented to describe the facts and figures of datasets before and after retrieval.

3.1 Waseem And Hovy Dataset

In literature, existing datasets are used to investigate various domains including abusive language, harassment, hate speech and cyberbullying [Lee et al., 2018b,Tahmasbi and Rastegari, 2018]. These datasets have been retrieved from different social media platforms which are either social media blogs and websites. For instance, Twitter, Yahoo, Flickr, Facebook, Wikipedia, and some news websites. A few existing datasets are publicly available with labeled data, but a few datasets are not available with labeled data publicly.

In existing studies, different data labeling techniques have been performed to label the unlabeled datasets. Our focus of the study is based on Twitter data. Twitter datasets have been available which mostly used for the detection of abusive language and hateful speech. Therefore, we have used in [Waseem and Hovy, 2016] dataset to identify the abusive tweets.

The dataset in [Waseem and Hovy, 2016] has been retrieved from a large number of tweets corpus which contains 136,052 tweets. The annotated data contains 16,914 tweets and 1,239 unique users as detailed in Section 3.5. This dataset has been labeled into three classes which include sexism, racism and normal as shown in Table 3.1. It is mostly used for the identification of abusive language and publicly provided with the tweets ID's and labeled attributes according to the Twitter privacy policy. Therefore, we have decided to use the dataset [Waseem and Hovy, 2016] for the evaluation of our proposed appraoch.

3.2 Data Collection

We have collected the data for the detection of abusive language against existing Tweets IDs provided by [Waseem and Hovy, 2016] dataset. We have used several data collection strategies to collect the data using Twitter API and scraping. Twitter API presented with few limitations, therefore, we have also used Twitter Scraper to retrieve past tweets. The data collection strategies with different Twitter API using keywords, tweet ids, and user ids, etc.

However, we have used tweet ids and user ids of the existing dataset to collect data as required for the proposed approach. We have used three distinct data collection strategies including tweets information, user information, and past tweets. We have used scraping to fulfill the limitation of the search API for collecting past tweets.

3.2.1 Crawling API and Scraping

Twitter provides public and paid APIs to collect its data. Public APIs have limits for acquiring data. It has two different APIs, one is based on the past data and another is based on real-time data, these APIs are called Search API [1] and Streaming API [2], respectively. The paid API includes Twitter Firehose [3], it guarantees access to 100% tweets.

We have mainly used the Search API for collecting data. Search API allows to search data using words, usernames, geo-locations, etc. It provides limited access to collect tweets from past 7 to 15 days. Using the user-timeline API function, we can get up to last 3200 tweets per user. In our proposed approach, we need data from the past. To make the results comparable with existing approaches, we have to retrieve data using existing tweet ids provided in a dataset [Waseem and Hovy, 2016]. However, the tweets and user-timeline can be retrieved using the Search API, but past tweets are difficult to collect.

Conversely, we have used the Twitter scraper to collect past tweets because to recover the limited access of API. The Twitter scraper is not limited to collect the data, we can collect data from Twitter scraper [4] by using keyword or phrases, username and ids. However, we have used the Twitter scraper to collect past tweets from the existing tweet ids.

[1] `https://web.archive.org/web/20190809210308/https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets` Last accessed data October 12, 2020

[2] `https://web.archive.org/web/20200111022320/https://developer.twitter.com/en/docs/tweets/filter-realtime/overview` Last accessed data October 12, 2020

[3] `https://web.archive.org/web/20190809210314/https://developer.twitter.com/en/docs/tweets/compliance/api-reference/compliance-firehose`Last accessed data October 12, 2020

[4] `https://web.archive.org/web/20191207040547/https://github.com/taspinar/Twitterscraper`last accessed date October 12, 2020

3.2.2 Tweet- and User-based Collection

Tweet based collection retrieves the characteristic of the tweet which have been configured in different attributes. These attributes of tweets information have been provided on Twitter which includes user ids, text, user and entities as shown in Figure 3.1. Tweet-based is used to collect the tweet information by using different attributes of tweets such as keyword or phrases, username, specific hashtags, tweet id and user id.

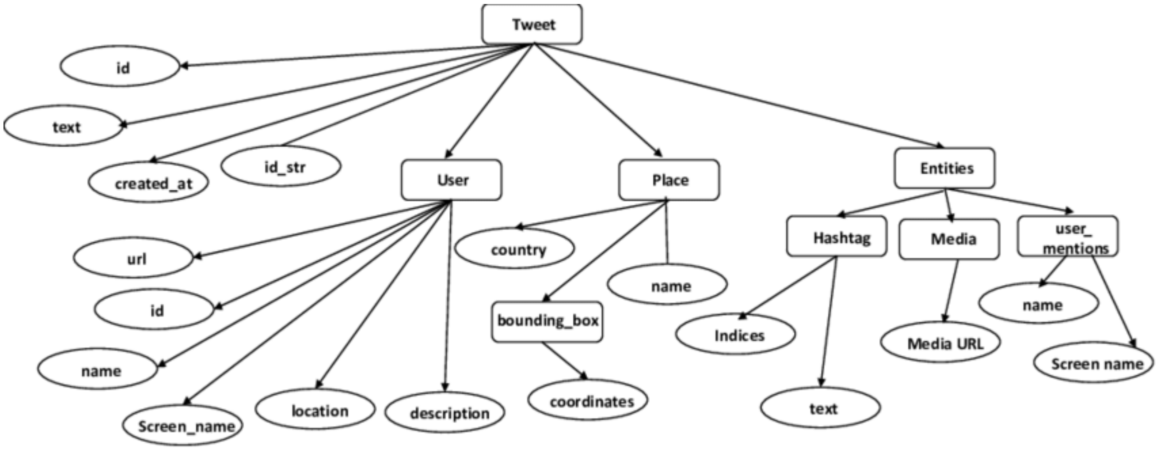


Figure 3.1: Tweets attributes or information

User-based collection retrieves the characteristic of user profile attributes. These attributes have been provided by Twitter API. User-timeline is used to retrieve user related attributes. The user attributes include user id, screen name, location, user description, followees and followers in Figure 3.1.

After further collections, a few numbers of tweets and user data are not acquired (see in Table 3.2). It may be deleted or publicly inaccessible.

3.2.3 Past-Tweet Collection

Past-tweets collection has been collected through tweet ids in the dataset. The past tweets against every tweet ids are acquired through Twitter API with limited access and limited data. But, Twitter API has been restricted for collecting old tweets and time request responses because of the search API limitation to access tweets from past 7 to 15 days. We have used users ids to retrieve the past tweets but the tweets are tweeted within a few years ago then do not contain the required tweets. Therefore, the user-timeline has limits in getting previous tweets. Therefore, we have used the Twitter-Scraper to collect past tweets against every current tweet. Twitter Scraper has been used for the collection of data, where we can easily retrieve all previous tweets If the user and tweets are publicly accessible and reachable. Unfortunately, every tweet of past tweets has not been retrieved because many of the users and their tweets do not exist on Twitter anymore.

3.3 Data Processing

In this section, we discuss the Twitter data format and the data processing in the dataset.

3.3.1 Data Format

Twitter provides data collection from Twitter. By default JavaScript Object Notation (JSON)[5] is used to serve the Twitter data. JSON format contains data in form of key/value pairs, where the value can be another nested JSON object or a list. An example of JSON record obtained using Twitter API is shown in Figure 3.2. In this example, "text" is the key and "Creating a Groc..." is its value. The values of the keys named "indices" are lists containing IDs of respective indices.

3.3.2 Data Collection Process

We have used the dataset in [Waseem and Hovy, 2016] which also contained the tweets ids and its labels attributes discussed in Section 3.1, Consequently, we have been re-

[5] `https://web.archive.org/web/20200119021743/https://www.json.org/json-en.html` Last accessed data October 12, 2020

```
[{
  "created_at": "Thu Jun 22 21:00:00 +0000 2017",
  "id": 877994604561387500,
  "id_str": "877994604561387520",
  "text": "Creating a Grocery List Manager Using Angular, Part 1: Add
&amp; Display Items https://t.co/xFox78juL1 #Angular",
  "truncated": false,
  "entities": {
    "hashtags": [{
      "text": "Angular",
      "indices": [103, 111]
    }],
    "symbols": [],
    "user_mentions": [],
    "urls": [{
      "url": "https://t.co/xFox78juL1",
      "expanded_url": "http://buff.ly/2sr60pf",
      "display_url": "buff.ly/2sr60pf",
      "indices": [79, 102]
    }]
  },
  "source": "<a href=\"http://bufferapp.com\"
rel=\"nofollow\">Buffer</a>",
  "user": {
    "id": 772682964,
    "id_str": "772682964",
    "name": "SitePoint JavaScript",
    "screen_name": "SitePointJS",
    "location": "Melbourne, Australia",
    "description": "Keep up with JavaScript tutorials, tips, tricks",
    "url": "http://t.co/cCH13gqeUK",
    "entities": {
      "url": {
        "urls": [{
          "url": "http://t.co/cCH13gqeUK",
          "expanded_url": "http://sitepoint.com/javascript",
          "display_url": "sitepoint.com/javascript",
          "indices": [0, 22]
        }]
      },
      "description": {
        "urls": []
      }
    },
    "protected": false,
    "followers_count": 2145,
    "friends_count": 18,
    "listed_count": 328,
    "created_at": "Wed Aug 22 02:06:33 +0000 2012",
    "favourites_count": 57,
    "utc_offset": 43200,
    "time_zone": "Wellington",
  },
}]
```

Figure 3.2: Structure of a JSON document

trieved the data as requiring the proposed Approach. We have collected the tweets from Twitter API [6] through *\Get\ statuses\ lookup* and *\Get\ user\ lookup* service which allows acquiring the tweets attributes through the request of each tweet ids or user ids. For example, we sent the request to get Twitter data through call API.

`https://api.Twitter.com/1.1/statuses/lookup.json`

`https://api.Twitter.com/1.1/user/lookup.json`

In the tweets collection, we have encoded the data into comma-separated values (CSV) format with proposed attributes. The attributes are included tweet ids, time and date, tweet text and entities like mentions, hashtags, URLs and retweet. The user-related attributes are user id, username, user descriptions, followers, followees, listed and favorites as shown in Figure 3.1

3.4 Data Filtration

The dataset provided in [Waseem and Hovy, 2016] is obtained. The data acquired is imbalanced as discussed in Section 3.5. Therefore we filter the data to overcome the missing values and imbalance data. First, we merge the classes sexism and racism into a single class abusive, because both the classes represent abuse. These classes have also only a few tweets as shown in Table 3.2. The class normal is converted into non-abusive as shown in Table 3.3.

3.5 Data Statistics

The dataset in [Waseem and Hovy, 2016] has been frequently used for the detection of abusive language 3.1. Table 3.1 shows the original data which contain 16,914

[6] `https://web.archive.org/web/20191225062951/https://developer.Twitter.com/en/docs/tweets/search/overview`Last Accessed date October 12, 2020

tweets. This dataset is annotated. The dataset [Waseem and Hovy, 2016] contains 35.3% abusive tweets [Wiegand et al., 2018], 7.3% unique users, 20.0% sexism, 10.1% racism and 68.3% normal tweets.

Class	Tweets	Users
Sexism	3,383	613
Racism	1,972	9
None	11,559	614
Totals	16,914	1236

Table 3.1: Existing tweets and users

After data collection of tweet and user timeline, a few numbers of data could not be retrieved. After crawling the available tweets and their user information, 65.2% data have been remained as shown in Table 3.2.

Class	Tweets	Users
Sexism	2699	203
Racism	261	1
None	8069	463
Totals	11029	667

Table 3.2: Tweet and user API

The original dataset contains 11,029 tweets. We filtered this dataset as discussed in Section 3.4. After filtration, the data contains 7,412 tweets and 662 unique users. We have merged the classes racism and sexism as abusive and normal as non-abusive, hence making the abuse identification problem as a binary classification problem. The statistics of the final dataset are shown in Figure 3.3.

Class	Tweets	Users
Abusive	2960	204
Normal	4452	458
Totals	7412	662

Table 3.3: Merged datasets

3.6 Summary

Twitter provides its real-time data to process and investigates innovations and detection. However, it provides the API to acquire the data from the Twitter data repository. We have used the search API to collect the data. Search API has limited access to retrieve data. In this API, it has some limitations to crawl the data. We have also used Twitter scraper to resolve the limitations of the search API. We used the exiting dataset which is commonly used for the detection of abusive language and hate speech. We performed filtration to reduce imbalanced and missing data. Lastly, we provided the statistics about the dataset.

Chapter 4

Proposed Method and Experimental Setup

We describe the proposed method and experimental setup in this chapter. First, we have discussed the proposed methods in Section 4.1. Proposed approach, We describes the pre-processing, content-based approach, contextual feature and classifications and work-flow of our proposed approach. First, the pre-processing is performed to construct the dataset for further processing. Secondly, elaborates the content-based approach is used in different features. Third, different the contextual features have been used which include the tweets features, user features and window-based features. Finally, supervised learning algorithms are used to evaluate the detection of abusive tweets. We discuss the experimental setup in Section 4.2. First, we implemented the state-of-the-art. Second, we have used cross-validation to split the dataset into testing and training. Finally, we have elaborated the evaluation measure to evaluate the proposed approach.

4.1 Proposed Method

First, the discussion is about the preliminary steps of our proposed Approach. These steps are based on content and context features. We have used three different combinations of contextual features with the content of tweets to detect the abusive language. The contextual features include tweet-, user- and window-based features. We evaluated the detection of abusive tweets using supervised machine learning according to the feature sets.

4.1.1 Pre-processing

After finalizing the dataset according to the proposed approach, pre-processing has been performed to prepare the data for further processing. First, we have converted the uppercase into lowercase. We performed some replacement and removal of words and tags, etc. Example of attributes performed pre-processing are in Figure 4.1

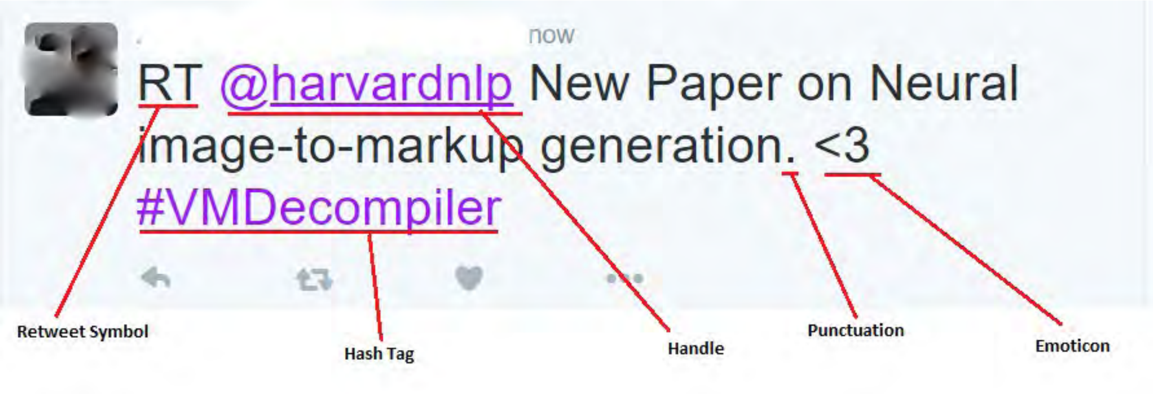
We replaced the abbreviations, misspelling words into their original words and replaced the other form of the words into expanded form using stemming and lemmatization techniques (see in Figure 4.1). We used the porter stemmer[1] and wordnet lemmatizer[2].

Afterwards, we removed the unwanted stopwords, words, special characters, symbols, punctuation and emojis. For example, the overview of pre-processing to remove the unwanted attributes in Figure 4.1. Removed the duplicate words and the special character for example (@, #, -,$). In textual features, the hashtags, mentions, retweets and URLs have been removed.

After pre-processing, the data is prepared for further processing to extract the textual content and context features. For example, overview before and after performing

[1] `https://web.archive.org/web/20190627112454/http://snowball.tartarus.org/algorithms/porter/stemmer.html` Lasted access date October 12, 2020

[2] `https://web.archive.org/web/20190711193825/https://www.machinelearningplus.com/nlp/lemmatization-examples-python/` Lasted access date October 12, 2020

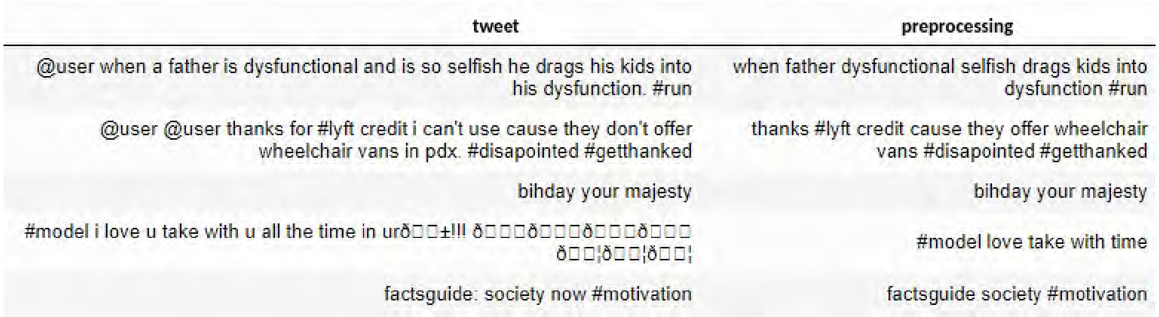


now

RT @harvardnlp New Paper on Neural image-to-markup generation. <3 #VMDecompiler

Figure 4.1: Example of pre-processing attributes

pre-processing in Figure 4.2.



tweet	preprocessing
@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunctional selfish drags kids into dysfunction #run
@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #lyft credit cause they offer wheelchair vans #disapointed #getthanked
bihday your majesty	bihday your majesty
#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ ð□□¦ð□□¦ð□□¦	#model love take with time
factsguide: society now #motivation	factsguide society #motivation

Figure 4.2: Overview of before and after performing pre-processing

4.1.2 Content-based approach

In this section, the content-based approach is based on the tweets T of different users. Pre-processing has been performed to prepare tweets T for further process as discussed above Section 4.1.1. Afterwards, we have used statistical features on content. Tweets T are represented as:

$$T = \{t_1, t_2, \ldots, t_n\}$$

Where T is a set Tweets, t is a tweet, n is the maximum the number of tweets However, every tweet t contains the number of words W. Words W are represented as:

$$W = \{w_1, w_2, \ldots, w_m\}$$

Where W is the number of words in tweet t and m is the maximum number of words in every tweet t.

Accordingly, the words W have been transformed into statistical values or vectors using statistical features $stat_F$. The statistical features create the feature into machine-understandable which converted the tokens or terms into vectors including n-gram, Frequency-Inverse Document Frequency (Tf-Idf) and word embedding. We have used the uni-gram with Tf-Idf. Every word W in tweets are specified the Tf-Idf weights which are represented as:

$$Tf - idf(w,t) = tf_{w,t} * idf_w$$

Term frequency $tf_{w,t}$ is represented as:

$$tf_{w,t} = \frac{n_{w,t}}{\sum_k n_{k,t}}$$

where w denotes terms (words) appear in document (tweets)t, and n denotes that there are total k the number of words w_n in tweets t.

Inverse document frequency idf is represented as:

$$idf_w = \log \frac{|T|}{|t : t_w \in t|}$$

where T denotes tweets contain in the dataset which has already been defined above.

4.1.3 Contextual Features

Context features F are the factors of additional characteristics of tweets and their users. In our proposed Approach, the contextual features F are based on combinations of features. Performance of each attribute in the feature sets is evaluated individually. On the behalf of their performance, we have added the attributes one by one into each feature sets. Accordingly, we categorise the contextual features F into three sets: tweet features F_t, user features F_u and window-based features F_w as shown in Table 4.1.

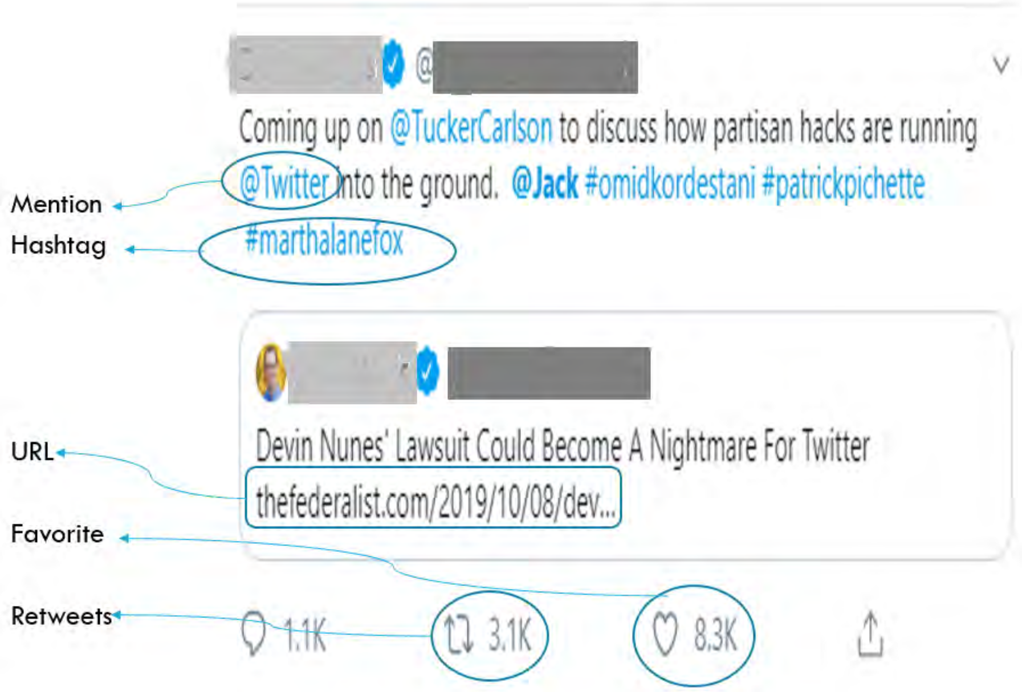
Contextual Features	**Features**	**Description**
Content Features	Retweets	Number of retweets to other tweets
	Favorite	Number of tweets are favorite.
	Mentions	Number of mentions in tweets.
	Hashtags	Number of hashtags in tweets
	URLs	Number of URLs in tweets
User Features	Follower	Number of users followed to the user
	Followees	Number of users followed by the user
	Favorite	Numbers of favorites to user profile
	List	Number of listed to user profile
	Description	The user describes themselves
Window-Based Features	Previous Tweets	Takes past tweets which is based on sliding window
	Temporal Tweets	Takes past tweets which is based on temporal window

Table 4.1: Combination of contextual feature sets

Tweet Features

Tweet features F_t are the combination of characteristics of tweet attributes. These features help the detection of abusive tweets using some popularity and tweet features. The tweet features F_t =$\{f_{t1}, \ldots, f_{tn}\}$ which includes the Mentions Men, Hashtags

$Hash$, URLs, Retweets RT and favorite Fav (Table: 4.3). Where f_t denotes individual feature attributes and n denotes the number maximum features. We have used



[redacted] @[redacted]

Coming up on @TuckerCarlson to discuss how partisan hacks are running @Twitter into the ground. @Jack #omidkordestani #patrickpichette #marthalanefox

[redacted] [redacted]

Devin Nunes' Lawsuit Could Become A Nightmare For Twitter

thefederalist.com/2019/10/08/dev...

Figure 4.3: Example of tweet features

the counts of these features which is represented as:

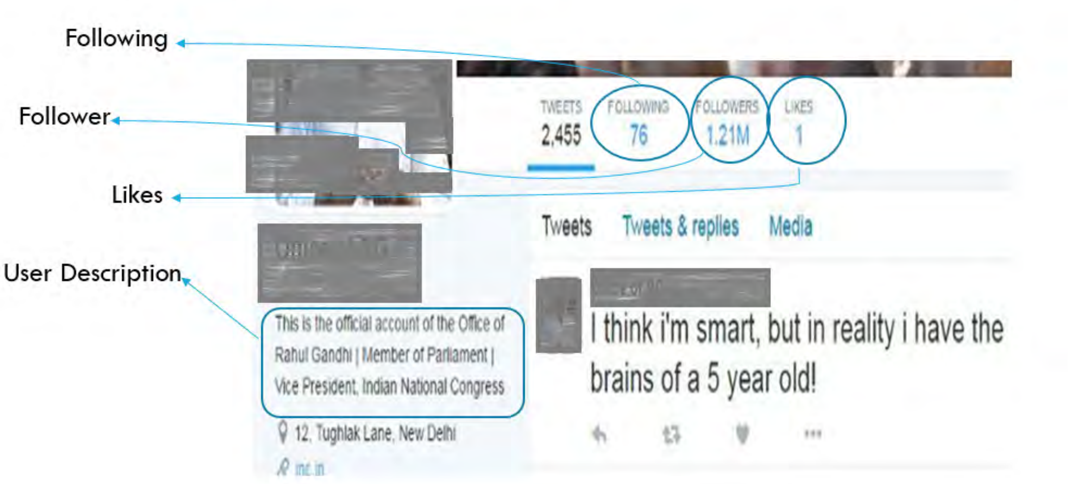
$$F_t(count) = \{count(RT), count(Fav), count(Men), count(Hash), count(URLs)\}$$

Where $F_t(count)$ denoting the counts of Tweet features F_t, RT denotes the Retweets, Fav denotes the Favorite, @ denotes the mentions, # denotes the hashtags and URL denotes the URLs which contain in tweets.

In these features, we have used two popularity features Pop_Fea which are the Retweets RT and Favorites Fav attributes. These attributes describe tweet popularity based on their counts.

User Features

User features F_u are the combination of characteristics of user profile information. These combinations of features F_u describe the popularity and connections of users which include followers and followees, favorites, like, listed and user descriptions (See in Figure 4.4.



This is the official account of the Office of Rahul Gandhi | Member of Parliament | Vice President, Indian National Congress

12, Tughlak Lane, New Delhi

inc.in

[redacted]

I think i'm smart, but in reality i have the brains of a 5 year old!

Figure 4.4: Examples of user features

In these combinations of features F_u, we used their counts which have been represented as:

$$F_u(count) = \{count(Fr), count(Fe), count(fav), count(like), count(listed)\}$$

Where $F_u(count)$ denotes the count values of Tweet features, Fr denotes followers, Fe denotes the followees and fav denotes the user favorites.

In these features, we also used user descriptions, it contains text in which describe user him/herself in a few words. First, we have also performed pre-processing as discussed in Section 4.1.1 in this feature. Moreover, statistical features are used the same as above Section 4.1.2. we have converted the token in datasets into vectors using uni-gram with Term frequency-inverse document frequency (Tf-Idf) weight. It has been discussed briefly in Section 4.1.2.

$$U_{dis} = Tf - idf(w, ud) = tf_{w,ud} * idf_w$$

Where U_{dis} denotes the users description features, w is the terms in document ud, ud denotes the content of user descriptions.

Lastly, user descriptions U_{dis} features combine with the counts features $F_u(count)$. These features have been defined as:

$$F_u = \{F_u(count), U_{dis}\}$$

Window Features

Window features F_w have been used which are based on past tweets pt of every tweet ct in a dataset. In these features, content or tweets text of past tweets within the window has been extracted. First, we have also performed pre-processing on past tweets as discussed in Section 4.1.1. Consequently, statistical features have performed including uni-gram with Tf-Idf have as described in Section 4.1.2.

$$F_W = Tf - idf(w, pt) = tf_{w,pt} * idf_w$$

Window-based features have two different features including previous window and temporal features (Figure 4.5).

The purpose of these features is to confirm that the current tweets are either abusive tweets or normal tweets. These features detects and clear the ambiguious tweets. Table 4.2 shows the abuses comprising in several tweets when the conversions on Twitter performed within specific content and time. Sometimes, the malicious person misuses other user accounts and abusive tweets are being tweeted. These

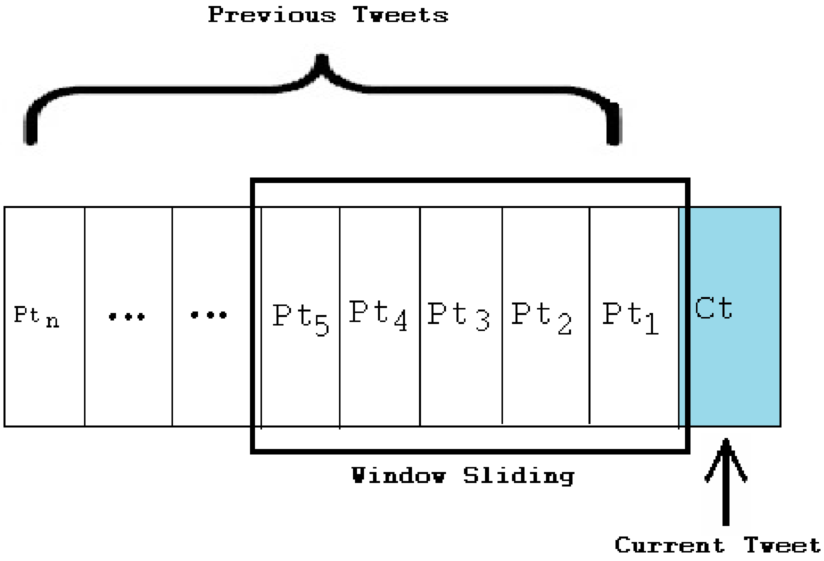


Figure 4.5: Window-based features of previous tweets

features help the detection of abusive language more accurately and efficiently.

Previous Tweets Window

We have used past tweets *pt* of every current tweet *t* as features within the sliding window. The past tweets have been taken through the different variance of sliding window s_{win}. The sliding window s_{win} is specific ranges of past tweets. We have arranged different range of sliding window s_{win} including 5, 10, 15 and 20.

However, check past tweets *pt* of every tweet *t* if past tweets are not available then moves next tweet *t* and check again and again. If the past tweets *pt* are available, sliding window s_{win} is applied, the past tweets *pt* text is taken within the window and concatenated the past tweets (see Algorithm 1) which are represented as:

$$pt(t) = \{pt(t_1), pt(t_2), \ldots, pt(t_n)\}$$

Where *pt* is past tweets and t is each current tweet.

Temporal Tweets window

We have used the previous tweets *pt* based on temporal window F_{tw}. In this fea-

Time	Previous tweets
7:38:05	This shit is so real b**** http://t.co/jLRYTM5gVi
7:34:44	@User1 You can D*** OFF now ðŸ˜,ðŸ˜
7:31:28	@User1 I'm ok with that ðŸ˜,ðŸ˜,
7:28:50	Slipping in to weekend mode like http://t.co/tTjX41FVyV
7:26:43	Your baby can't even clear the bong fmd
7:24:30	*drinks friends beer*
7:20:58	Why won't Twitter just shoot to the top of my TL like a normal cunt?! F*** you
7:18:52	My b** just swallowed my m**
7:18:16	I just ate a plate of loi hoosi and now nothing I'm wearing fits
5:11:08	@User2 I catch a train to work, that's nothin
4:44:08	Domestic violence victims wish they saw white and gold
4:40:30	User2 what if I just g**** h*** it and st*** my d*** in its cash slot? Is it a GAYTM?
4:36:24	Just used the ANZ GAYTM *y** b*****
4:18:35	Your daily horoscope: F*** OFF
4:03:48	@User3 u ain't got shit on me Corbyn
3:47:30	You all wear ugly dresses on the reg tho.. Let's be honest

Table 4.2: Example of previous tweets

Algorithm 1: Previous Tweets Window

```
Input: TWEETS
Input: $s_{win}$
Output: $F_{pw}$
foreach $t$ in TWEETS do
    $m \leftarrow min(s_{win}, n)$ ;   ▷ $n$ is the number of previous tweets by the same user of the tweet $t$
    $Pt \leftarrow \{pt_1, \ldots, pt_m\}$ ;   ▷ $pt_i$ is the previous $i^{th}$ tweet of $t$.
    $F_{pw} \leftarrow Concat(F_{pw}, Pt)$ ;   ▷ $f_{pw}$ have used past tweets within sliding window-based as feature.
end
```

ture, we have organized the time window $Time_{win}$ and have taken the past tweets pt content within the time window $Time_{win}$. This feature F_{Tw} is useful to check the confirmation of the current tweet t using its past tweets within the specific time window. We have arranged three different intervals of time windows $Time_{win}$ including 6 hours, 12 hours, 2 days respectively.

First, we have taken the date and time of current tweets $Time_t$ and every past tweet $Time_{pt}$ of each current tweet then the times have been converted into second. After the conversion of times, we have computed the time difference $Time_{Diff}$ between the time of current tweet $Time_t$and time of its previous tweets $Time_{pt}$. Therefore, we have checked the time difference $Time_{Diff}$ with time window $Time_{win}$, if the time differences $Time_{Diff}$ contained within time windows $Time_{win}$ then we take previous tweet contents pt (Algorithm: 2). This feature has been represented as:

$$pt(t) = \{pt(time_{win1}), pt(time_{win1}), \ldots, pt(time_{winn})\}$$

However, we concatenated the past tweets content pt within the time windows $Time_{win}$.

4.1.4 Supervised Machine Learning

We have used supervised learning algorithms to evaluate the detection of abusive tweets. However, we used two different supervised learning algorithms including classifications methods and ensemble methods with its specification or parameter in Table 4.3.

4.1.5 Work-Flow of Proposed Method

In this section, the discussion is about the flow of the proposed approach. It contains various steps that have discussed above Section 4.1. The proposed approach started

Algorithm 2: Temporal Tweets Feature

```
Input: TWEETS
Input: $Time_{win}$
Output: $F_{tw}$
foreach $t$ in TWEETS do
    $Pt \leftarrow \{pt_1, \ldots, pt_n\}$ ; ▷ $pt_i$ is the previous $i^{th}$ tweet of $t$ by the current tweet.
    foreach $p$ in PT do
        $Time_t \leftarrow Get(Time_of(t)_{seconds})$; ▷ A time $Time_t$ of current tweet $t$ converted into $second$.
        $Time_{pt} \leftarrow Get(Time_of(pt(i \rightarrow n)_{\mathbf{seconds}})$ ; ▷ The times $Time_pt$ of all previous tweets $Pt$ converted into $second$.
        $Time_{Diff} \leftarrow Difference(Time_t, Time_{pt})$ ; ▷ The difference $Time_{Diff}$ of current tweet time $Time_t$ and all its previous tweets time $Time_pt$
        if $\boldsymbol{Time_{Diff}} <= Threshold(\boldsymbol{Time_{win}})$ then
            $p(Time_{win}) \leftarrow Get(p(1) \rightarrow p(Time_{win}))$ ; ▷ $p(Time_{win})$ is the previous tweets $p$ within $Time_{win}$ temporal window.
        $F_{tw} \leftarrow Concat(F_{tw}, p(Time_{win}))$ ; ▷ $F_{tw}$ have used past tweets $pt$ within temporal window-based as feature.
```

from the dataset, pre-processing, content features, and contextual features, supervised learning and evaluation and results respectively. Figure 4.6 shows the workflow of the proposed approach.

Dataset

The dataset has been used for the detection of abusive tweets. We used the existing dataset in [Waseem and Hovy, 2016]. After the data collection, we filtered the data because it was highly imbalanced and significantly existed the missing values. (see in Section 3.2). After the preparation of the dataset moves forward next step.

Pre-processor

Pre-processing is performed to remove the noise in dataset for further processing. Performed some removing and replacement processes in pre-processing as discussed

Supervised Learning	**Classification**	**Parameter**
Linear Classifications	SVM	Kernel = linear and C = 1
	Logistic Regression	Regulation "penalty= L2" , random number generator "random_state = 42", optimization problem "solver = liblinear" and multi_class =auto
	Naive Bayes	MultimonialNB with additive smoothing and alpha =1.0
Ensembles Methods	Random Forest	Number tree in forest "n_estimators =35" and Random number generator "random_state =42"
	AdaBoosting	Number tree in forest "n_estimators =100", sub_sample = 1 and learning_rate =1.0
	GradBoosting	number tree in forest "n_estimators =35", learning_rate =1.0 and Random number generator "random_state =42"
	XGBoost	
	LightBoost	Number tree in forest "n_estimators =100", num_leaves =31, learning_rate =1.0 and Random number generator "random_state =42"
	Bagging	Number tree in forest "n_estimators =100", and Random number generator "random_state =42"

Table 4.3: Classification methods and their parameters

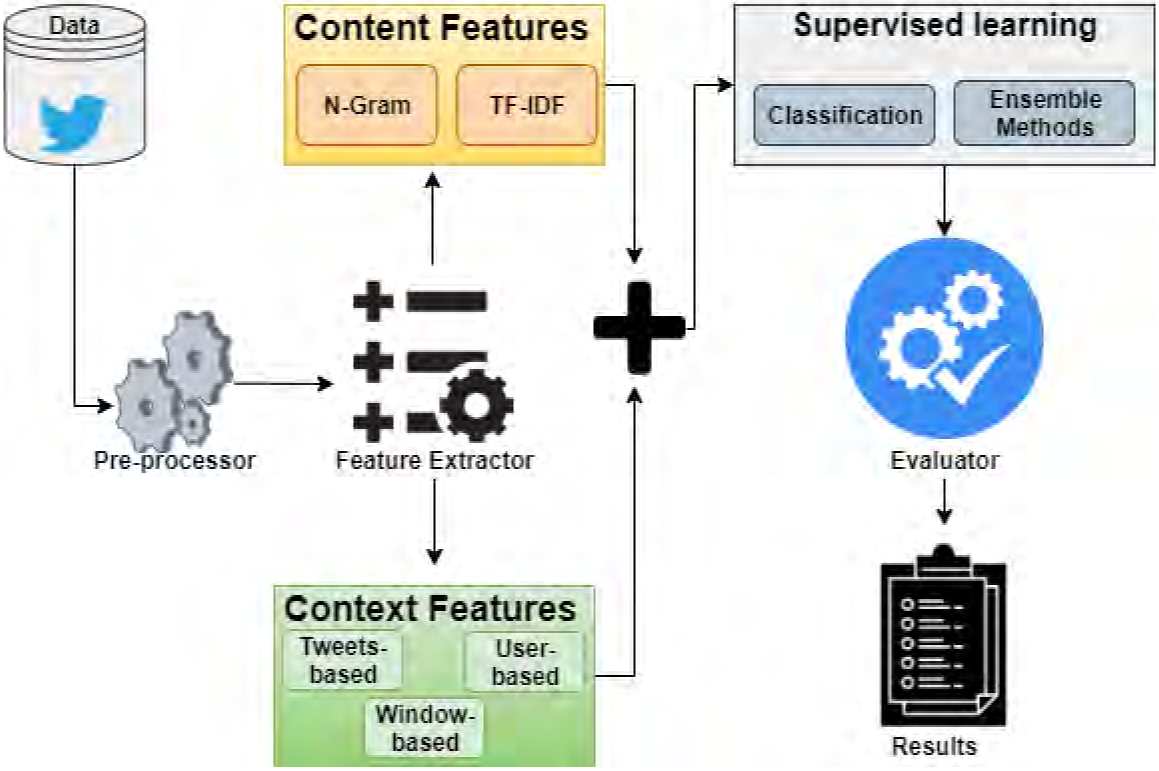


Figure 4.6: Workflow of proposed approach

in Section 4.1.1.

Feature Extractor

After pre-processing, the dataset can be used for next processes which are feature extractions. We have extracted two different main features including content and context features. Finally, concatenated the features and move toward the classifying process.

Content features

The statistical features are performed on contents features. uni-gram and Tf-If are used to convert the word into vector according to their weight for the evaluation process.

Context features

After the previous step, we have used different context features that belong from the tweets' information, users' information and past tweets the context features known as tweets features, user features and window based features (see in Section 4.1.3). These context features have been used individually, Afterwards, these individual features have concatenated which is proposed approach.

We have used three different context features to detect the abusive tweets which are such as:

- **Tweet features** have been taken with different sub-categories of tweet attributes which contains the retweets, favorite, mention, hashtag and URL.
- **User features** have been taken with different sub-categories of user attributes which contain the followees, followers, favorites, likes and user descriptions.

- **Window-based features** have two sub-features which are previous window feature and Temporal feature.

Supervised Learning

We performed supervised machine learning with their parameters. We have used linear classification as well as ensemble methods which discussed detail in Section 4.1.4. The classification including SVM, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, AdaBoost, GradientBoost, XGboost, LightBoost and Baging. Applied these methods move towards the evaluation steps.

Evaluation

The evaluation processes have been performed Stratified-Kfold cross-validation and evaluation metrics. Stratified-Kfold split data into train data and test data discussed in Section 4.2.3. Different evaluation measurements are used to evaluate the performance. According to these measurements, move forward the next step to analyze the performance.

Result

According to the evaluator, the results have been illustrated by the individual and combined features through concatenator which is our proposed approach. Compared the results with each other. Lastly, the results of the proposed approach are compared with state-of-the-art as discussed in Chapter 5.

4.2 Experiment

In this section, the experiment process including state-of-the-art discussion, cross-validation and evaluation metrics. First, we implemented the state-of-the-art containing two baseline papers. These baseline papers are compared with the proposed approach. Secondly, we used cross-validation to distribute the dataset into training data and testing data. Finally the discussion about the evaluation measures are used to evaluate the performance of techniques.

4.2.1 State-of-the-art Approach

We have discussed the features set and algorithms of state-of-the-art. We have implemented the base papers and evaluated the result by incorporating the proposed dataset as discussed Section 3.1. We have chosen two base papers from our existing approach which are "Automated Hate Speech Detection and the Problem of Offensive Language [Lee et al., 2018b]" and "Comparative Studies of Detecting Abusive Language on Twitter [Davidson et al., 2017]".

Paper1: "Automated Hate Speech Detection and the Problem of Offensive Language [Lee et al., 2018b]"

In this paper, the authors used tweet content with some additional features. The tweet contents were used the statistical features which are uni-gram with Tf-idf. The statistical features also used in proposed textual content-based features. The additional features are the characteristics of tweets content which contained the sentiment score, word count, character count and syllable count. They have used the machine learning classification to evaluate for the detection process which includes Linear-SVM, Naive Bayes, Logistic Regression, Random Forest and Gradient Boosting Classifier.

Paper2: "Comparative Studies of Detecting Abusive Language on Twitter [Davidson et al., 2017]"

In this paper, authors have used tweets content which performs into word-level and character-level with Tf-idf weight. In word-level, they have used uni-gram, bi-gram, tri-gram and tetra-gram. In character-level, researchers have used (3 to 8)-gram. They have used supervised machine learning and Neural Network. The classifications include Linear-SVM, Naive Bayes, Logistic Regression. Thus, The neural network includes the Convolutional Neural network (CNN) and Recurrent Neural Network (RNN). Neural Network methods also performed word and character level.

In the Convolutional Neural network (CNN), for word-levels, three convolutional layers have been used with sizes 1, 2 and 3 respectively. the parameter of CNN me activation function "Relu" and "Max-pooling" and Character-level CNN have uses seven layers with size 3 to 8 with Max-pooling.

In Recurrent Neural Network (RNN), authers used GRU with 50 hidden layers and one encoding layer, loss function "cross-entropy" and "sigmoid" and Adam optimizer. They have also used Bi-directional RNN as the baseline. GRU is the Algorithm of RNN.

4.2.2 Implementation

We have implemented our proposed approach in python programming language. We used Scikit-Learn as a machine learning library. We implemented a crawler using tweepy to extract data from twitter using Twitter API. To pre-process the data, we used Scikit-Learn with NLTK packages. Afterwards, custom code was written to extract features. Supervised learning algorithms mostly available in Scikit-Learn have been used for evaluation. The flow of implementation is described in section 4.1.5.

4.2.3 Data Cross-Validation

The data split techniques to distribute the dataset into testing and training data. In the machine learning aspect, it is known as cross-validation. There are different data techniques which include holdout method, K-fold cross-validation and Stratified K-fold cross-validation.

However, we have used K-fold cross-validation to split training and testing data but it does not performs well because of the large number of imbalanced labels in the dataset. Therefore, we have used Stratified K-fold cross-validation for data split because in this method it takes all labels in every iteration of the fold as shown in Figure 4.7 and it performs better than cross-validation. We have used 10-fold to distribute data into training and testing.

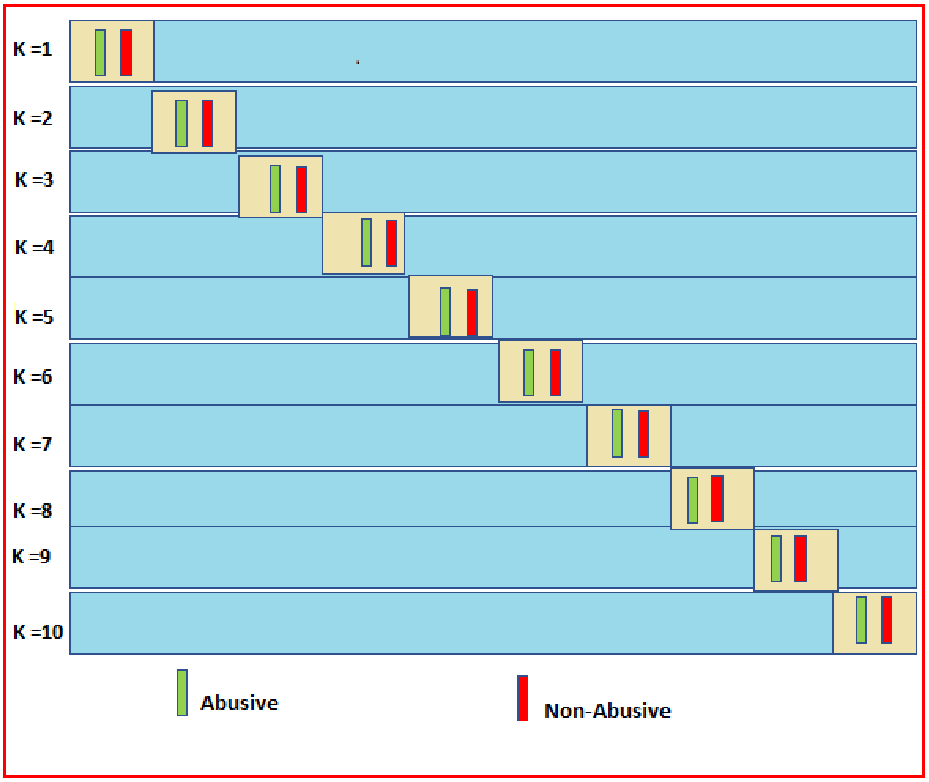


Figure 4.7: Stratified K-fold cross-validation

4.2.4 Evaluation Measure

We have used evaluation metrics to compute the performance of the proposed methods which include accuracy, f1-score, precision, and recall. The evaluation metrics have been computed through the attributes of the confusion matrix. The confusion matrix contains two dimensions of classes which are actual and predicted. The attributes of the confusion matrix contains True Positive (TP), False Positive (FP), True Negative

(TN) and False Negative (FN) as shown in Table 4.4.

We have defined as the attributes of confusion matrix according to the classes in the dataset which are categories into abusive and non-abusive (See Table: 4.4):

- **True Positive (TP):** Defines correct prediction of predicted and actual abusive tweets class "Correct prediction".

- **True Negative (TN):** Defines correct prediction of predicted and actual non-abusive tweets class "Correct prediction".

- **False Positive (FP):** Defines correct prediction of predicted abusive but actual non-abusive tweets class "Wrong prediction".

- **False Negative (FN):** Defines correct prediction of predicted non-abusive but actual abusive tweets class "Wrong prediction".

Predicted			
Actual		**Non-Abusive**	**Abusive**
	Non-Abusive	**TN**	**FP**
	Abusive	**FN**	TP

Table 4.4: Confusion Matrix

Accordingly, we have formalized the evaluation matrix which is defined as Precision, the ratio between correct prediction classes with predicted positive which have represented mathematically as:

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

Recall, the ratio between correct prediction classes with the actual positive which is represented mathematically as:

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

F1-score, the harmonic average of precision and recall which is represented mathematically as:

$$F1 - score = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

Accuracy, the ratio between the correct predicted with whole matrix which is represented mathematically as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.3 Summary

In this chapter, first, we have discussed the key motivation and then the preliminary steps of the proposed approach. Tweet contents have been used statistical features. The context features have also been contained the tweets features, user features and window-based features. Each feature is categorized into different which are extracted from attributes of tweets, users and its past tweets. The window features are the characteristics of past tweets of every current tweet. The window features are two different features of past tweets which are the previous window-based and temporal features. These features have been evaluated individually and combined with supervised machine learning. We have described the architecture of our proposed methods. Afterwards, we have discussed the experiment process which has state-of-the-art approaches, cross-validation to separate the data into training and testing. in the end, we have discussed the evaluation measure to check the performance of the proposed methods.

Chapter 5

Results and Discussion

In this chapter, we discuss the results of the proposed approach. We analyzed the results into individual features, and combined the individual features. Compared the context with each other individually. The proposed approach is compared with state-of-the-art and content-based approach as discuss in Section 5.1. Moreover, we discuss some discussions of results in Section 5.2.

5.1 Results

The categorized and investigated of abusive language detection have been proposed in recent studies. The studies classified the abuse detection in the form of keyword-based, content-based and context-based. Our proposed method is based on three different contextual features containing tweet content, user-based and window-based features with content. We considered state-of-art and explore the combination of context features.

We have illustrated the evaluation into five different steps such as: 1) Examine the individual contextual features. 2) Analyze different intervals of the sliding window and temporal window. 3) Combined the important factors of contextual features. 4) Analyzed the proposed approach. 5) Comparisons.

5.1.1 Individual Features

The results of content and context features are performed individually. First, we have elaborated on the results of content-based approach which is the basic part of detection the abusive tweets of our approach.Moreover, the discussion is about the performance of the contextual features which include tweet, user and window-based features.

First, we have analyzed the results of the contextual features with content, the performance of individual features is better as compared to content-based approach. Table 5.1 shows the consequence of tweet features and user features as well as a content-based approach. The highest performance of content-based has 77% accuracy and 76% its highest F1-score in Naïve Bayes. In contextual features, the tweet features have 78% highest accuracy and 77% its highest F1-score in SVM. The User features have 85 % highest accuracy is GradientBoost Classifier and 84% its highest F1-score.

We have also described the window-based features in which the results have been illustrated according to the specific window. Window-based features contained two different window-based features including previous window and temporal window as discussed in Section 4.1.3.

Previous Window Features

The previous window have different intervals of window range as shown in Figure 5.1. We have evaluated the classifications on every window, and compared with each other.

We choose the best window range for the proposed approach. We have computed the average of each window sliding ranges (5,10,15 and 20) which are 79.48, 79.51, 79.61 and 79.41 respectively. Accordingly, we have selected the sliding window 15 past tweets for the proposed methods because this sliding window performed better as compared to other windows. The highest accuracy of sliding window is 81.6 in

Classifiers		**Prec ision**	**Recall**	**F1-Score**	**Accu racy**
Content-based Approach					
Classifications	**SVM**	78%	76%	75%	76.20%
	NB	**80%**	**78%**	**76%**	**77%**
	LR	78%	76%	74%	75.60%
Ensemble Methods	**RF**	75%	75%	73%	74.60%
	ADABoost	79%	74%	71%	74.30%
	GradBoost	80%	75%	72%	74.60%
	Bagging	71%	72%	71%	71.50%
	XGBoost	80%	74%	71%	74%
	LightGBM	80%	74%	71%	73.40%
Tweet Features					
Classifications	**SVM**	**78%**	**78%**	**77%**	**78.0%**
	NB	75%	75%	75%	75.4%
	LR	77%	77%	76%	77.0%
Ensemble Methods	**RF**	76%	77%	76%	76.5%
	ADABoost	74%	74%	74%	74.5%
	GradBoost	76%	76%	75%	76.0%
	Bagging	73%	74%	73%	73.5%
	XGBoost	76%	76%	75%	76.0%
	LightGBM	77%	77%	76%	77.0%
User Features					
Classifications	**SVM**	86%	**85%**	**84%**	84.6%
	NB	86%	84%	84%	84.4%
	LR	85%	84%	84%	84.0%
Ensemble Methods	**RF**	85%	84%	84%	84.0%
	ADABoost	87%	85%	84%	84.7%
	GradBoost	87%	85%	84%	**85.0%**
	Bagging	84%	83%	82%	82.8%
	XGBoost	**88%**	85%	84%	85.0%
	LightGBM	86%	84%	83%	84.2%

Table 5.1: Results of individual features

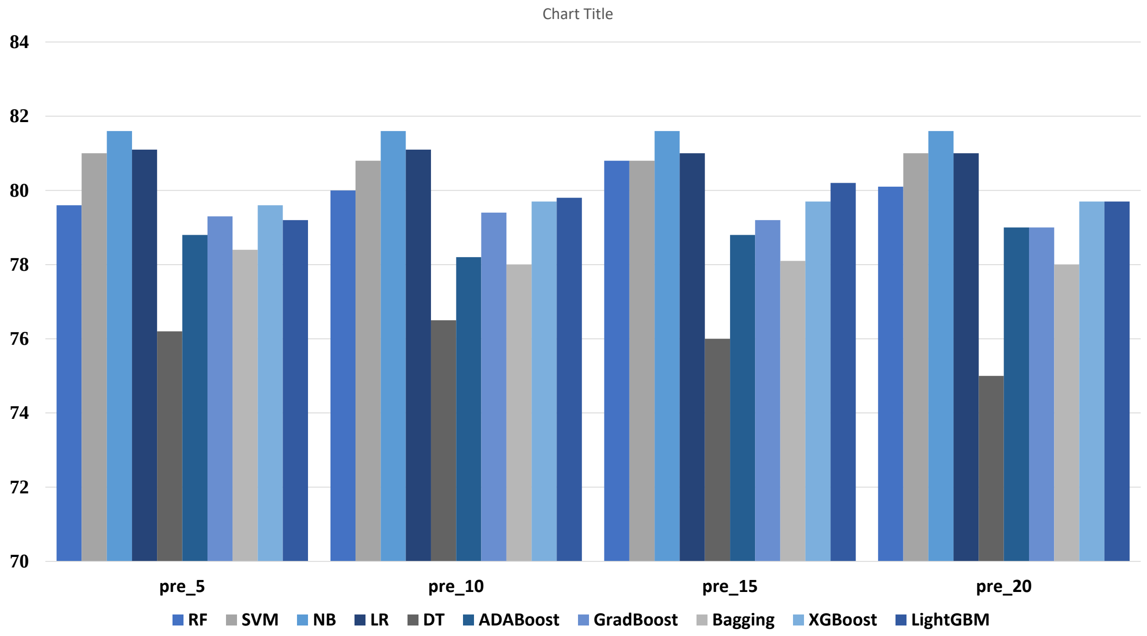


Figure 5.1: Variation of previous windows

Naïve Bayes Classifier as shown in Figure 5.1.

Temporal Window Features

First, we discussed the temporal window of past tweets. The different intervals of the time window have been used in Figures 5.2. We have arranged three different time intervals and found the average of each time window according to the classification methods.

Accordingly, we choose the best time window for the proposed methods. We have computed the average of each window sliding ranges (6 Hours,12 Hours and 2 days) which are 78.8, 78.3 and 79.3 respectively. Therefore, we have selected the 2 days time window because this time window perform better as compared to other. The highest accuracy of this temporal window is 82% in logistic regression (see in Figure 5.2).

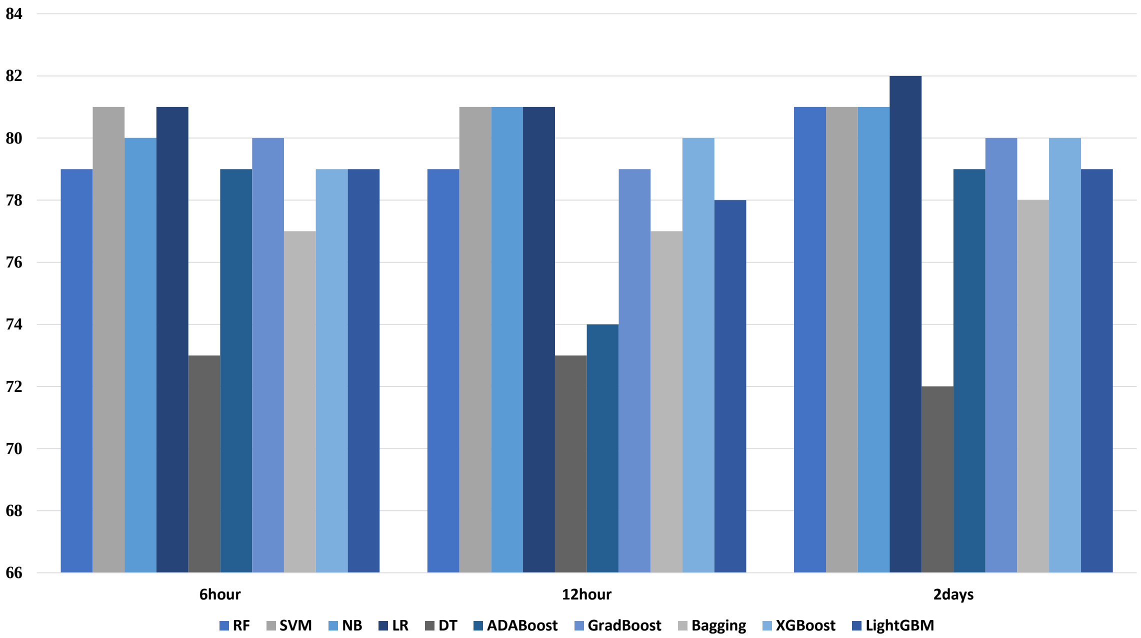


Figure 5.2: Variation of time windows

5.1.2 Combined Contextual Features

The performance of combining the contextual features which are the proposed approach. We have selected the features Which are based on their importance according to the pattern. Table 5.2 shows the results of combined features which out-performed the detection of abusive language than others.

These context features help the detection more accurately and confirmly. In consequence, the ensemble methods and linear classifications performed better. However, the highest accuracy of our proposed approach is 86% in Logistic regression Method.

Accordingly, we have elaborated and analyze the performance for the combinations of features. In these combinations, each contextual features performed better. The proposed approach is based on the performance of individual features and then combined these features are based on importance.

Table 5.3 shows the precision, recall and accuracy of individual classes of proposed method. The class abusive has 83% highest F1-score in LightBoost classifier and abusive class has 91% highest F1-score in AdaBoost classifier.

Proposed Approach Combined the contextual features with content-based approach					
Supervised Learning	Classifiers	Precision	Recall	F1-Score	Accuracy
Classification	SVMLinear	85%	84%	84%	84.0%
	Naïve Bayes	**88%**	85%	**85%**	85.3%
	Logistic Regression	88%	**86%**	85%	**86.0%**
	Random Forest	86%	85%	85%	85.0%
Ensemble Methods	AdaBoost	85%	84%	83%	84.0%
	GradientBoost	88%	85%	85%	85.3%
	Bagging	81%	81%	81%	81.4%
	XGBoost	88%	85%	85%	85.4%
	LightGBM	86%	84%	84%	85.0%

Table 5.2: Results of the proposed approach

Proposed Approach (Combined all contextual features)							
Supervised learning	Classifiers	Abusive			Non-abusive		
		Precision %	Recall %	F1-score	Precision %	Recall %	F1-score %
Classification	SVM	91	**75**	82	**85**	95	90
	NB	97	70	81	83	**99**	90
	LR	96	72	82	84	98	90
Ensemble Methods	Random Forest	96	72	82	84	98	90
	AdaBoosting	96	72	82	84	98	**91**
	GradientBoost	**98**	65	78	81	99	89
	XGBoost	97	71	82	83	99	90
	LightBoost	97	72	**83**	84	98	91
	Bagging	88	75	81	85	93	89

Table 5.3: Results of classes of the proposed approach

5.1.3 State-of-the-Art

We have evaluated two baseline papers on our dataset. The baseline papers are

- Automated Hate Speech Detection and the Problem of Offensive Language [Davidson et al., 2017].
- Comparative Studies of Detecting Abusive Language on Twitter [Lee et al., 2018b].

Table 5.4 illustrates the performance of baseline papers which we have implemented. The highest results of both papers are the first base paper; "Automated Hate Speech Detection and the Problem of Offensive Language [Davidson et al., 2017]" is 79.2% highest accuracy and the second baseline "Comparative Studies of Detecting Abusive Language on Twitter [Lee et al., 2018b]" is 75.1% highest accuracy.

Comparison	Base paper1: [Lee et al., 2018b] (Word-level)	Base paper1: [Lee et al., 2018b] (Character-level)	Base paper2: [Davidson et al., 2017]
SVM	68.70%	75.10%	69%
NB	64.90%	65.60%	74.80%
LR	64.50%	70.80%	**75.90%**
DT			67.90%
RF	69.10%	70.60%	74.70%
GradBoost	63.10%	65.30%	
CNN	77.30%	**75.60%**	
RNN	**79.20%**	56.30%	

Table 5.4: Comparison of the proposed approach with baseline papers

5.1.4 Comparison of Individual Features

We compared the results of the individual features with each other to evaluate the performance and importance of context features. Figure 5.3 illustrates the difference between each feature where the user features outperform as compared to content-based approach, tweets and window-based features. Consequently, the user features contained better pattern that easily distributes the abusive and non abusive tweets from the user features than other features.

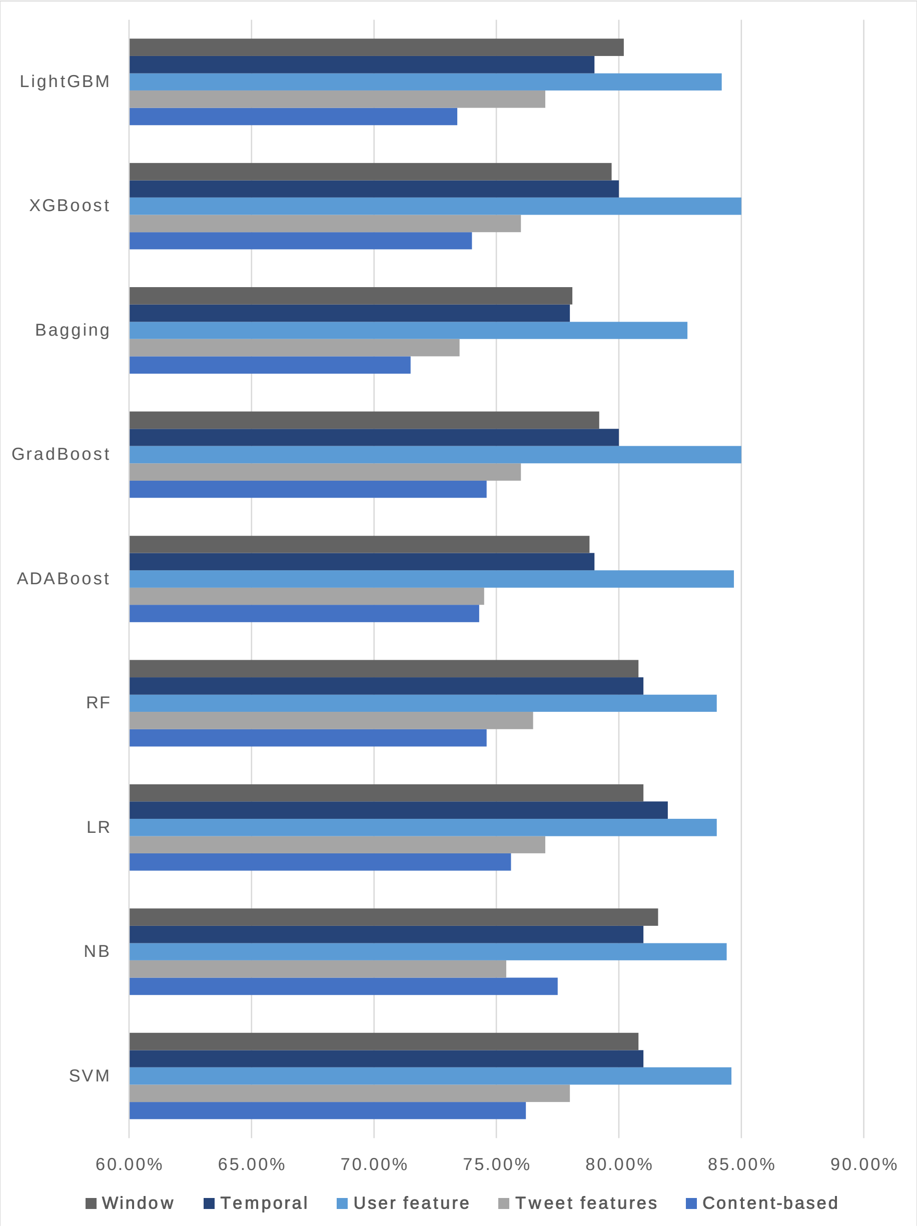


Figure 5.3: Comparison of the content-based with individual features

5.1.5 Comparison Proposed Approach with Content-based Approach

We have analyzed the results of the content-based and the proposed approach. The proposed approach outperforms as compared to content-based approach. The Figure 5.4 shows the difference between content and context which the highest difference is 8%. However, the contextual features help the abuse detection of content-based.

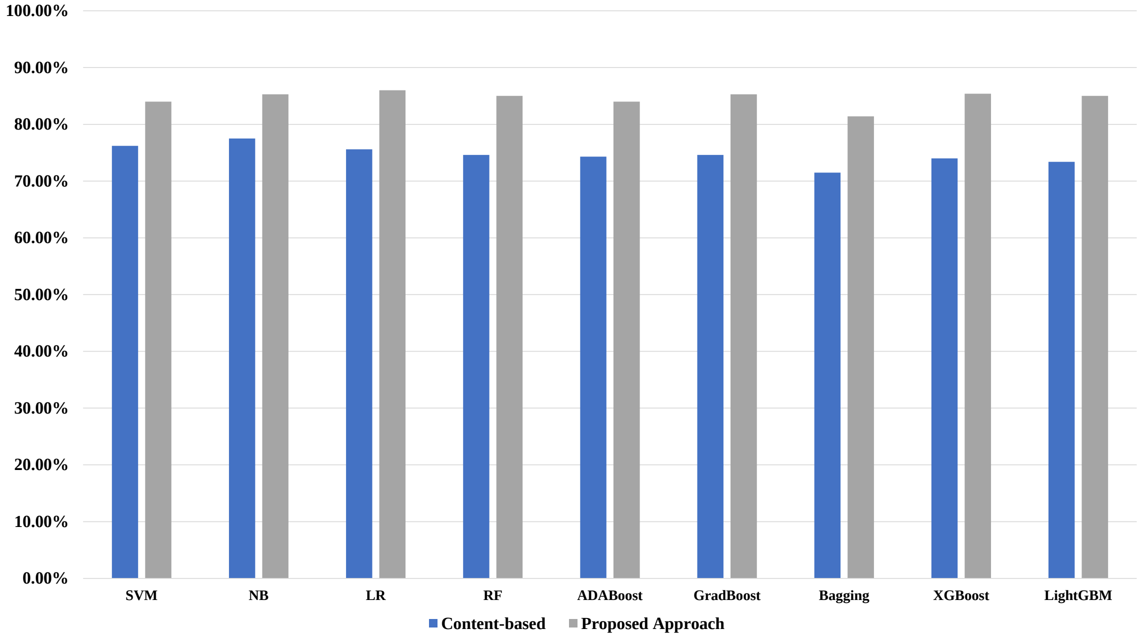


Figure 5.4: Comparison of the proposed approach with content-based approaches

5.1.6 Comparison with State-of-the-art

Finally, we have compared the performance of the proposed approach (combined content-based, user-based and window based features) with state-of-art as discussed in in Figure 5.5.

We have found significant difference between the accuracy of proposed approach and state-of-the-arts. The highest difference between the proposed method and baseline papers are 9% to 10% difference. However, the combination of the contextual

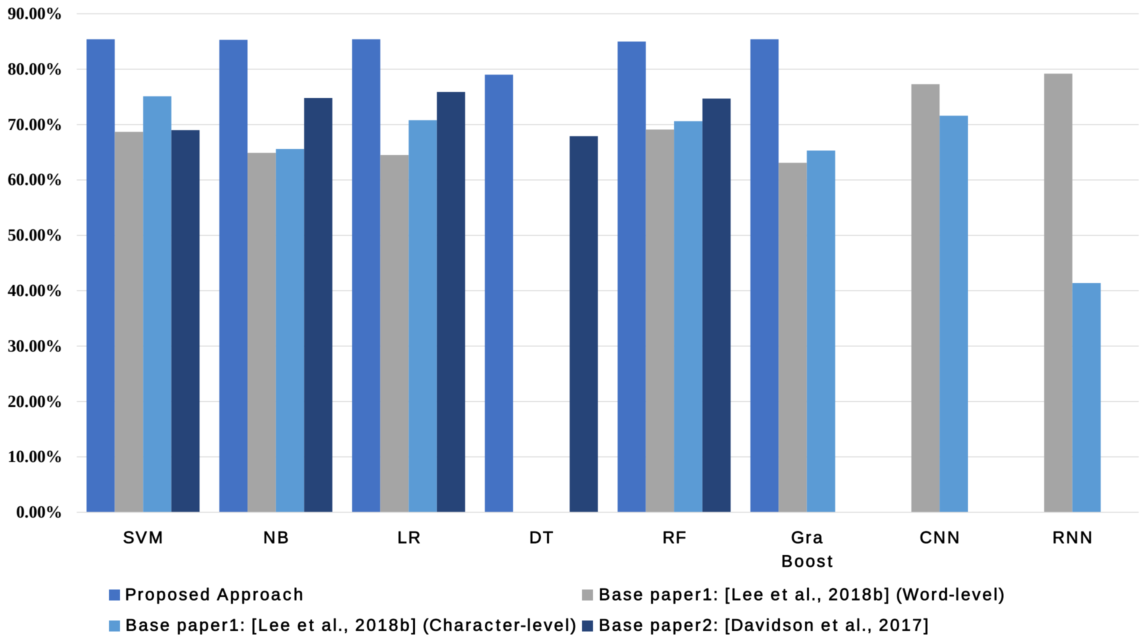


Figure 5.5: Comparison of the proposed approach with state-of-the-art

features approach outperforms the state-of-the-arts.

5.2 Discussions

In this section, we discuss the performance and important factor of abusive tweets detection. The discussion is about the results of each context features performs better than content-based approach and state-of-the-art. We discuss the contextual features support the performance of detection. The content-based does not fulfil the detection of abusive tweets. Moreover, we discuss the combination of features affects the detection.

RQ1: Do Contextual Features perform better than content-based features for the detection of abusive tweets?

The contextual features perform better as compared to content-based approach. These features help the content more prominently if the features have available to simplify the abuse and non-abuse properties. In content-based, sometimes the content cannot identify because the contents contain ambiguous words and sentences, and malicious users misused the other accounts. Therefore, this situation can be handled with use of the context features which help the detection of abusive tweets. The results of the contextual features are better than content-based. **RQ2: Which combinations of features are most significant to accurate the detection of abusive tweets?**

We evaluated the performance of individual features and acquired the important features of every feature set in our proposed approach. We analyzed and implemented the combination of individual features where we have selected the important features for our proposed methods. However, we combined different features of each part of the context features in which mostly features are based on their popularity. For example, followees, following retweets, etc. A few features based on their tweets and users which are included mentions, hashtags, URLs and User-descriptions.

In this case, the user features outperform than other features because all the attributes of user features show the popularity values well define an except User-descriptions. User features attributes simplify the abusive and non-abusive better. for example, the abusive users have the insignificant number of followers and list, etc.

In tweets features, the combination of tweets information features are based on tweets content and popularity attributes. Therefore, tweet features do not perform better. These features do not well to fulfill the pattern of abusive or non-abusive. But, some features generate over-fitting including mentions, hashtags and URLs. However, it effected the performance of tweets features.

In window-based features, these combinations of features are based on past tweets of current tweets in a dataset. These features can further perform well if all the past tweets are extracted. Unfortunately, the past tweets have not been available and inaccessible of every tweet.So, we acquired ta few number of past tweets but the performance of abusive detection within window-based is better. The purpose of these features are user to confirm and accurate detection of abuse in tweets, if the abusive tweets contain in previous tweets within time.

Gradually, we combined the context features which performed more effectively and accurately. These context features improve the performance of abusive language detection.

5.3 Summary

We have summarized the performance of our proposed method, where we illustrated the consequences of individual contextual features. The contextual features compared individually and co Individually, user features are more effectively as compared to other features. We identified the important combination of each features. the performance of the proposed approach is outperformed as compared to content-based.Furthermore, we have chosen two baseline papers, implemented and evaluated it on our filtered dataset but the proposed method outperform the state-of-the-art.

Chapter 6

Conclusions and Future Work

In this chapter, we discuss the conclusions, limitations and future direction of this thesis. Section 6.1 concludes this thesis. Section 6.2 discusses the limitations of the thesis and finally Section 6.3 lists down possible future directions.

6.1 Conclusions

The goal of this thesis is to detect abusive language using context-based features along with content-based features. In the previous studies, it has been detected in mainly three different ways which are word-, content- and context-based detection. These techniques have some limitations. Word-based detection faces challenges when the new words occur that are not available in wordlist or dictionary. Another limitation in the word-based approach is the confusing words that are in datasets which can not identify them accurately. In content-based detection, the limitations include the confusing words in contents which may lead to inaccurate detection. Recently, the context features including different additional features, like tweets features, user features and network features are used to detect abuse. Accordingly, the models depend on the patterns of features that can be easily distributed into either abuse or non-abuse. Mostly, the studies were not fulfilled to identify the patterns of features.

In our proposed approach, we have used three different contextual features that have been taken from tweets information, user information and past tweets. We combine various features which include popularity features and other features. In window-based features, we use two different features which include the previous sliding window and temporal window. We have considered many different intervals of the sliding window and temporal window. We use the best case of these intervals in the proposed approach. We have combined these context features and evaluated them by using Supervised Machine Learning algorithms. The algorithms include linear classification and ensembles methods.

We use an existing dataset to evaluate the performance of abusive tweets detection. The dataset developed in [Waseem and Hovy, 2016] has been used in our proposed work. In this dataset, we acquire different strategies for data collection. We have performed the filtration on the dataset also because the dataset is highly imbalanced and has missing data.

However, we have analyzed the results of our proposed method in which first, we have checked individually the combination of features and compared them. Afterwards, we have evaluated the combinations of features in which the performances are outperformed as compared to content and individual context features. We have compared our proposed methods with two baseline methods, the result show that our proposed method outperforms the state-of-the-art methods.

6.2 Limitations

In the research study, we use an existing dataset. As per Twitter's policy, tweets cannot be published in a dataset, but tweet ids can be published. Due to this policy, a few tweets were not available. In our proposed method, we retrieved past tweets, past tweets but a large number of past tweets of each tweet were not available. Therefore, we face the issue in window-based features, if the past tweets are available then these features may perform even better in detecting the abusive tweets. A few individual features affect significantly detecting the abuse but a few of these features are slightly affected for the detection of abuse. Due to unavailability of data, window-based features may suffer in identifying patterns for detection abusive tweets

6.3 Future Work

In this study have used many different contextual features. We can extend these features as follows:

- First, we can use different contextual features which include images and videos. We can transform the image content and video content into textual form.
- Second, we can use URLs and use the contents of referring webpages.
- Finally, we can also work on predicting abuse, instead of detecting it.

Bibliography

[Abozinadah and Jones, 2017] Abozinadah, E. A. and Jones, J. H. (2017). A statistical learning approach to detect abusive twitter accounts. *ACM International Conference Proceeding Series*, Part F130280:6–13.

[Baroni et al., 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

[Cecillon et al., 2019] Cecillon, N., Labatut, V., Dufour, R., and Linarès, G. (2019). Abusive language detection in online conversations by combining content-and graph-based features. *CoRR*, abs/1905.07894.

[Chakrabarty and Gupta, 2018] Chakrabarty, T. and Gupta, K. (2018). Context-aware attention for understanding twitter abuse. *CoRR*, abs/1809.08726.

[Chatzakou et al., 2017a] Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. D., Stringhini, G., and Vakali, A. (2017a). Hate is not binary: Studying abusive behavior of #gamergate on twitter. *CoRR*, abs/1705.03345.

[Chatzakou et al., 2017b] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017b). Measuring #gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1285–1290, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Chen et al., 2017] Chen, H., McKeever, S., and Delany, S. J. (2017). Abusive text detection using neural networks. In *Proceedings of the 25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7 - 8, 2017*, pages 258–260.

[Chen et al., 2018] Chen, J., Yan, S., and Wong, K. C. (2018). Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, 0:1–10.

[Choudhury and Breslin, 2010] Choudhury, S. and Breslin, J. G. (2010). User Sentiment Detection: A YouTube Use Case. *The 21st National Conference on Artificial Intelligence and Cognitive Science (AICS 2010).*

[Clarke and Grieve, 2017] Clarke, I. and Grieve, J. (2017). Dimensions of Abusive Language on Twitter. *First Workshop on Abusive Language Online,*, pages 1–10.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

[Diwanji, 2018] Diwanji, S. (2018). Responsibilities of abusive behavior. `https://www.statista.com/statistics/886180/india-public-opinion-on-social-media-abuse-liability/`. [Online; updated 23-sep-2019].

[Dollarhide, 2019] Dollarhide, M. (2019). Social Media Definition. `https://www.investopedia.com/terms/s/social-media.asp`. [Online; updated 02-may-2019].

[Duggan, 2017] Duggan, M. (2017). Online harassment 2017. pew research center.

[ElSherief et al., 2018] ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. M. (2018). Peer to peer hate: Hate speech instigators and their targets. *CoRR*, abs/1804.04649.

[Fehn Unsvåg and Gambäck, 2018] Fehn Unsvåg, E. and Gambäck, B. (2018). The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

[Founta et al., 2018a] Founta, A., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., and Leontiadis, I. (2018a). A unified deep learning architecture for abuse detection. *CoRR*, abs/1802.00385.

[Founta et al., 2018b] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018b). Large scale crowdsourcing and characterization of twitter abusive behavior. *CoRR*, abs/1802.00393.

[García-recuero, 2016] García-recuero, Á. (2016). Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications Alvaro To cite this version : Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications. *WWW*, pages 305–309.

[García-Recuero et al., 2018] García-Recuero, Á., Morawin, A., and Tyson, G. (2018). Trollslayer: Crowdsourcing and characterization of abusive birds in twitter. *CoRR*, abs/1812.06156.

[Gaydhani et al., 2018] Gaydhani, A., Doma, V., Kendre, S., and Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and TFIDF based approach. *CoRR*, abs/1809.08651.

[Gitari et al., 2015] Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

[Golbeck et al., 2017] Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., and Wu, D. M. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 229–233, New York, NY, USA. ACM.

[Hamilton et al., 2017] Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. *CoRR*, abs/1706.02216.

[Jindal and Liu, 2008] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA. ACM.

[Kshirsagar et al., 2018] Kshirsagar, R., Cukuvac, T., McKeown, K. R., and McGregor, S. (2018). Predictive embeddings for hate speech detection on twitter. *CoRR*, abs/1809.10644.

[Kwon et al., 2018] Kwon, S., Liang, P., Tandon, S., Berman, J., Chang, P.-j., and Gilbert, E. (2018). Tweety holmes: A browser extension for abusive twitter profile detection. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18, page 17–20, New York, NY, USA. Association for Computing Machinery.

[Lee et al., 2018a] Lee, H. S., Lee, H. R., Park, J. U., and Han, Y. S. (2018a). An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113(December 2017):22–31.

[Lee et al., 2018b] Lee, Y., Yoon, S., and Jung, K. (2018b). Comparative studies of detecting abusive language on twitter. *CoRR*, abs/1808.10245.

[Madisetty and Sankar Desarkar, 2018] Madisetty, S. and Sankar Desarkar, M. (2018). Aggression detection in social media using deep neural networks. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127.

[Mathur et al., 2018] Mathur, P., , Shah, R., , Sawhney, R., , and Mahata, D. (2018). Detecting Offensive Tweets in Hindi-English Code-Switched Language. *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 1–9.

[Nobata et al., 2016] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 145–153.

[Park and Fung, 2017] Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *CoRR*, abs/1706.01206.

[PEW, 2019] PEW, R. C. (2019). Social media definition. `https://www.pewresearch.org/internet/fact-sheet/social-media/`. [Online; updated 12-jane-2019].

[Pitsilis et al., 2018] Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Detecting offensive language in tweets using deep learning. *CoRR*, abs/1801.04433.

[Rajadesingan et al., 2015] Rajadesingan, A., Zafarani, R., and Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 97–106, New York, NY, USA. ACM.

[Ribeiro et al., 2018a] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., and Jr., W. M. (2018a). Characterizing and detecting hateful users on twitter. *CoRR*, abs/1803.08977.

[Ribeiro et al., 2018b] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., and Jr., W. M. (2018b). "like sheep among wolves": Characterizing hateful users on twitter. *CoRR*, abs/1801.00317.

[Saeed et al., 2019] Saeed, Z., Abbasi, R. A., Razzak, M. I., and Xu, G. (2019). Event detection in twitter stream using weighted dynamic heartbeat graph approach [application notes]. *IEEE Comp. Int. Mag.*, 14(3):29–38.

[Sharma et al., 2018] Sharma, A., Nandan, A., and Ralhan, R. (2018). An Investigation of Supervised Learning Methods for Authorship Attribution in Short Hinglish Texts using Char & Word N-grams. *CoRR*, abs/1812.1.

[Tahmasbi and Rastegari, 2018] Tahmasbi, N. and Rastegari, E. (2018). A Socio-contextual Approach in Automated Detection of Cyberbullying. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 1(4):1–22.

[Wang et al., 2016] Wang, Q., She, J., Song, T., Tong, Y., Chen, L., and Xu, K. (2016). Adjustable time-window-based event detection on twitter. In *Web-Age Information Management - 17th International Conference, WAIM 2016, Nanchang, China, June 3-5, 2016, Proceedings, Part II*, pages 265–278.

[Wang and Goutte, 2017] Wang, Y. and Goutte, C. (2017). Detecting changes in twitter streams using temporal clusters of hashtags. In *Proceedings of the Events* *and Stories in the News Workshop*, pages 10–14, Vancouver, Canada. Association for Computational Linguistics.

[Waseem et al., 2017] Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *CoRR*, abs/1705.09899.

[Waseem and Hovy, 2016] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

[Watanabe et al., 2018] Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access*, 6(c):13825–13835.

[Wiegand et al., 2018] Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). Inducing a Lexicon of Abusive Words – a Feature-Based Approach. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

[Zampieri et al., 2019] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

[Zhang and Luo, 2018] Zhang, Z. and Luo, L. (2018). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *CoRR*, abs/1803.03662.

[Zimmerman et al., 2018] Zimmerman, S., Kruschwitz, U., and Fox, C. (2018). Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).