Using Machine Learning to Predict the Targets of Hate Speech on Social Media

By

Sahrish Khan

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

Dedicated to

My Baba who supported me at every step of my life and in my educational career. Along with, to my supervisor, Dr. Rabeeh Ayaz Abbasi, for all the guidance and knowledge provided during the process.

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: September 8, 2020

Sahrish Khan

Abstract

Social media facilitates people having a diverse set of personalities to communicate with each other. People are free to communicate with others without any limitations. Occasionally such a communication results in to use of hate speech against others. Hate speech is the use of violent, aggressive, and offensive language. Though social media websites do not allow the use of hate speech, but the size of these platforms makes it nearly impossible to manage all their content. Consequently, several studies have been conducted for automatically detecting hate speech on social media. Focus of these studies is to detect the hateful content. Majority of these studies ignore predicting the target of the hate speech on social media. Focus of this study is to predict targets of hate speech. In this regard, firstly, a new balanced Hate Speech Targets Dataset (HSTD) is developed. HSTD contains tweets labeled for targets and non-targets of hate speech. Secondly, a novel framework Hate–speech Targets Prediction FrameworK (HTPK) is proposed to predict the targets of hate speech on social media.For this purpose, we have used machine learning algorithms. There are many algorithms used for prediction in machine learning but we have applied the algorithms used in binary class prediction to HTPK and chose the algorithms that performed best. Comparison with state-of-the-art methods shows that HTPK performs better than these methods.

Acknowledgments

All praises for ALLAH Almighty, for his countless blessings on me and his sacred Prophet Muhammad (P.B.U.H). Indeed I would not have been able to get any idea, about the problem and the solution presented in this dissertation, without ALLAH's will. It is indeed His guidance that enlightened my path in writing this dissertation. I want to thank my parents, Dadi Ammaa (Grandmother) and my brothers, who think that I have the caliber to do MPhil from this institute.

Changing my perspective towards university, first of all, I want to thank my supervisor Dr. Rabeeh Ayaz Abbasi, for his valuable guidance and continuous encouragement. At times when I totally lost hope and could not see any further, it was my supervisor who cleared the dust created by my disillusions and helped me to see out of the box in that situation of disguise. I am obliged to my supervisor for his support, motivation, continuous assistance and detailed review of this dissertation. Indeed his efforts have a vital role in this work. I would like to take this opportunity to thank everyone at my department, specially my teachers, for giving me a life time experience. I feel blessed to have teacher like Dr. Shuaib Karim, Dr. Onaiza Maqbool, Dr. Muddassar Azam Sindhu, Dr. Ghazanfar Farooq, Dr. Akmal Saeed Khattak, Dr. Umer Rasheed and Dr. Khalid Saleem. Finally, I would like to express my gratitude to all the people that supported my research in this topic..

To all the researchers that make their work freely available to others.

To the friends and roommates for the good caring words!

To all the people that have the courage to don't accept all the norms, and keep dreaming and fighting against oppression.

To the guard uncles, who give me countless prayers each time I greet them.

Thanks and Regards,

Sahrish Khan

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the age of constantly growing complexity and volume of the World Wide Web, social media has become an important part of life through which millions of users are able to express their current feelings, thoughts and beliefs. On one hand, it empowers the users to share their personal opinions at any topic. On the other hand, anti-social behavior like online trolling, cyber bullying, harassment and hate speech are also common on social media. Furthermore, misuse of freedom of speech on social media has become an issue all over the world. With the spread of online social networks hate speech has become even more serious. Hate speech in the context of social media is not only causing tension among people, but its effects can also lead to serious real-life controversies[1]. Unfortunately, the digital world can be used by haters as a safe haven to spread hate, but it causes unhappiness for the victims and affects their lives[2]. As the Online–Hate–Prevention–Institute (OHPI) CEO Andre Oboler observed "The longer the content stays available, the more damage it can inflict on the victims and empower the perpetrators. If you remove the content at an early stage, you can limit the exposure. This is just like cleaning litter, it doesn't stop people from littering but if you do not take care of the

[1] https://web.archive.org/web/20191101222256/https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/, September 8, 2020

[2] https://web.archive.org/web/20191228175948/https://www.adl.org/resources/reports/murder-and-extremism-in-the-united-states-in-2017/, September 8, 2020

problem it just piles up and further exacerbates"[3].

Therefore, social media like Facebook and Twitter provide mechanisms to report hateful contents [Twitter, 2019, Facebook, 2019]. However, on social media there are users from diverse backgrounds using different vocabularies related to hate speech. Thus, the task of hate speech detection remains challenging due to plenty of hateful content, the unavailability of benchmarks and lack of efficient approaches. For this reason, several studies have been conducted for detecting hate speech on social media [Djuric et al., 2015, Davidson et al., 2017, Bohra et al., 2018, Pratiwi et al., 2019]. To address these issues machine learning approaches have been used. The focus of this study is on exploring machine learning algorithms (ML) to predict the targets of hate speech on social media. We have used ML techniques because many tasks in prediction, classification, decision making and other fields that involve human intelligence have been automated by the use of ML technique.

1.1 Motivation

The challenge is not just to detect hate speech, but also, who is being targeted and why? There is limited research which addresses the issue of predicting and identifying the victims of hate speech. However, the targets of hate speech have been identified from hater posts in [Silva et al., 2016, ElSherief et al., 2018a]. The hater is a user who uses hate speech in his post against other users. Example of hate speech is given in the following tweets.

- "@user1 I hate you **nagger**, you make life difficult"
- "@user2 UN-TAG you **retard**"
- "@user3 idiot **ghetto**"

[3] `https://web.archive.org/web/20190813181243/https://ohpi.org.au/ohpi-quoted-in-a-unesco-report-on-online-hate/`, September 8, 2020

These tweets sound hateful and demeaning the user victim of the tweets. In above example the target has been identified by hater's post by analyzing these words **nagger**, **retard** and **ghetto**. In previous studies, the target has been identified and analyzed from this type of posts. Identifying the target of hate speech is a good contribution in the research field. However, predicting hate speech target is also an important issue that has not been worked out in the literature to the best of our knowledge. There may be various reasons for not addressing this problem. One of the reasons is the absence of benchmark datasets. To solve this problem, we have used supervised machine-learning (ML) algorithms, because supervised ML algorithms are helpful in resolving prediction issues. Though, such ML algorithms require labeled data which is difficult to have. Although, we have not found a label dataset about the tweets of hate speech victims in previous studies [Sharma et al., 2018b, Waseem and Hovy, 2016, ElSherief et al., 2018a, Sharma et al., 2018b, Tulkens et al., 2016, Ross et al., 2016, Park and Fung, 2017, Founta et al., 2018, McHugh et al., 2019]. So, in focusing on this problem we have addressed the issue of dataset in our study. The goal of this study is not only to predict the target of hate speech and to gather a target related dataset, but also to find out the reasons behind being the target of hate speech. In this study, we analyze that who can and cannot become the target of hate speech on the basis of their own tweet. We have chosen twitter because of its mostly publicly available data as compared to other social media like Facebook where posts are mostly private. Due to the absence of publicly available datasets that include those posts through which people become the target of hate speech, we create Hate Speech Target Dataset (HSTD) through Twitter API.

This work provides several important findings. Like, mostly people become the target of hate speech because of their own posts. If the posts contain anger, suspiciousness, criticism, hateful content, abusive speech and low emotional awareness then there are chances that in the replies of such posts the user can become the target of hate speech.

1.2 Research–Questions (RQs)

The following RQs have been address in this study.

RQ1: Can we develop a labeled dataset of hate speech targets?

RQ2: Which algorithms perform better for hate speech target prediction on Twitter?

RQ3: Can part of speech tags be useful in hate speech target prediction?

RQ4: Can we predict the target of hate speech by their tweets content?

RQ5: Why common users on social media are becoming the victims of hate speech?

1.3 Research Contributions

Regarding the RQs, the purpose of this study is to develop a framework that predicts the target of hate speech on Twitter. To achieve this we have set various goals which are as follows:

- **Finding Research Gap:** Do a thorough analysis of existing studies and identify the limitations contained therein and study the concepts in relation to these studies.
- **Developing Dataset:** Develop a dataset that can be used to achieve the goal of this study.
- **Designing a Framework:** Designing a framework that efficiently predicts the targets of hate on Twitter.
- **Evaluation:** By utilizing well–known bench-marks, evaluate the proposed frame work and compare the efficiency with existing approaches (state–of–the–art).

1.4 Outline of the Thesis

The remaining thesis is organized as follows:

Chapter 2 Provides a background of concepts and terms used in this thesis. Finally, it discuss the machine learning techniques and performance measures.

Chapter 3 reviews the literature regarding hate speech detection on social media. It describes the methodology used for systematic literature review. Afterwords, analysis is performed on data extracted from related res each work. Furthermore, issues regarding hate speech are explained by dividing them into groups. The last part of this chapter highlights the research gap.

Chapter 4 describes the dataset collection and annotation process. It provides an overview of the proposed Hate–speech Targets Prediction Framework (HTPK). Furthermore, data preprocessing and feature extraction techniques have been explained in detail.

Chapter 5 provides an overview of the state–of–the–art methods. Furthermore, experiments and evaluation measures have been conducted to verify the HTPK.

Chapter 6 concludes this thesis by addressing the research contributions and describes the future work.

1.5 Summary

This chapter presents an overview of social media, as it is a platform which made communication easier but it can also be used by haters as a safe place to spread hate. For this reason, several studies have been conducted for detecting hate speech on social media. Furthermore, this chapter describes the motivation and contribution of the study.

Chapter 2

Background

This chapter describes some terms and concepts that have been used in the literature and in our study on tackling hate speech issues on social media. To illustrate these concepts, this chapter is divided into three sections. Section 2.1 provides an overview of social media, particularly Twitter. Furthermore, the definitions of hate speech are narrated in the Section 2.2 from various sources. Section 2.3 and Section 2.4 define machine learning approaches and evaluation methodologies used to solve hate speech issues on social media.

2.1 Overview of Twitter

Social media is a medium of online communication it allows content sharing, collaboration and community based interaction in real time. Various platforms constitute social media, such as social bookmarking, social networking, forums, wikis and microblogging. Some of the most commonly used social media are Twitter, Facebook, LinkedIn, Wikipedia, Reddit and Google+ [1]. If viewed in terms of hate speech detection on social media, most of the studies [Gaydhani et al., 2018, Alfina et al., 2017, ElSherief et al.,

[1] `https://web.archive.org/web/20190330161339/https://whatis.techtarget.com/definition/social-media/`, last accessed on MAR 30, 2019

2018b] have used Twitter dataset for this purpose. Twitter is a microblogging site where users communicate in form of short texts called tweets[2]. There are some features in the tweet that make it unique from the rest of the social media posts. These features have attracted attention of researchers. The features are as follows:

- **General Tweets:** Short text that is posted on Twitter in the form of a tweet, and contains a photos or/and video[3].
- **Retweets(RT) and RT Count:** Retweet means a Twitter user shares tweet of another user. Through this they are giving credit to the original user. In addition, the original user's tweet also shows how often it is retweeted.
- **@Mentions:** In this a user mentions another user in his tweet with his username preceded by the @ symbol.
- **#Hashtag:** This is way in which the user is flagging a specific subject or topic to a tweet. Consequently, other users can search for this specific subject and perceive tweets on this specific subject.
- **@Replies and its counts:** In a tweet it can be seen how many times it has been replied and which users have replied.
- **Favorite counts:** This means how many users liked the tweet.

We have explained these features in the Table 2.1 with examples.

2.2 Hate Speech Definition From Various Sources

The concept of hate speech needs to be clearly stated so that suitable algorithms are developed to detect it. It is important to recognize that finding a complete and and

[2] `https://web.archive.org/web/20191028112704/https://www.lifewire.com/what-exactly-is-twitter-2483331/`, September 8, 2020

[3] `https://web.archive.org/web/20190927174207/https://help.twitter.com/en/using-twitter/types-of-tweets/`, last accessed on SEP 27, 2019

Table 2.1: Features Examples of Tweets

Features	**Example**
General Tweets	"@username1 go on my snap–chat fam "
Retweets(RT) and RT Count	"@ username2"RT@username1 go on my snap–chat fam"
@Mentions	"@username3@username1 is a nice person, follow her"
#Hashtag	#culturenight, #Obamacare etc

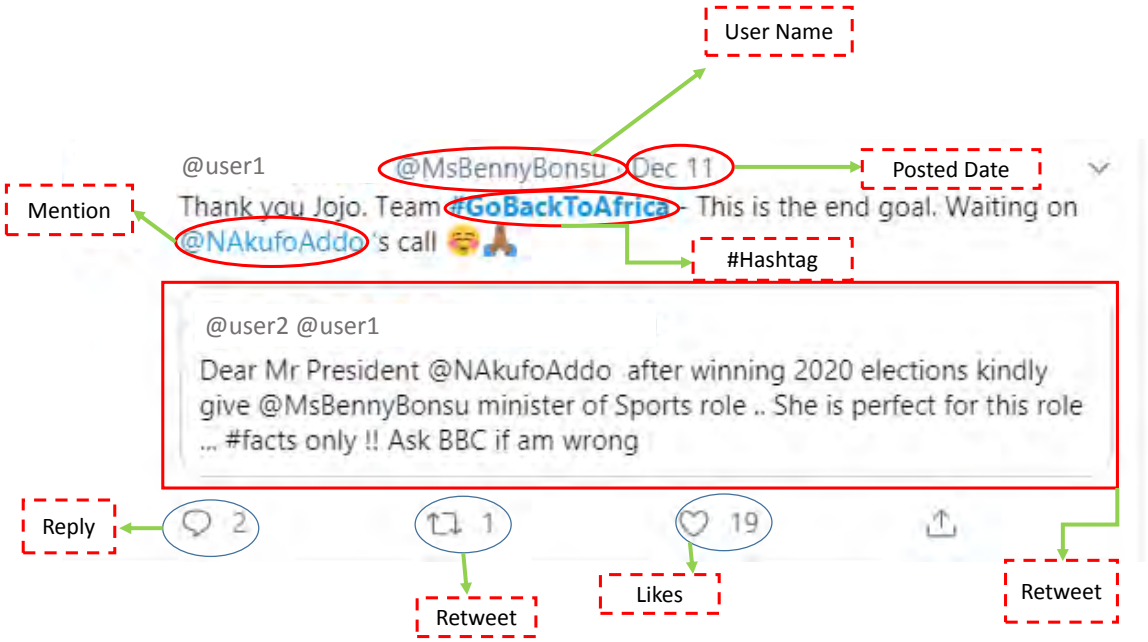

Figure 2.1: Features of Tweet

comprehensive definition or explanation of hate speech is a difficult task, because it is an ambiguous term and it needs different interpretations for itself. Various sources have described hate speech in their own way as shown in Table 2.2. Furthermore, we have analyzed hate speech definitions that make it clear that its purpose revolves around three points which are as follows:

- **Hate Speech (HS) is to threaten or attack:** Common point of all the definitions described in the Table 2.2 is that HS is intended to threaten or attack someone or a group.

- **HS is to provoke to hate or violence:** In addition, some definitions define that HS purpose is promote or provoke to hate and violence [Chetty and Alathur, 2018,

Table 2.2: Definitions of Hate Speech

Sources	Definitions
Twitter [2019]	"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories" [Twitter, 2019]
Facebook [2019]	"Direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation" [Facebook, 2019]
ECHR [2019]	"All forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin" [ECHR, 2019]
Chetty and Alathur [2018]	"Hate speech is any speech, which attacks an individual or a group with an intention to hurt or disrespect based on identity of a person" [Chetty and Alathur, 2018]

Twitter, 2019, ECHR, 2019]

- **HS has particular victims:** Coupled with, all the definitions highlight that HS has particular victims [Chetty and Alathur, 2018, Twitter, 2019, ECHR, 2019, Facebook, 2019]

We have described all these points in the Table 2.3. In short, we can say that hate speech is intended to humiliate, ridicule, and disrespect the targeted victims.

Table 2.3: An analysis of the content in HS definitions

Source	HS is to threaten or attack	HS is to provoke to hate or violence	HS has particular victims
Twitter	✓	✓	✓
Facebook	✓	–	✓
ECHR	✓	✓	✓
Chetty and Alathur	✓	✓	✓

2.2.1 Hate Speech Identification and Detection

For hate speech identification various social media platforms datasets have been used like Twitter, Facebook, Whisper, etc. Most of the studies have chosen twitter for dataset collection because of its public availability as compared to other social media platforms like Facebook where posts may be private. Additionally, tweet has some features like hashtag, location, mention, replies and retweet, which make Twitter unique among other social media platforms.

2.2.2 Why Study Identification and Prediction of Hate Speech Target?

Hate speech involves two parties, on is a hater and other one is victim. A hater is a social media user who uses hateful words in their post for another user, based on race, religion, ethnicity, gender, disability, gender or for some other personal reasons. And the user for whom this hate speech is being used is called victim or target of hate speech. In both of these parties, the target of hate speech is highly affected because his self respect is being hurted. Although, there has been considerable research by researchers on the detection of hate speech. But at the same time it is very important to identify and predict the victims of hate speech from their own posts. It is a enough motivation for us to work on the target's prediction. Because we haven't reviewed any work in literature to do on it yet.

2.2.3 Target Analysis and Identification

Hate speech can occurs in several formats and shapes victimizing at number of minorities and groups [Silva et al., 2016]. A large scale study has been conducted on hate speech target by using Whisper and Twitter Silva et al. [2016]. In this study they captured the frequency of common victims of hate speech on these platforms. The common types or groups of targets analyzed in this study are described below.

- **Behaviour** sensitive and insecure people.
- **Race** nigga, black and white people.
- **Class** generally people who belongs to lower class i.e. redneck, poorer.
- **Ethnicity** Jews, Pakistani, Indian.
- **Gender** generally belonging to discrimination towards women, sexism and transgender.
- **Disability** retarded people etc
- **Religion** Hinduism, Judaisms, Muslims etc
- **Physical** beautiful people, obese people
- **Other** shallow and alcoholic people

We have shown the frequency of hate categories found in data in Figure 2.2.

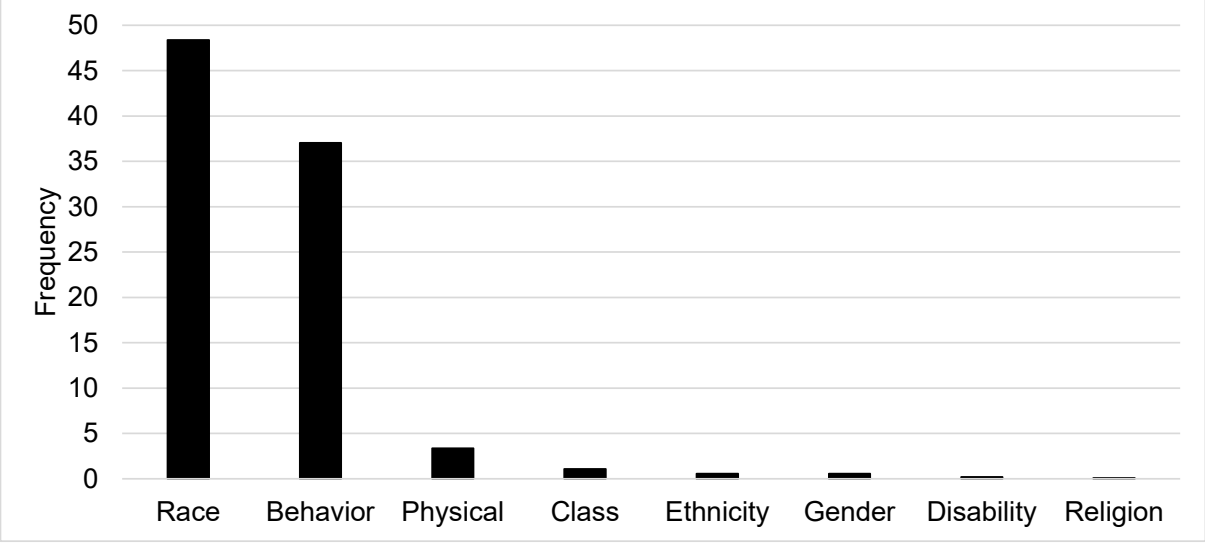

Figure 2.2: Various types of hate categorize in literature [Silva et al., 2016]

2.3 Classifications Approaches

While analyzing the sate of the art, our focus has been on the papers in which algorithms are being used. We have seen that in most studies, Machine Learning (ML) algorithms have been used to detect hate speech from text. Some studies have also used Deep learning (DL) algorithms, which are the specialized kind of ML. From the provided data, the ML classifiers identify patterns and then perform particular task by deciding on this information. Furthermore, in the regard of hate speech detection, supervised learning of ML algorithms has been used in existing studies. They are discussed as follows:

Supervised ML Approaches

We analyzed in the state of the art that supervised ML approach is being used to detect hate speech from text. This approach builds learning model and requires label data to train this model. The training–data contains set of instances that have various patterns and then their classes are created based on these patterns and each class is assigned a label. Consequently, the supervised ML approach makes predictions of new unseen data using a model which is obtained from training–data. These supervised ML algorithms have been used in existing studies: Support Vector Machine (SVM) [Gaydhani et al., 2018], Logistic Regression (LR) [Waseem and Hovy, 2016, Davidson et al., 2017, Gaydhani et al., 2018], Naive Bayes (NB) [Gaydhani et al., 2018], Decision Trees (DT), Gradient Boosting (GB) and Random Forest (RF) [Bouazizi and Ohtsuki, 2016, Alfina et al., 2017].

DL Approaches

Deep Learning is a specific from of ML and approach of Artificial Intelligence (AI). Deep learning is useful in scenario where training dataset is huge. In the regard of hate speech detection, this approach has also been used in existing studies. The most commonly

used algorithms are Recurrent Neural Networks (RNN) [Zhang et al., 2018], Convolutional Neural Network (CNN)[Zhang et al., 2018], Long Short-Term Memory network (LSTM)[Santosh and Aravind, 2019] and CNN+RNN [Badjatiya et al., 2017, Zhang et al., 2018, Chen et al., 2018, Mubarak and Darwish, 2019]

2.4 Evaluation Methodologies

It is necessary to evaluate the ML algorithms used for solving problems. For this purpose number of evaluation measures have been used in different studies. Most of the evaluation measures use a so called confusion matrix.

Confusion Matrix (CS):

This is used as a performance measure in ML problems. It is about describing the summary of prediction outcomes from classification problem. It explains number of incorrect and correct prediction values for each class. Table 2.4 describes the four various

Table 2.4: Confusion Matrix for Evaluation

Predicted Values	**Actual Values**	
	Negative	**Positive**
Negative	True Negatives (TN)	False Positives (FP)
Positive	False Negatives (FN)	True Positives (TP)

values used in the confusion matrix.

- **True Negatives (TN):** indicates the accurately classified non hate speech instances.
- **False Negatives (FN):** denotes the instances classified as non–hate speech which are actually hateful.
- **False Positives (FP):** is the number of classified hate speech which actually non–hate speech instances.

- **True Positives (TP):** is the number of accurately classified hate speech examples.

These values are helpful in measuring accuracy, recall, prediction and f1–measure[4] that have been used for evaluation in the literature.

- **Accuracy (A):** It is used to evaluate the overall efficiency of ML algorithms.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2.1}$$

- **Recall (R):** It describes as the ratio of the original positives which are accurately anticipated as positive.

$$Recall = \frac{TP}{FN + TP} \tag{2.2}$$

- **Precision (P):** Actually predict the ratio of positive cases.

$$Precision = \frac{TP}{FP + TP} \tag{2.3}$$

- **F–Measure:** It is used to calculate the harmonic mean of recall and precision.

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{2.4}$$

In addition to these measures, ROC (Receiver Operating Characteristics), AUC (Area Under The Curve) have been used in literature.

[4] `https://web.archive.org/web/20191023061722/https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62?gi=5269cf582112/`, September 8, 2020

2.5 Summary

The terms and concepts are explained in this chapter that have been used in the literature and in our study on detecting hate speech issues on social media. For this purpose, an overview of Twitter is provided. Afterwards, the definitions of hate speech from various sources are clarified. Finally, ML algorithms and evaluation measures are described.

Chapter 3

Related Work

This chapter presents systematic literature review in order to clearly describe the work done in the literature on hate speech issues on social media, specifically Twitter. Section 3.1 describes the systematic literature review. In the Section 3.2, we have analyzed the issues that have been worked out in the literature with regard to hate speech and we classify these issues. Finally, in Section 3.3 we describe the research gap.

3.1 Systematic Literature Review

The purpose of systematic literature review is to analyze and provide an overview of the research topics related to hate speech on social media, particularly those addressed in the field of computer science.

3.1.1 Methodology

A systematic literature review has been carried out to understand the research addressing hate speech detection on Twitter, we also focus on the targets of hate speech. Furthermore, we have analyzed feature extraction techniques, machine learning algorithms and evaluation measures used in detecting hate speech. The literature review method-

ology is based on the method proposed by Kitchenham et al. [2010]. Figure 3.1 gives an overview the whole process.

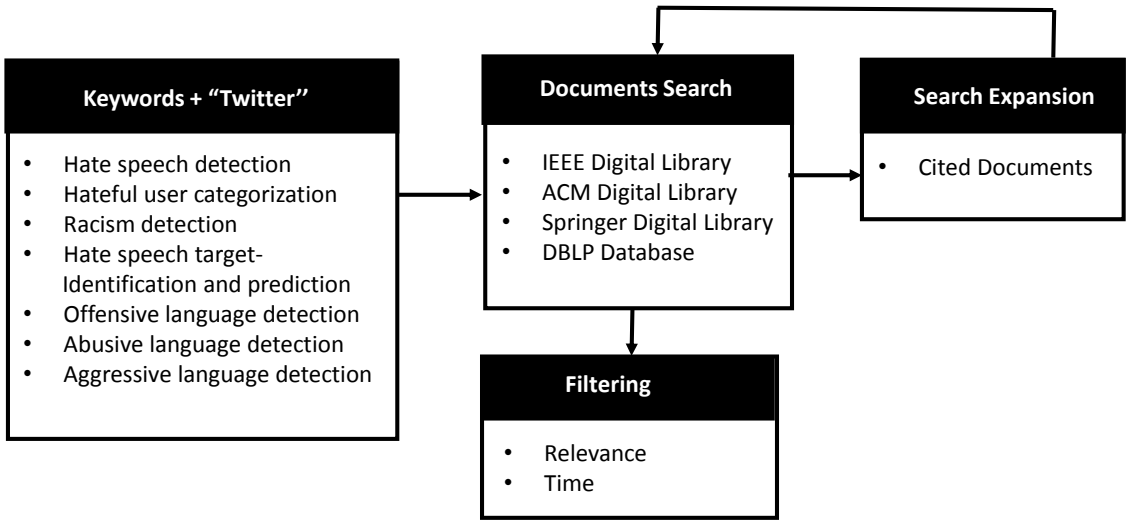

Figure 3.1: Methodology to carry out the systematic literature review

- **Selection of Keywords:** As we mentioned in the Section 2.2, there is no standard definition of hate speech. As a result, we searched documents based on various keywords. These keywords meet the popular definitions of hate speech. These keywords were used in conjunction with the keyword "Twitter".

- **Documents Search:** In order to collect the documents about hate speech, we have used IEEE Digital Library[1], ACM Digital Library[2], Springer Digital Library[3] and DBLP Database[4]. We have chosen these digital libraries, because most of the research articles related to computer science are indexed in them.

- **Search Expansion:** Using the keywords mentioned in the methodology, we search for documents in the mentioned digital libraries. Then we explore the citations of

[1] `https://web.archive.org/web/20200113182527/https://ieeexplore.ieee.org/Xplore/home.jsp/`, last accessed on September 8, 2020

[2] `https://web.archive.org/web/20200108081534/https://dl.acm.org//`, last accessed September 8, 2020

[3] `https://web.archive.org/web/20200112083957/https://link.springer.com//`, last accessed on September 8, 2020

[4] `https://web.archive.org/web/20200112070340/https://dblp.uni-trier.de//`, last accessed on September 8, 2020

these documents. Subsequently, we search for the documents in which they are cited. And then in terms of the keywords, we have extracted the documents we find relevant for our study. We have extracted only those documents in which the provided keywords were relevant to our research.

- **Filtering:** We have gathered relevance based, time based, and text-based technical and theoretical documents from 2010 to 2019. We started collecting documents from 2010 because we couldn't find documents discussing hate speech on Twitter prior to that.

3.1.2 Data Extraction

After collecting the number of documents, we have extracted 66 documents relevant for our study, most of them are from 2016 to 2019. The following measures have been used to analyze them:

- Approaches Used
- Temporal analysis
- Citation analysis
- Distribution of keywords
- Social platforms used
- Languages and dataset used
- Data preprocessing techniques analysis
- Features extraction techniques
- Classification methods
- Evaluation measure

We depict these measures in Figure 3.2. Their details are given in the following sections.

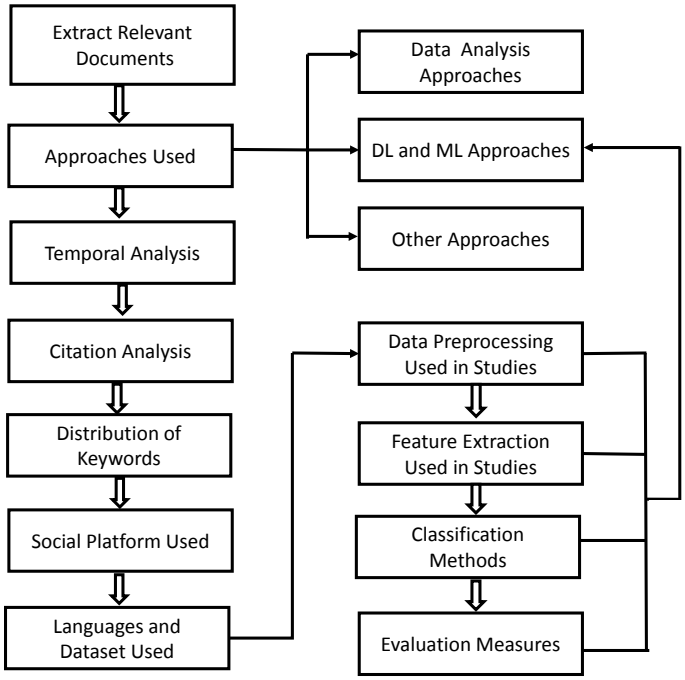

Figure 3.2: Data extraction for literature review

3.1.3 Approaches Used

We have categorized the hate speech detection related documentation based on the approaches used in literature as follows:

- **Data Analysis Approaches (DA):** Here we have analyzed the data used as well as the methods that are used to develop the new datasets.
- **DL and ML Approaches:** In ML and DL, various feature extraction techniques have been used with algorithms and number of evaluation measures are used to evaluate the performance of approaches.
- **Other Approaches (OA):** These approaches do not belong to either ML, DL and DA.

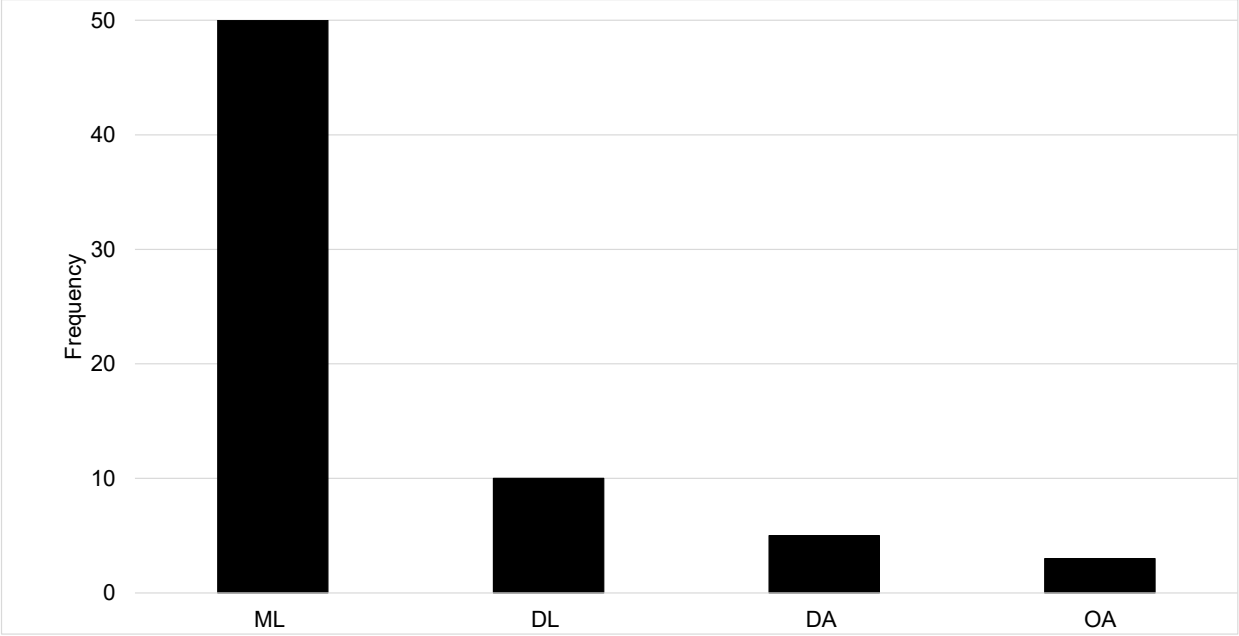

Figure 3.3: Frequency of approaches used for hate speech detection

Though, these approaches are well described. We have analyzed that some of these approaches have also been used in combination. As in some studies the DA, DL and ML approaches have been used together [Waseem and Hovy, 2016]. Figure 3.3 shown that ML is the most used approach compared to DL, DA and OA.

3.1.4 Temporal Analysis

We temporally analyze the hate speech literature and visualize the cumulative frequency of publications Figure 3.4, publications regarding hate speech are increasing from 2010 to 2019. Furthermore, we observe a steep growth 2015 on words.

3.1.5 Citations Analysis

We extract the citations of each document from Google Scholar[5], while analyzing documents for hate speech and we state the five most cited paper in the Table 3.1.5. In

[5] `https://web.archive.org/web/20200105114819/https://scholar.google.com//`, last accessed on September 8, 2020

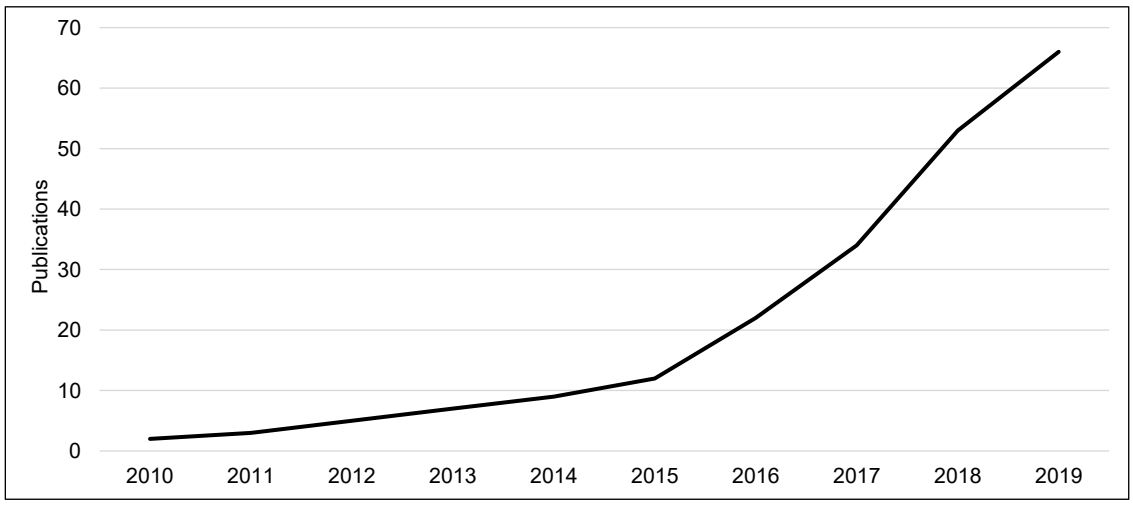

Figure 3.4: Temporal Frequency of publications regarding hate speech on social media

addition, we have analyzed that these papers were more cited in the year of 2018 and 2019. Furthermore, the purpose of analyzing these citations was to illustrate how much work has been done in the literature regarding hate speech on social media.

Table 3.1: Most cited publications in the field of computer science

Reference	**Citations**	**Title**
Davidson et al. [2017]	343	Automated hate speech detection and the problem of offensive language
Waseem and Hovy [2016]	294	Hateful symbols or hateful people? predictive features for hate speech detection on Twitter
Djuric et al. [2015]	229	Hate speech detection with comment embeddings
Badjatiya et al. [2017]	201	Deep learning for hate speech detection in tweets
Silva et al. [2016]	105	Analyzing the targets of hate in online social media

3.1.6 Distribution of Keywords

We extract the keywords provided in the documents and then we categorize them based on their domain. We ignore the documents which do not explicitly define keywords. These categories and their related keywords are described in Table 3.2.

Additionally, we create a word cloud for all of these keywords. As shown in Figure

Table 3.2: Distribution of keywords in various studies

Categories	Keywords
Social Networks	Facebook, Online social networks, Micro Blogging, microblogging Websites, Twitter, Social media, online posts, social networking and Internet
Hate–Speech (HS) and Offensive Language	Offensive language, hate speech, hateful offensive expressions, misuse of freedom of speech, abusive language, cyberbullying, violent language, hateful language, cyber conflicts, sarcasm, aggressive language, violence, violent language, flame detection, sensitive topics, Hate speech recognition, hate expressions, demeaning words profane words, social tension and hateful contents
Natural Language Processing (NLP)	Linguistic analysis, task analysis, comments analytical, multiclass sentiment analysis and sentiment analysis
Features Extraction	Feature extraction, named entities, topic extraction, web mining trending world events, topic similarity , tweets, text summarization, hashtags, pattern-based approach
Machine Learning (ML)	Text classification, Pattern classification, Naive Bayes, Support Vector Machines, J48,classification, Supervised Learning, RBF kernel, and Machine learning
Deep Learning (DL)	Long Short-Term Memory network (LSTM), Neural language models, Recurrent Neural networks (RNN), Convolutional Neural Network (CNN) and CNN+RNN

3.5, It can be seen that machine learning, hate speech, offensive language and social language keywords have a higher frequency. In terms of categories, the keywords with the highest frequency are shown in Figure 3.6. We have analyzed that in most publications the keywords of hate speech category are given. Then, the keywords that are most mentioned are ML, social networks and NLP category. Coupled with, all other keywords are correlated with hate speech topic.

3.1.7 Social Platforms Used

As we have clearly stated in Chapter 2 that If viewed in terms of hate speech detection on social media, most of the studies [Gaydhani et al., 2018, Alfina et al., 2017, ElSherief et al., 2018b, Waseem and Hovy, 2016, ElSherief et al., 2018b] have been used Twitter for this

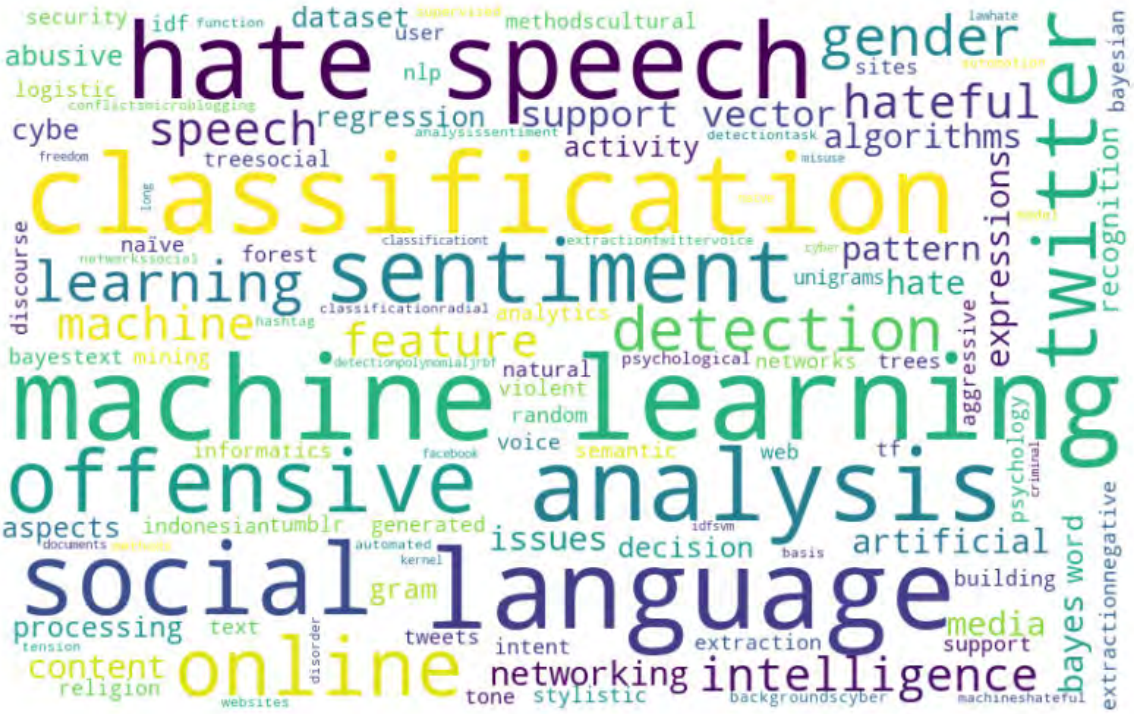

Figure 3.5: Keywords frequency

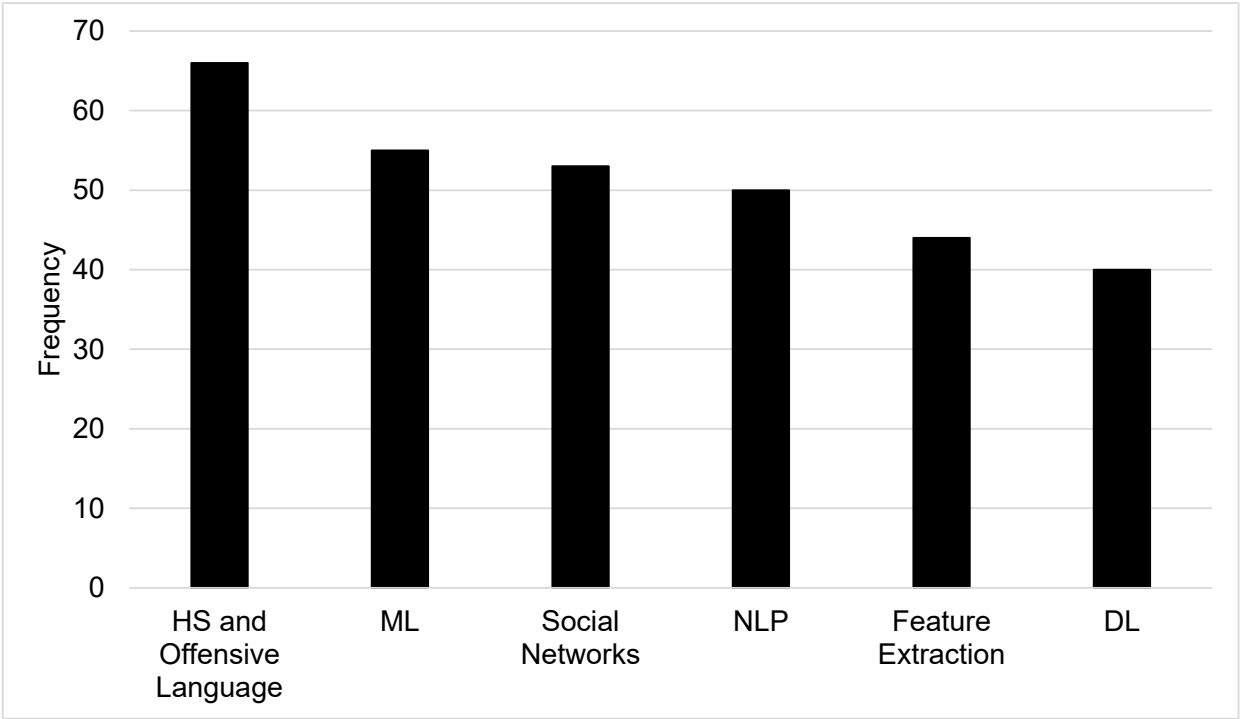

Figure 3.6: Categories of keywords

purpose. There are some features in the tweet that make it unique from the rest of the social media posts. This is why researchers prefer Twitter dataset. In addition, Facebook [Vigna et al., 2017], Whisper [Silva et al., 2016], YouTube [Anagnostou et al., 2018] and sites [Kurniasih et al., 2018] have also used. As shown Figure 3.7, the most commonly used social media platforms is the Twitter.

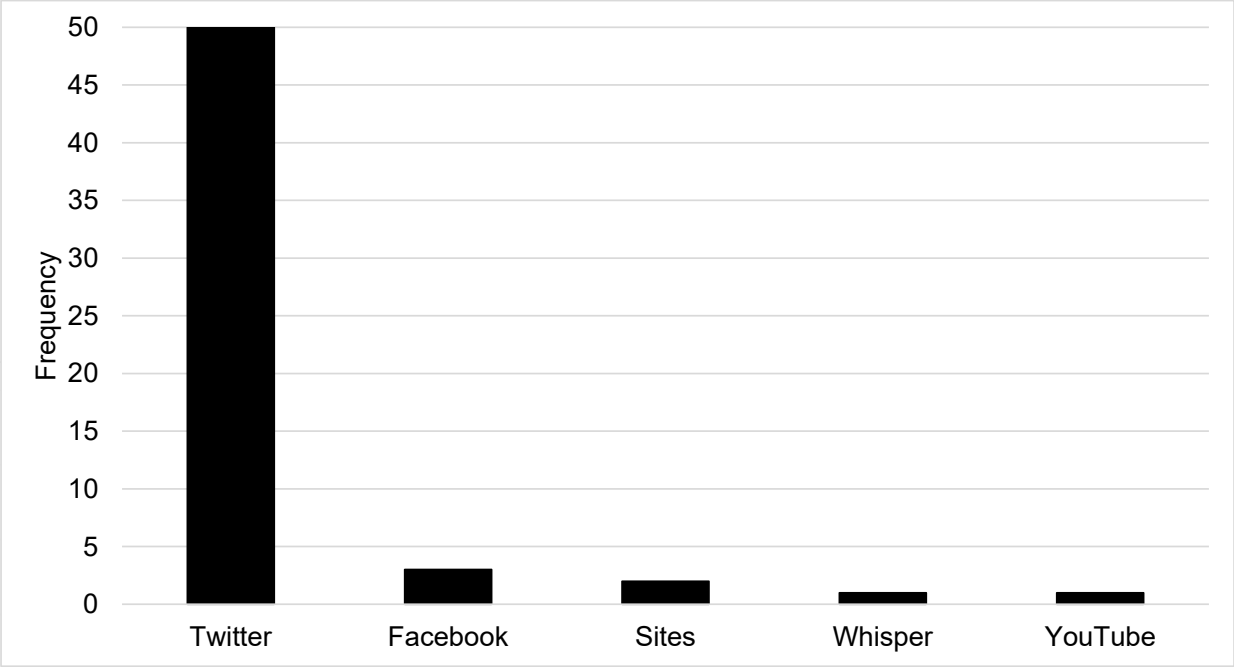

Figure 3.7: Social media platforms used in literature regarding hate speech detection

3.1.8 Languages and Datasets

In this study, it is analyzed that the dataset has a particular significance in all the studies conducted in the field of computer science regarding hate speech or its target detection, prediction and identification. Furthermore, various social media platforms have been used for the collection of dataset as described in Section 3.1.7. In addition, one issue that we notice regarding the hate speech datasets is their availability. The proposed methods and results in the previous studies are hard to compare in absence of benchmark. Correspondingly, the absence of publicly available datasets may be due to privacy issue or content contained in the dataset such as abusive and offensive words. Although there is

a decent collection of data utilized in the previous studies as described in Table 3.3. The language most commonly used in the datasets for detecting hate speech is English because it is the most spoken language on the internet[6]. Other languages have also been used. We have shown the frequency of all these in Figure 3.8, as can be seen that the most commonly used language is English.

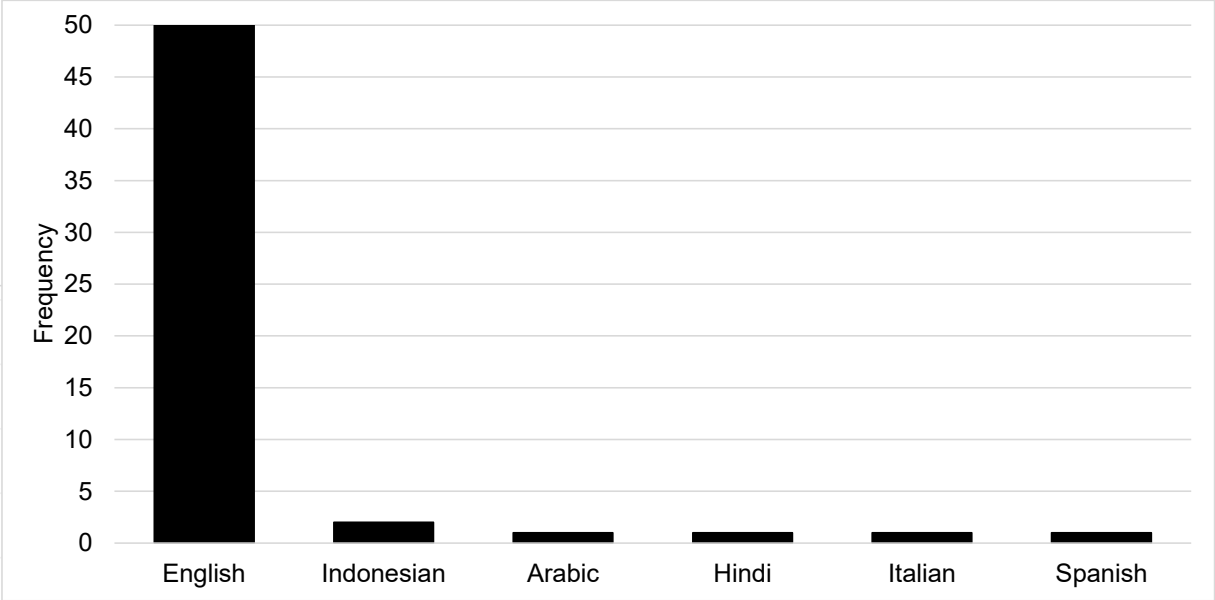

Figure 3.8: Frequency of languages used in literature

3.1.9 Data Pre–Processing

Data pre–processing is a process which process input data into a from which is useful for ML algorithms. The data preprocessing techniques play an important role in improving the performance of any ML algorithm. That's why it is necessary to have clean data for ML algorithms. There is too much language diversity on Twitter in some languages, Noise is a major issue on Twitter due to the informality, containing useless characters, short sentences, abbreviations and slang. The following preprocessing techniques are used in various studies.

[6] `https://web.archive.org/web/20191116202241/https://speakt.com/top-10-languages-used-internet//`, last accessed on September 8, 2020

Table 3.3: Corpus and datasets used in existing studies for hate speech and target detection

Paper	Name	Number of Classes Used	Language	Availability
[Sharma et al., 2018c]	Harmful speech	threatening, extremism, trolling	English	Not available
[Waseem and Hovy, 2016]	Hate speech Twitter annotations	Racist, sexist	English	GitHub
[Waseem and Hovy, 2016]	Hate speech Twitter annotations	Racist, sexist, both,neither	English	GitHub
[Davidson et al., 2017]	Offensive language and hate speech	Not mentioned	English	Available for Community
[Mubarak and Darwish, 2019]	Abusive language detection	Obscene and Clean	Arabic	Not Available
[Pratiwi et al., 2019]	Indonesian tweets	Hate Speech(HS), Non–HS	Indonesian	Non available
[Santosh and Aravind, 2019]	Hindi-English	hate, non–hate	Hindi–English	GitHub
[Bohra et al., 2018]	Hindi-English	hate, non–hate	Hindi–English	GitHub
[Nugroho et al., 2019]	Hate speech identification	HS, Offensive, Not–offensive	English	Available for Community

- **Lowercasing:** It is a process in which a stream of text is converted into lowercase. This technique helps to improve the performance of classification as it makes the data easier to understand and it decreases the dimensionality. Not using this technique can cause problem such as "HATE", "HaTe" and "hate" being considered several words. This technique is used in [Alfina et al., 2017, Gaydhani et al., 2018, Plaza-del Arco et al., 2019b]. Not all of irrelevant characters (e.g. !?&%) are helpful in classification tasks, so they are eliminated as well as seen in [Zhang and Luo, 2018a, Stammbach, 2019]

- **Tokenization:** Tokenization means splitting the words. Tokenization may different depending upon the language. This is a common technique for preprocessing, which is used in the recent publications [Gaydhani et al., 2018, Plaza-del Arco et al., 2019b, Watanabe et al., 2018a].

- **Lemmatization:** Lemmatization is a morphological examination of words. This means any word is lemmetized by going to its root. Such as the word like, liking, likely, liked lemmetize into "like". It is used in [Watanabe et al., 2018a].

- **Stemming:** In stemming, words are reduced to their base form such as "studies" to "studi", "likes" to "like". The use of this technique improves the classification of text. As it reduces the data dimensionality. It is used in [Gaydhani et al., 2018].

- **Stopwords removal:** Stopwords are the most common words used in language and are the part of almost every sentence (e.g: "and", "a" , "this"). Therefore, these words are not considered good for text classification. They have been removed in [Alfina et al., 2017, Gaydhani et al., 2018, Mishra et al., 2018].

- **Punctuation removal:** Generally punctuation is not considered important for text classification. So it has been removed in [Alfina et al., 2017].

3.1.10 Features Extraction

We have analyzed that most studies have used features extraction to improve the performance of classifiers. The most common features are described below.

- **Token Frequency Based Features:** The features that are initially used in some studies to address these issues were Bag of Words (BoW) [Burnap and Williams, 2016] and dictionaries [Dadvar et al., 2013], which are the type of token frequency features. But it has been analyzed that these features could not perceive the context of the phrase. Therefore, Ngram is used to overcome this drawback, which performed better than BoW in terms of performance [Alfina et al., 2017, Nobata et al., 2016].

  In addition, in some studies TF-IDF (Term Frequency – Inverse Document Frequency) is used as a feature extraction [Gaydhani et al., 2018]. a TF-IDF is a vector that described how important a word in the dataset.

- **Topic based Features:** In addition, topic based features have used such as topic extraction or topic similarity [Liu and Forss, 2014, Liu and Forss, 2015]. The topics in the document are identified using this feature such as religion, race, etc.

- **Sentiment based Features:** This feature is used to identify text sentiment. It checks and uses the different opinion degrees defined in any text such as positive, negative and neutral. [Bouazizi and Ohtsuki, 2016, Agarwal and Sureka, 2017].

- **Linguistic Study Features:** The Linguistic study features are those in which the sentence structure is used to obtain information such as part of speech (PoS) tagging [Watanabe et al., 2018b], rule based approach [Haralambous and Lenca, 2014] and typed dependencies [Burnap and Williams, 2016]. The use of this features make it easier to understand the sentence context.

- **Sentence based Features:** These features include counts of emoticon, capital letters, punctuations, hashtags, mentions, URLs and message length [Davidson et al., 2017, Watanabe et al., 2018a, Davidson et al., 2017, Alfina et al., 2017, Rodríguez et al., 2019].

- **Ontological Features:** Ontological features have also been used in some studies. As Waseem et al. [2017] presented the topology regarding abusive language detection. Their topology categorizes abusive language into four types depending upon their implicit/explicit and direct/generalized nature.

Features used in the literature are summarized in Table 3.6.

3.1.11 Classification Methods

Machine learning (ML) algorithms are of particular significance when it comes to detecting hate speech. In literature, studies have taken advantage of ML algorithms in detecting and identifying hate speech and its targets. With in ML, supervised ML algorithms have been used more regularly. Some of these algorithms are the ones that perform well on Twitter dataset in terms of performance and these algorithms are Logistic Regression (LR) [Waseem and Hovy, 2016, Davidson et al., 2017, Gaydhani et al., 2018], Support Vector Machine (SVM) [Gaydhani et al., 2018], Random Forest (RF) [Bouazizi and Ohtsuki, 2016, Alfina et al., 2017], Naive Bayes (NB) [Gaydhani et al., 2018], Decision Tree (DT) [Pratiwi et al., 2019, Aulia and Budi, 2019], Gradient Boosting (GB) [Ribeiro et al., 2018] and AdaBoost (AB) [Nugroho et al., 2019].

Coupled with, in the regard of hate speech detection, DL approach has also been used in existing studies. The most commonly used algorithms are Recurrent Neural Networks (RNN) [Zhang et al., 2018], Convolutional Neural Network (CNN) [Zhang et al., 2018], Gated Recurrent Units (GRU) [Zhang et al., 2018], Long Short-Term Memory network (LSTM) [Santosh and Aravind, 2019] and CNN+RNN [Badjatiya et al., 2017, Zhang

et al., 2018, Chen et al., 2018, Mubarak and Darwish, 2019]. To sum up, we have analyzed that the most commonly used ML algorithms with better performance are SVM, NB, RF and LR. Similarly, for DL, LSTM, CNN, GRU and RNN are reported in literature. We have depicted a summary of all these algorithms that have been used in several studies to detect hate speech on Twitter.In Figure 3.9 it shown that SVM, NB and RF are the most commonly used algorithms in literature.

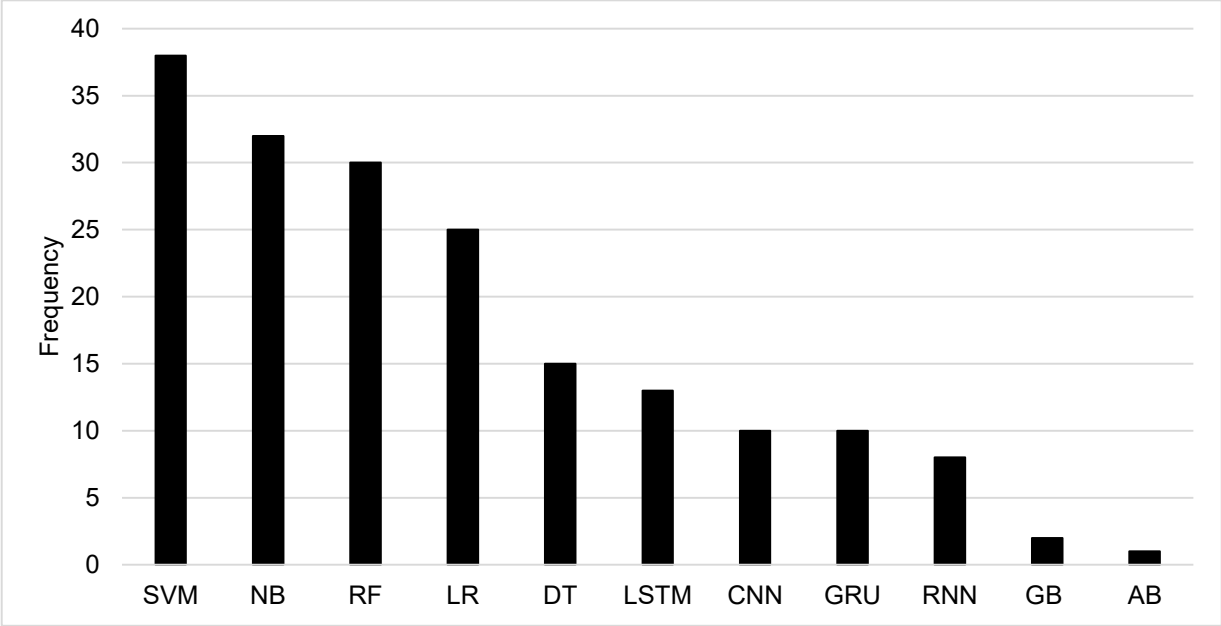

Figure 3.9: Frequency of algorithms used in literature regarding hate speech detection on social media

3.1.12 Evaluation Measures

Various evaluation measures have been used in hate speech regarding studies to quantify the performance of described algorithms and feature extraction. These measure in include confusion matrix, Accuracy (A) [Alfina et al., 2017, Gaydhani et al., 2018, Santosh and Aravind, 2019], Precision (P) [Gaydhani et al., 2018, Badjatiya et al., 2017, Davidson et al., 2017, Silva et al., 2016, Burnap and Williams, 2016, Zhang et al., 2018], Recall (R)[Gaydhani et al., 2018, Badjatiya et al., 2017, Davidson et al., 2017, Burnap and Williams, 2016, Zhang et al., 2018, Şahi et al., 2018, Santosh and Aravind, 2019], Weighted

average (W–a) [Alfina et al., 2017], weighted micro (W–m)[Badjatiya et al., 2017] and F1–score (F1) [Gaydhani et al., 2018, Badjatiya et al., 2017, Davidson et al., 2017, Burnap and Williams, 2016, Zhang et al., 2018, Santosh and Aravind, 2019] and AUC (Area Under The Curve) [Djuric et al., 2015]. Evaluation measures are used in literature are summarized in the Table 3.6.

3.2 Categories of Problems Related to Hate Speech

In this literature review we have studied issues related to hate speech. We categorized them into the following:

- **P1:** Detect hateful tweet
- **P2:** Detect and categorize hateful user
- **P3:** Identify and analyze targets of hate from hater's tweets

The three problems mentioned above are being addressed in various studies according to the literature review.

3.2.1 *P1:* Detect Hateful Tweet

If ones looks hate speech detection issue on Twitter, different approaches have been proposed to detect hate speech from tweets in order to address this problem [Plaza-del Arco et al., 2019a, Pratiwi et al., 2019, Bohra et al., 2018, Şahi et al., 2018, Sharma et al., 2018c]. For this purpose, labeled data is used that contains two or three classes (hate, non–hate, neutral) and most of the data is taken from Twitter[Waseem and Hovy, 2016, Sharma et al., 2018a, Tulkens et al., 2016, Ross et al., 2016]. To address this problem, ML and DL algorithms have been used along with the use of various features sets, such as ngram, BoW, content based, etc. this increase the efficiency of algorithms [Burnap

and Williams, 2016, Gaydhani et al., 2018, Liu and Forss, 2015]. We have illustrated this problem in the Table 3.5 by example.

3.2.2 *P2:* Detect and Categorize Hateful Users

In this problem hate full users have been categorized and detected on Twitter. For this purpose, the content shared by users, their activities and network structure are examined and compared with common users. Consequently, the results show that these users have recent accounts, increasing number of followers and their posts contain more profane and negative words [Ribeiro et al., 2017, 2018].

3.2.3 *P3:* Identify and Analyze Targets of Hate Speech from Hater's Tweets

This problem is about identifying and analyzing hate speech targets from hater's posts. The studies addressed this problem by analyzing the hater's posts. [ElSherief et al., 2018b]. For example this is a hater's tweet "@user1 is idiot" in this tweet "@user1" is targeted by the hater.

3.3 Research Gap

We found a research gap in literature to predict the targets of hate speech based on victims tweets, we term this problem as problem **P4**. This problem is unique among the problems of the rest of the studies and has not been worked out. It was a motivation for us to work on the target's prediction. Because we were unable to find any work in literature addressing this particular issue. Table 3.4 summarize the literature and problems addressed. To illustrate the concept of these issues, given examples of these problems in the Table 3.5. In addition, we have depicted the frequency of problems related to hate speech in Figure 3.10. It is observed that, most of the work in literature is on the detec-

Table 3.4: Comparison of proposed work with existing studies

Reference	Detect Hateful Tweets	Detect Hateful Users	Detect or Predict Targets from Haters' Tweets	Predict Targets based on own Tweets
[Watanabe et al., 2018a]	✓			
[Alfina et al., 2017]	✓			
[Gaydhani et al., 2018]	✓			
[Badjatiya et al., 2017]	✓			
[Davidson et al., 2017]	✓			
[Silva et al., 2016]			✓	
[Zampieri et al., 2019b]			✓	
[Zhang et al., 2019]			✓	
[ElSherief et al., 2018c]			✓	
[Burnap and Williams, 2016]	✓			
[Zhang et al., 2018]	✓			
[Ribeiro et al., 2018]		✓		
[Ribeiro et al., 2017]		✓		
[Zhang and Luo, 2018b]	✓			
[Sharma et al., 2018c]	✓			
[Aulia and Budi, 2019]	✓			
[Zhang et al., 2019]	✓			
[Zampieri et al., 2019a]	✓			
[Şahi et al., 2018]	✓			
[Sharma et al., 2018b]	✓			
[Warner and Hirschberg, 2012]	✓			
[Rodríguez et al., 2019]	✓			
[Plaza-del Arco et al., 2019a]	✓			
[Waseem and Hovy, 2016]	✓			
[Pratiwi et al., 2019]	✓			
[Santosh and Aravind, 2019]	✓			
[Bohra et al., 2018]	✓			
[Nugroho et al., 2019]	✓			
Proposed Framework				✓

Table 3.5: Categories of problems related to hate speech

Problem	**Purpose of Studies**	**Examples**	**Detection, Analyzing, Prediction**
P1	Detect hateful tweet	@user2 replying to "@usr1 **shut up idiot**"	Hate speech detection
P2	Detect and categorize hateful user	**@user2** replying to "@usr1 shut up idiot"	On the basis of profile and tweet content. @user2 is a hater
P3	Identify and analyze targets of hate from hater's tweets	@user2 replying to "**@usr1** shut up idiot"	Analyzing target @user1 is a target
P4	Predicting the targets of hate speech by victims tweets	**"@user1 needs to learn how to spell thread"**	Predicting whether on the base of this tweet @user1 becomes the target of hate speech or not

tion of hate speech. P2 and P3 are rarely addressed. Finally, we have summarized the literature in the Table 3.6.

Table 3.6: Summary of hate speech related work in existing studies

Reference	Problem	Algorithms	Features	Evaluation
[Alfina et al., 2017]	P1	NB, SVM, RF	Ngrams	W–a, F1
[Gaydhani et al., 2018]	P1	NB, LR, NB	TF–IDF, Ngrams	P,R, F1, A
[Badjatiya et al., 2017]	P1	LSTM, CNN	TF–IDF, BoW	P, R, W–m, F1
[Davidson et al., 2017]	P1	LR, NB, RF, SVM	TF–IDF, No of Hashtag, Retweet, mention, URLs, Sentiment lexicon	P, R, F1
[ElSherief et al., 2018b]	P3	SAGE	Named Entity Recognition	Linguistic Analysis
[Silva et al., 2016]	P3	Manually analyzed	–	P
[Burnap and Williams, 2016]	P1	SVM, RF, DT	BOW, Ngrams	P, R, F1
[Zhang et al., 2018]	P1	CNN+GRU	–	P, R, F1
[Ribeiro et al., 2018]	P2	AB, GB, GS	user, glove	AUC, A, F1
[Ribeiro et al., 2017]	P2	AB, GB	User features	
[Zhang and Luo, 2018a]	P1	CNN+GRU, CNN+sCNN	Word2Vec	F1
[Şahi et al., 2018]	P1	NB, SVM, RT, J48, RF	Ngrams, TF–IDF, No of char	P, R, F1
[Kwok and Wang, 2013]	P1	NB	Ngrams	A
[Djuric et al., 2015]	P1	LR	paragraph2vec	AUC
[Nobata et al., 2016]	P1	SkipGram–Model	Punctuation, Ngrams, PoS, length	P, R, F1
[Aulia and Budi, 2019]	P1	SVM, LR, RF, DT	TF–IDF, Ngrams, Sentiment	W–a, F1
[Pratiwi et al., 2019]	P1	SVM, LR, NB, RF, DT	Word n-grams, Char n-grams, Hate Code	F1
[Santosh and Aravind, 2019]	P1	SVM, RF, LSTM	sub-word	A,R, F1
[Bohra et al., 2018]	P1	SVM, RF	Ngrams, Lexicon Punctuation, Negations	A
[Nugroho et al., 2019]	P1	RF, AB, Neural Network	–	A, P, R, F1

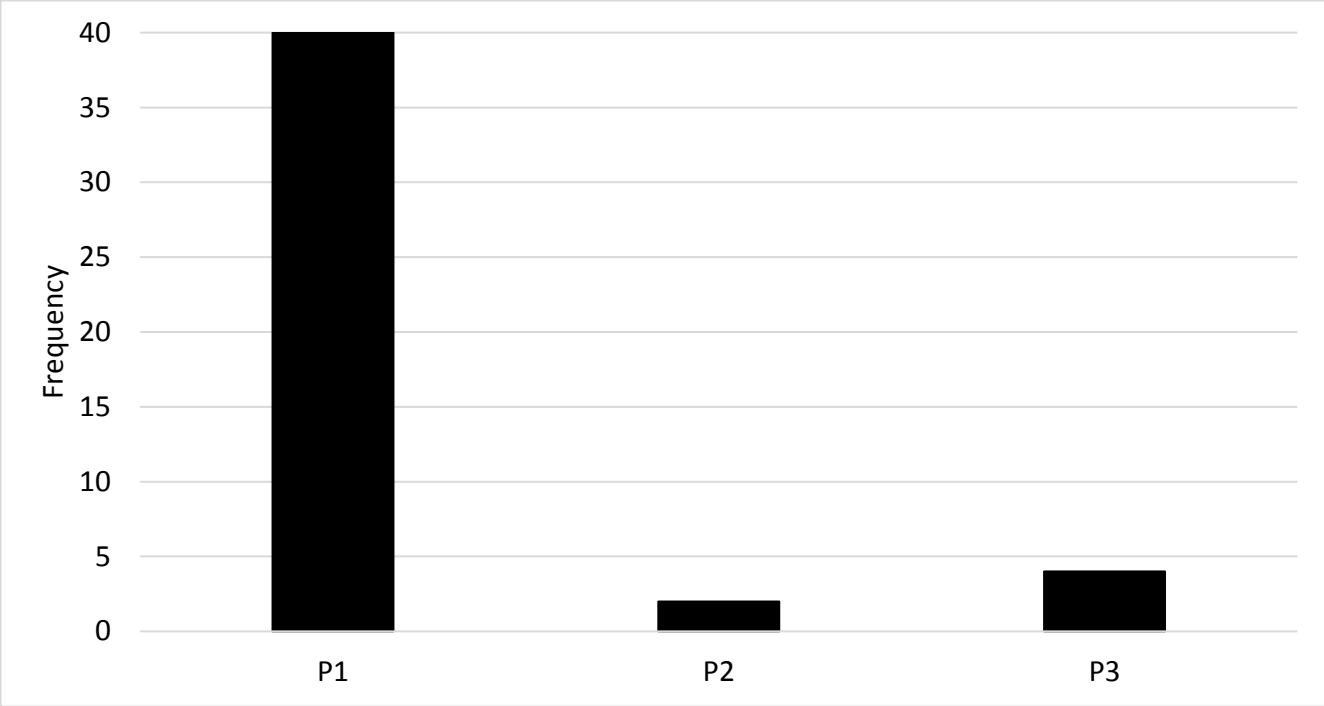

Figure 3.10: Frequency of problems addressed in literature

3.4 Summary

The purpose of this chapter is to explore issues related to hate speech. For this purpose, we have conducted systematic literature review in order to analyze the literature regarding hate speech on Twitter. For this purpose, we have used the methodology described by Kitchenham et al. [2010]. First, we have extracted hate speech documents from digital libraries. Then we have analyzed the approaches, social media platforms, languages, datasets, preprocessing techniques, features extraction techniques, classification methods and evaluation measures used to address hate speech issues on Twitter and social media in these papers.

The issues that we have found are language diversity, unavailability of datasets and lack of common benchmark. Due to which there is a lack of comparative studies. For this reason it is difficult to decide which approach is the best in terms of performance. Furthermore, we have analyzed the work on hate speech issues in literature and categorize it into research problems **P1**: detect hateful tweet, **P2**: detecting and characterize

hateful users and **P3**: Identify and analyze targets of hate speech from hater tweets. As a result, we have found the research gap that is to predict the targets of hate speech on Twitter based on victims tweets this is a unique problem among the the problems of the rest of the studies and has not been worked out to the best of our knowledge.

Chapter 4

Dataset and Proposed Framework (HTPK)

The purpose of this chapter is to describe the dataset and framework. In Section 4.1 we have clearly explained the collection and annotation of the dataset on Twitter. It should be clear that the dataset we have collected is a dataset of hate speech targeting. Furthermore, we have outlined our proposed Hate–speech Target Prediction FrameworK (HTPK) in Section 4.2. HTPK consists of a five steps which includes dataset preprocessing, features extraction, features concatenation, parameter tuning of machine learning algorithms, validation and to predict target and non–target of hate speech.

4.1 Dataset

RQ1: Can we develop a labeled dataset of hate speech targets?

There is no doubt that a good contribution of work is being done to detect hate speech. For which different datasets have been used [Waseem and Hovy, 2016, Sharma et al., 2018a, Tulkens et al., 2016, Ross et al., 2016] but there are still some issues that we have not seen resolved in the literature, one of is to predict the target of hate speech. To address this issue, we analyzed different datasets, but we could not find any dataset that

provides the posts of the hate speech victims, on the contrary, existing datasets refer to the haters' posts. Therefore, we didn't use existing datasets in our study. We have developed a new Hate Speech Targets Dataset (HSTD) to resolve target prediction problem. The statistics of HSTD dataset are shown in Table 4.1. We have described the HSTD collection and annotation in detail in the section 4.1.1.

Table 4.1: Statistics of the dataset

Users		**Tweets**	
Targets	1500	Results in hate	1500
Non-Targets	1500	Did not results in hate	1500

4.1.1 Dataset Collection and Annotation

Dataset collection has been a bit difficult task for us. Because hate speech victims data is the key to our research work, while, the data have been used in the previous studies is about hater's post. After analyzing existing datasets mentioned in literature, the dataset has used in the [ElSherief et al., 2018b] paper is closest related to our research problem. This dataset has been about haters' posts. It analyzes the target of hate speech through the hater's posts. For example: "@name **nigga** get out from my country". This tweet has a nigga target, so this is how target was analyzed. There are two types of tweets in their dataset, generalized hate and direct hate.

- **Generalized hate:** It is the language of hate by which an individual group is targeted for hate speech. e.g religion, ethnicity, nationality etc.
- **Directed hate:** It targets specific people or entity with hate speech. An instance is: @user you are *idi*t.

They have used two methods to collect these tweets from Twitter API. Which are key-phrase-based and hashtag-based. We took direct tweets in which the hater was replying to a victims. Then we tracked the hater tweets id and collected the tweets that have been

replied. We have collected 6000 tweets in total using the Twitter API. We could not collect most of these tweets some users has expired accounts and we did not have access to some users' accounts because of their privacy policy. We collect two tweets of the same user. There was one in which the user was becoming the target of hate speech. At the same time frame, we have collected second tweet in which the user was not becoming the target of hate speech. We have not taken tweets from verified users' accounts in our dataset because verified users are often become the victim of hate speech based on their personality.We did this step based on analysis, before working on this problem, the problem we were working on was the hate speech prediction for verified users.We took tweets from verified users and analyzed replies to these tweets. What we noticed after the analysis was that most verified users are becoming victims of hate speech in every tweet they make. The content of their tweets is not causing hate speech for them but they are becoming victims of hate speech because of their public personality. Therefore, in this study, we did not add the tweets of verified users to our dataset because in this problem, our focus has been on the content of the tweet. That is why we have collected the posts of common users. common users are users without verified blue checkmarks on their accounts. That is how we developed a 3000 tweets HSTD. The procedure of dataset collection is shown in Figure 4.1.

We have manually labeled the HSTD. In order to measure the non-target we collected tweets that are not replied. Witch this technique we have analyzed that thy are not becoming the victims of hate speech. Then we labeled all these users' tweets as non-target. We have used two labels in our dataset. Users who have been the target of hate speech labeled as "Targets" and tweets in which they are not targeted of hate speech are labeled as "Non-Targets". We have analyzed that a user does not have the following things in tweet is less likely to be a target of hate speech.

- Use of demeaning words
- Encourage hate speech

- Promotes tricky hashtags such as #BoycottStarbucks, #MuslimBan.
- Show misbehavior in post
- Attack on someone personality
- Criticizing on celebrity personality
- Defends racism

Unbalanced dataset may cause negative outcome on prediction performance [Ganganwar, 2012]. Therefore, we have developed the balance HSTD. 1500 tweets are labeled as 'Target' and remaining 1500 are labeled as 'Non-Targets'. The 'Target' tweets in it are the tweets that we have tracked with hater tweets. They are labeled as targets because users have become victims of hate speech in response to these tweets. Furthermore, ambiguous statements which become hate speech in a particular context are not covered, e.g., "I'll tell you if you do this". Similarly irony, taunts are not covered. Not all of them are covered because the tweets of the target users that we have taken are tracked from the hater's post. And in non-target, take the tweets of these target users in the same time period which are without reply. Therefore, we have not been able to cover all this in our study. Our final HSTD contains 3000 tweets.

4.2 Proposed Frame work

Predicting the target of hate speech on social media is a challenging task. Every human being has different ways to express his or her thoughts, it also includes hate speech. Thus, it is difficult to write the rules of hate speech target prediction by hand. Therefore, we proposed a Hate–speech Target Prediction FrameworK (HTPK) . Its basic goal is to predict the target of hate speech on Twitter. Essentially it involves five phases. Which are shown in Figure 4.2. These phases are discussed in detail in the following sections.

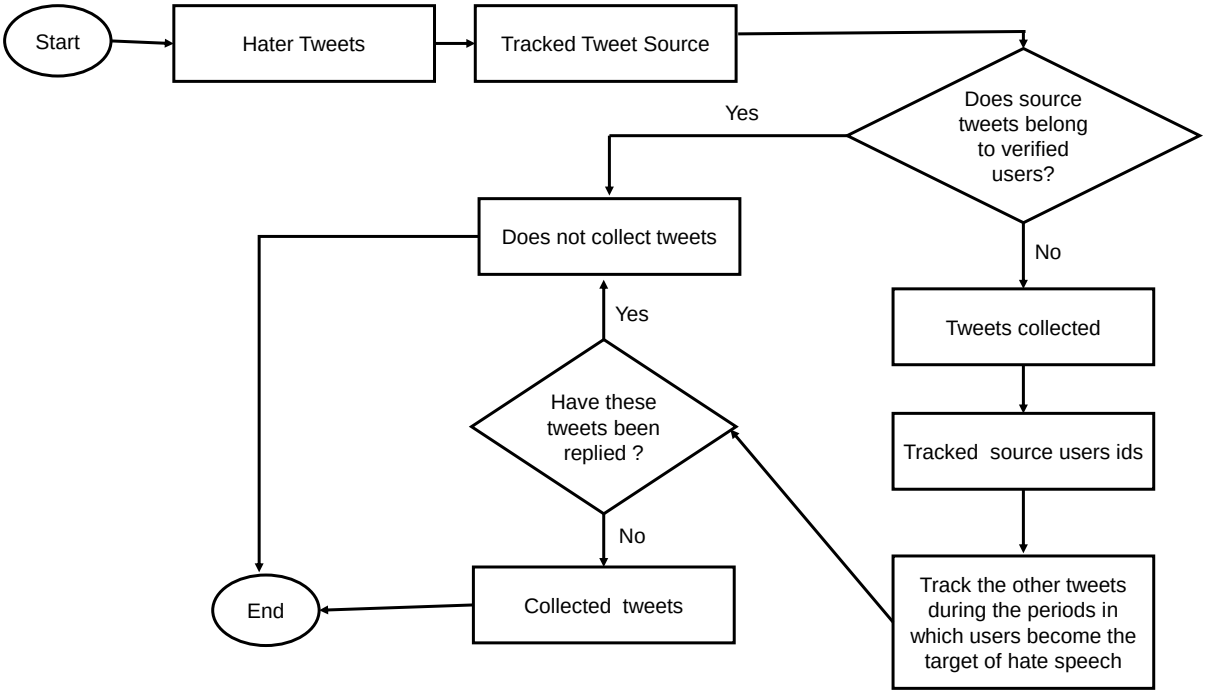

Figure 4.1: Dataset collection procedure

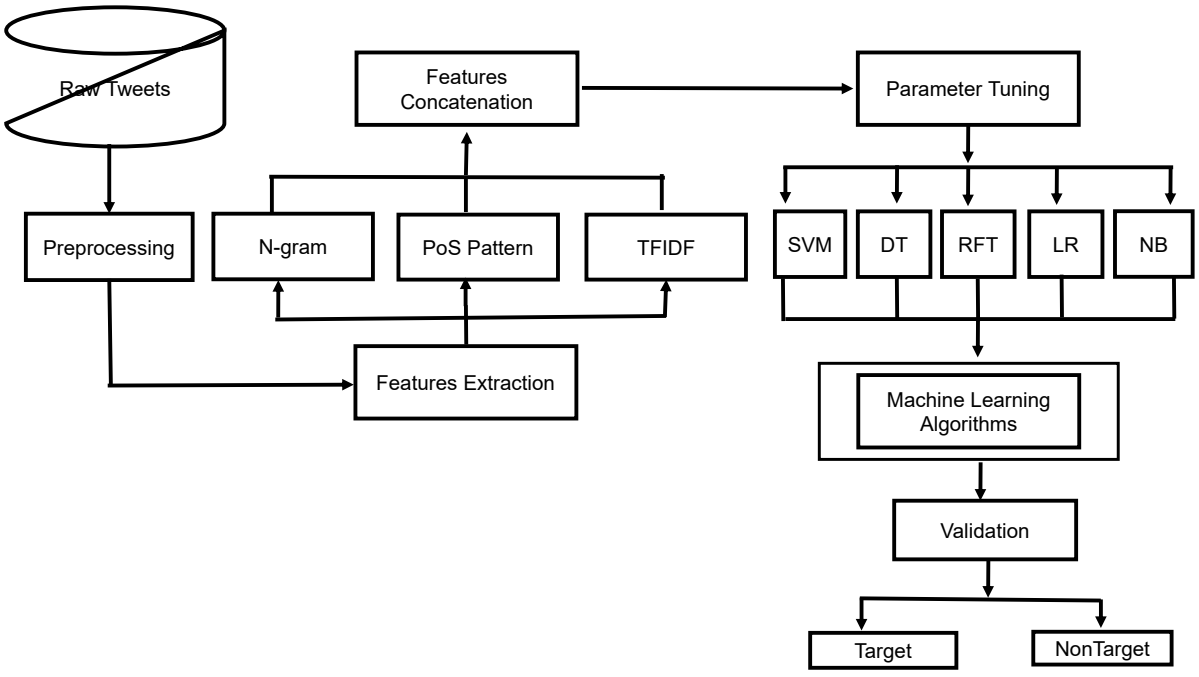

Figure 4.2: Hate Speech Target Prediction Framework (HTPK)

4.3 Data Pre-Processing

Data pre–processing is a process which process input data into a from which is useful for ML algorithms. The data preprocessing techniques play an important role in improving the performance of any ML algorithm. That's why it is necessary to have clean data for ML algorithms. There is too much language diversity on Twitter in some languages, Noise is a major issue on Twitter due to the informality, containing useless characters, short sentences, abbreviations and slang. It is important to preprocess data to reduce this noise. We have applied the following preprocessing techniques on our dataset.

- **Lowercasing:** In is a process in which a stream of text is converted into lowercase. This technique helps to improve the performance of classification as it makes the data easier to understand and it decreases the data dimensionality. Not using this technique can cause problem such as “HATE”, "HaTe" and "hate" being considered various words.

- **Irrelevant characters removal:** Not all of irrelevant characters (e.g. !?&%) are helpful in classification task, so they have been removed from our dataset.

- **Tokenization:** Tokenization means splitting the stream of text into meaningful tokens. Tokenization is different in every language. This is a common technique for preprocessing.

- **Stemming:** In stemming, words are reduced to their base form such as "studies" to "studi", "likes" to "like". Using this technique improves the classification of the text. As it reduces the data dimensionality.

- **Stop-words removal:** These words are words that do not contains useful information (e.g: "and", "a" , "this"). Therefore, these words are not considered good for text classification. They have been removed.

- **Emoticon removal:** Our dataset has a low number of emojis, so we removed them instead of translating them. Because of this, the accuracy of the model did not matter.

- **Punctuation removal:** Generally, punctuation is not considered important for text classification. So, it has been removed.

- **Remove URLs, Numbers, Double Spacing and Emails:** Raw tweets sometimes contain elements that are not effective for prediction problem, as emails, Numbers (e.g., 0, 1, 2, ..., 9) and URLs. Therefore, we have removed them from our dataset.

4.4 Features Extraction

In the related work, we have analyzed that most studies have used features extraction to improve the performance of classifier [Alfina et al., 2017, Gaydhani et al., 2018]. Therefore, after preprocessing the dataset, we have extracted the three features from tweets and then combined these features, so that these could be fed to ML algorithms by making meaningful representation of tweets. The following is a description of these features.

4.4.1 N-grams based Features

We trained machine learning algorithms, so that they can process natural language. Human can easily understand natural language, but machine learning algorithms cannot understand this language because they just take numeric inputs. Therefore, we need to create a language pattern that machine learning algorithms can easily understand. As each word has its own specific meaning but when we merge these words, it becomes even easier to understand their meaning. The words taken by BoW (Bag of Words) rely on the use of N-grams (N-grams is a sequence of words that is derived from text). Words

Table 4.2: Data pre-processing procedure for hate speech target prediction

Preprocessing Techniques	**Before**	**After**
Lowercase	The OFFICIAL videos o͡o32 are here, free@gmail.com, https://t.co :)	the official video o͡o32 is here, free@gmail.com, https://t.co :)
Remove punctuation	the official videos o͡o32 are here, free@gmail.com, https://t.co :)	the official videos o͡o32 are here free@gmail.com https://t.co :)
Remove special characters	the official videos o͡o32 are here free@gmail.com https://t.co :)	the official videos 32 are here free@gmail.com https://t.co :)
Remove numerals	the official videos 32 are here free@gmail.com https://t.co :)	the official videos are here free@gmail.com https://t.co :)
Remove stop-words	the official videos are here free@gmail.com https://t.co :)	official videos free@gmail.com https://t.co :)
Remove emoticons	official videos free@gmail.com https://t.co :)	official videos free@gmail.com https://t.co
Remove URLs	official videos free@gmail.com https://t.co	official videos free@gmail.com
Merge multiple spaces	official videos free@gmail.com	official videos free@gmail.com
Remove emails	official videos free@gmail.com	official videos
Tokenization	official videos	official|videos
Stemming	official|videos	official|video

N-grams study as tokens, set of word or words however, character N-grams study as set of character or characters. By using N-grams feature we can anticipate word t_n based on preceding n-1 words.

$$p(t_n|t_1, t_2, t_3, ..., t_{i-1}) \tag{4.1}$$

We have taken unigram and bigram and trigram in this study. For Example:["'His', 'replies', 'insult', 'target' ' war'" is a **Unigram**] and ["'His replies' , 'replies insult', 'insult target' , target war'" is a **Bigram**]. We have further described them in Figure 4.3 with an example. The reason for using unigram, bigram and trigram is that the size of HSTD is

sma

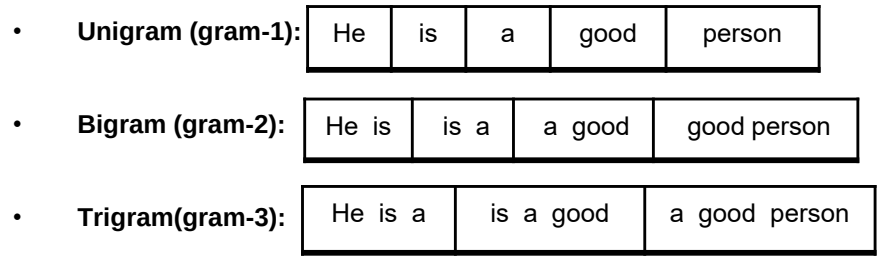

Figure 4.3: Representation of N-gram

4.4.2 TFIDF based Features

The TFIDF (term frequency-inverse document frequency) is the frequency of each word as per its inverse frequency in the dataset. This means that the weighted frequency value of the low occurrence tokens is higher than that of those with high occurrence. On the other hand, Term-Frequency (TF) has the absolute frequency of the terms in the dataset. TFIDF is the base feature used for experiments in the state-of-the-art [Gaydhani et al., 2018]. The TFIDF value for term t in dataset d can be calculated using (4.2) [Gaydhani et al., 2018].

$$tfidf(t,d) = tf(d,t) * idf(d,t) \tag{4.2}$$

To have the feature values in a range, TFIDF has to be normalized. L1 normalization is defined in (4.3) and L2 in (4.4).

$$L1_{norm} = \frac{u}{|u_1| + |u_2| + |u_3| + ... + |u_d|} \tag{4.3}$$

$$L2_{norm} = \frac{u}{\sqrt{u_1^2 + u_2^2 + u_3^2 + ... + u_d^2}} \tag{4.4}$$

4.4.3 Part Of Speech Tags for Pattern Extraction

We have used PoS (Part of Speech) tagging to analyzing the relationships among words and text content. Using PoS, we replace the original words in the tweets with their grammatical position in the sentence. In the parts of speech tags a parse tree is constructed, which is used to extract the relation among words based on their definition and context. In addition, PoS tagging is a process in which words in tweets are marked by their corresponding grammatical tags. More formally, allow Words w be includes in dataset with tags $t_a, ..., t_n$. We may assign tags:

$$A(t_g|w) = \frac{d(w, t_g)}{d(w, t_a) + ... + d(w, t_n)} \tag{4.5}$$

Where d(w,t_g) is frequency of w and t_g occur in the dataset. Thus, A($t_g|w$) is the probability context to tag the word ambiguous. For instance, the word insult can be used for both noun and verb. For instance, the following sentence:

- His replies are an **insult** to the target of war. **(noun)**
- Don't **insult** anyone. **(verb)**

To see the difference in the use of these terms, on has to look at both context and definition of the word. For this we used peen treebank PoS tagset. The words of the tweets in the dataset are assigned their corresponding tags. For this, a column named PoS tag vectorization is created in the dataset. Comparative terms are described in Table 4.3.

We have extracted the pattern of each tweet from target and non-target class using the part of speech tags vectorization. For instance, the following tweet "@Name you are idiot, who told you they wanted to be friends with you" ["NN-@Name PRP-you VBP-are JJ-idiot, WP-who VBD-told PRP-you PRP-they VBD-wanted TO-to VB-be NNS-friends IN-with PRP-you"]. It captures syntactical content. The frequency of the tags in the tweet is used as a feature.

Table 4.3: POS tags used and their descriptions

PoS Tags	**Description**
CD	CARDINALNUMBER
NNS, NN, NNPS, NNP	NOUN
PRP$, PRP	PRONOUN
VB,VBG,VBD,VBP,VBN,VBZ	VERB
DT	DETERMINER
MD	MODAL
RP	PARTICLE
UH	INTERJECTION
JJ,JJS,JJR	ADJECTIVE
FW	FOREIGNWORD
LS	LISTMARKER
TO	TO
CC	COORDCONJUNCTION
EX	EXISTENTIAL
RB,RBS,RBR	ADVERB
PDT	PREDETERMINER
POS	POSSESSIVEEND

4.5 Summary

In this chapter dataset collection, annotation process and the HTPK framework are defined. Furthermore, the preprocessing techniques applied to the HSTD are explained. In addition, features extraction techniques used in this study are explained in detail. Afterwards, experiments have been performed which are described in detail in Chapter 5.

Chapter 5

Experiments and Evaluation

The purpose of this chapter is to a evaluate the proposed framework. Section 5.1 describes the approaches used in state of the art and apply them to HSTD dataset. In Section 5.2 we tune the parameters of the algorithms used in proposed framework. Section 5.3 clearly discusses the results obtained from the propose framework. Finally, Section 5.4 provides reasoning on why people become victims of hate speech on Twitter.

5.1 State of the Art

We have taken two published papers [Alfina et al., 2017] and [Gaydhani et al., 2018] as the state-of-the-art. We have applied the techniques used in these papers to HSTD. We found these papers as effective state-of-the-art methods, as they provide us improvements by using machine learning algorithms. Furthermore, Gaydhani et al. [2018] reports accuracy in their results and the [Alfina et al., 2017] reports the weighted average of F-measure. The performance of the models is shown in Table 5.1 by applying the techniques used in the state-of-the-art to HSTD. These results are different form the base papers because we used HSTD. Therefore, we do not assure fully consistency in the methods used in experiments. In order to assure the viability degree of performance, we can make sure that the preprocessing techniques and features used are similar for

each comparison. The results obtained by using Alfina et al.'s approach show that we are getting fine F-measure on NB (Naive Bayes). NB is performing well on unigram feature. While the rest of the classifiers also perform well. We have observed that as we increase the N-gram performance of classifiers decreases that could be due to tweet's short length, lack of structure and informality, together with the existence of diminutives and typos. Therefore, it is difficult to find a set of tokens that occurring together. Consequently, Increasing N-gram reduces performance. All the classifiers are performing well on L1 regulation of TF-IDF. However, LR (logistic regulation) performs better on unigram and bigram (1,2) with TFIDF L1 than NB and SVM.

Table 5.1: Results of the state-of-the-art methods

[Alfina et al., 2017]			
Features	**SVM**	**RF**	**NB**
Word unigram	86.42	82.50	**88.78**
Word bigram	72.42	68.78	73.14
Char trigram	58.72	61.45	58.72
[Gaydhani et al., 2018]			
N-gram+TFIDF Norm	**LR**	**SVM**	**NB**
unigram + L1	89.10	90.42	88.44
(unigram and bigram) + L1	**90.75**	89.10	89.43
(unigram and trigram) + L1	88.11	87.45	89.10
(unigram) + L2	89.10	88.77	87.45
(unigram and bigram) + L2	89.76	89.76	89.10
(unigram and trigram) + L2	88.44	89.10	88.44

Features used in the state-of-the-art and proposed framework are summarized in Table 5.2.

Table 5.2: Feature extraction used in the state-of-the-art methods

Features used in state-of-the-art methods and the proposed framework			
Article	**N-gram**	**TF-IDF**	**PoS Patterns**
[Alfina et al., 2017]	✓		
[Gaydhani et al., 2018]	✓	✓	
Proposed Approach	✓	✓	✓

5.2 Hyper-Parameters Tuning

The Model optimizer and hyper-parameter values utilized in the training procedure are various for respective learning purpose. These should be used for optimization to enhance model efficiency in terms of fit, accuracy and generalizability to invisible data. To improve each hyper-parameter we can either make sense about reasonable values for this type of parameters on the basis of the learning purpose or complete grid-search on the model. Gird-search is a progressive process to optimize the hyper-parameters. In general, if the input of the learning task is high, they are computationally expensive. We only have text data in this study, so we have performed a complete gird-search to the models. Accordingly, for Naive Bayes We have examined the smoothing prior α for tuning. For Logistic Regularization we have used the C regularization parameter and the algorithms for optimization (solvers) saga ,liblinear and saga for performance tuning. Furthermore, The N estimator parameter for Random Forest, C parameter for Support Vector Machine and For Decision Tree criterion parameter are considered. Different values have been used in all these parameters and gird search has been used to select the best value of the parameter. And the parameter on which maximum accuracy was being gained has been considered in the final results. Figure 5.2 shows the Naive Bayes results after tuning. We have examined the smoothing parameter α for tuning with $\alpha > 0$ because it prevents zero probabilities. Naive Bayes performs poor for α value 0.01 giving 89.10% accuracy and performs better on α value 1 giving 93.06% accuracy. Figure 5.1 shows the Logistic Regression performance after tuning. We have considered C parameter for regulation and the algorithm use for optimization are (solvers) lbfgs, saga and liblinear for performance tuning. As it is clear from the Figure 5.1 that the accuracy of all the solvers at the value of 3 and 6 of C regularization are the same 92.40% and as the values of C Regularization are increasing the accuracy is decreasing. Here, we conclude that the values of C effect on accuracy because the change of solver does not matter.

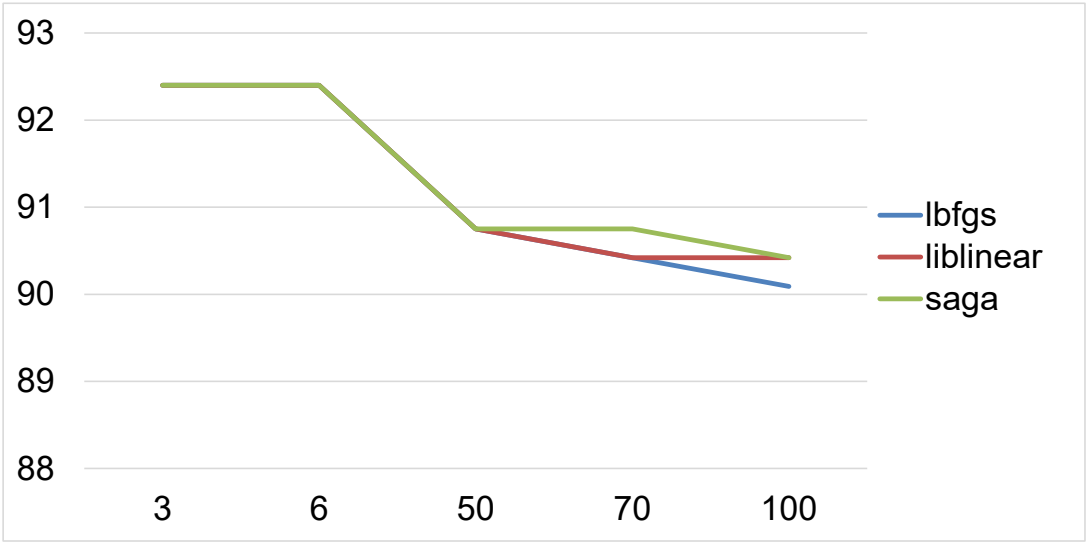

Figure 5.1: Results of Logistic Regression after tuning for N-grams for N=[1–2] and POS tags

5.3 Results and Discussion

After features extraction and parameter tuning, we move to our final experiments. The experiments are performed using the scikit learn. Scikit learn presents variation of classifiers according to the group of the algorithm (e.g: rule-based, decision tree-based, etc. We consider five well known machine learning algorithms used for prediction: Logistic regression, Support Vector Machines, Naive bays, Random Forest and Decision Tree. We performed 10 k-fold validation for the training of each model on HSTD. We used various key measures that include accuracy, recall, precision and F1 scores determined as: (where TN = True Negatives, TP = True Positives, FN = False Negatives and FP = False Positives) to evaluate prediction performance. These measures are defined in (5.1–5.4).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5.1}$$

$$Recall = \frac{TP}{FN + TP} \tag{5.2}$$

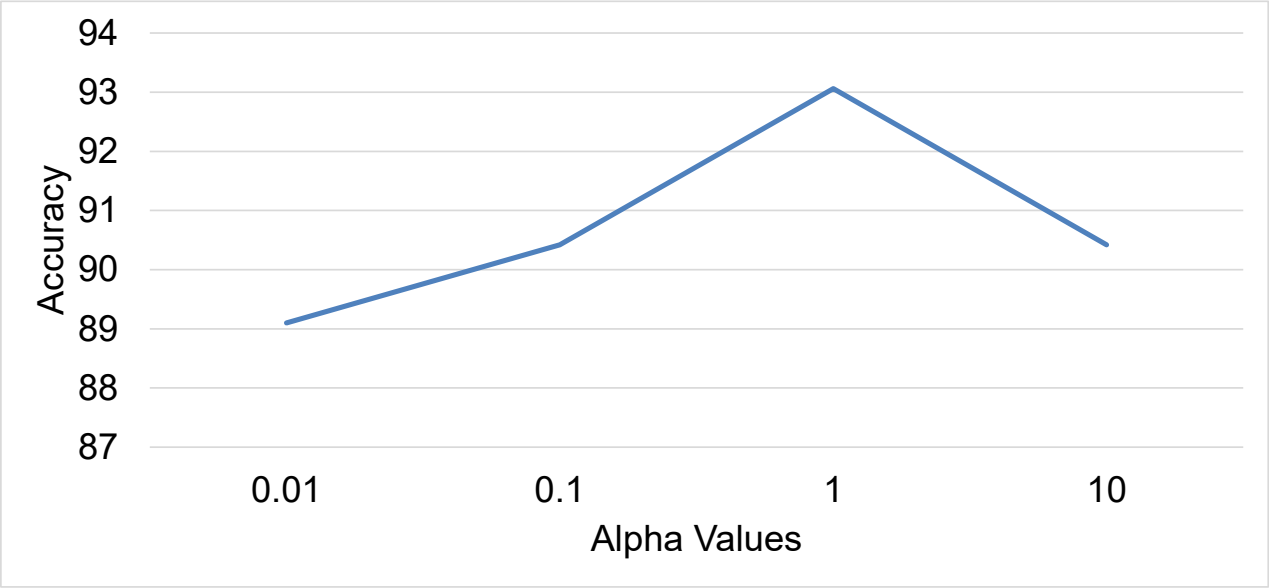

Figure 5.2: Results of Naive-Bayesian after tuning for N-grams up for N=[1–2] and POS tag

$$Precision = \frac{TP}{FP + TP} \tag{5.3}$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{5.4}$$

Accuracy measure has been used to balance between recall and precision, while recall is used to find the false negative cost and precision is used to determine the cost of false positives. Furthermore, the performance of each model is examined based on the combination of various features parameters. The performance of these models is compared. Table 5.3 shows that all the five algorithms perform remarkably better with combination of various features set. We have seen that word unigram (1, 1) with TFIDF and PoS show best results for low values of N, probably because of tweet's short length, lack of structure and informality, together with the existence of diminutives and typos. Therefore, it is difficult to find a set of tokens that occurring together. Furthermore, we observed that the efficiency of HTPK starts to decrease if we increase the range of character trigram where (n=1–3). On the contrary, word N-grams with PoS tags and TFIDF

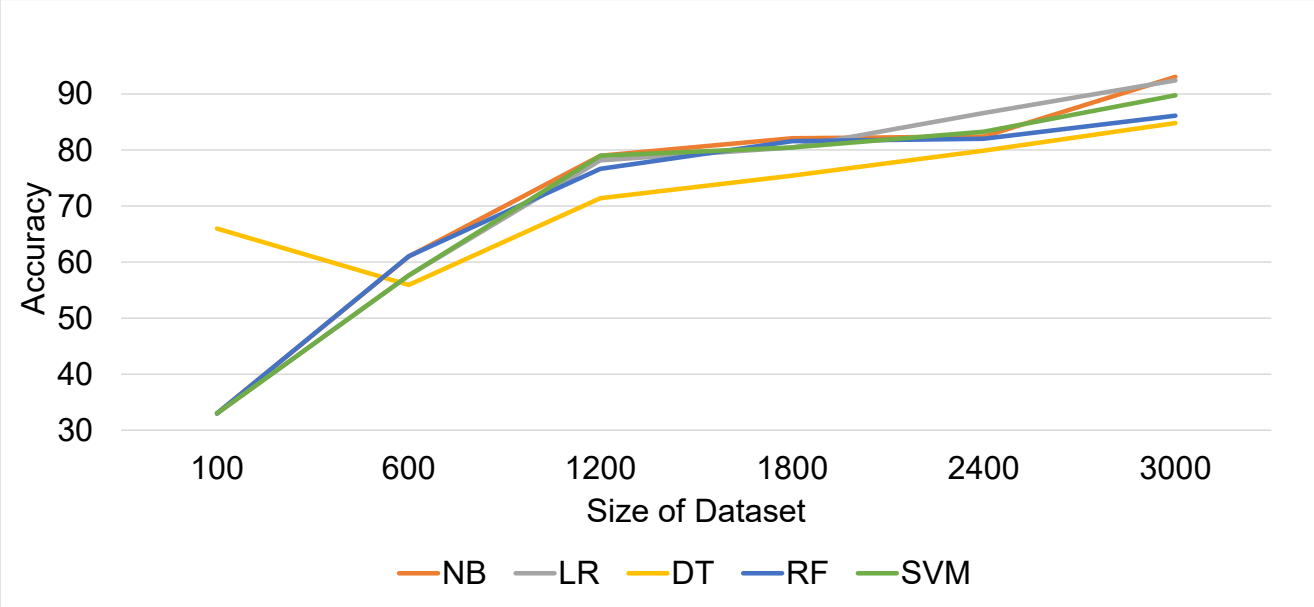

Figure 5.3: Accuracy Of HTPK on Different Size of Dataset

obtained better performance for lower values of N. Results in bold refer the best result of experiments. Target prediction improve by the features combination, with an improvements of up to 3% and 5% as compared to state-of-the-art methods [Alfina et al., 2017], [Gaydhani et al., 2018].

Table 5.3: Results for various combinations of features

Accuracy Metrics					
Features	**Linear SVM%**	**DT%**	**RF%**	**LR%**	**NB%**
word unigram + TFIDF + PoS Tags	84.76	85.47	87.12	89.76	91.08
word bigram + TFIDF+ PoS Tags	89.76	84.81	86.13	92.42	**93.06**
char trigram + TFIDF + PoS Tags	89.43	81.84	86.46	88.77	90.42

RQ2: Which algorithms perform better for hate speech target prediction on Twitter?
Almost all classifiers are giving good results on the combination of TF-IDF, PoS and bigram feature so using this feature we have calculated recall, precision and F-measure of all classifiers. Which are described in Table 5.4. As we can see in Table NB is performing the best in terms of accuracy. This is because NB gives better performance across multiple features as compared to the rest of the classifiers. And we have used a combination of different feature on HSTD. We combine these features using (5.5).

$$F_u = \{f_1, f_2, f_3\} \tag{5.5}$$

Where f_1, f_2, and f_3 are defined in (5.6), (5.7), and (5.8), respectively.

$$f_1 = p(t_n|t_1, t_2, t_3, ..., t_{i-1}) \tag{5.6}$$

$$f_2 = tfidf(t,d) = tf(d,t) * idf(d,t) \tag{5.7}$$

$$f_3 = A(t_g|w) = \frac{d(w,t_g)}{d(w,t_a) + ... + d(w,t_n)} \tag{5.8}$$

RQ3: Can part of speech tags be useful in hate speech target prediction?

In addition, tuning the parameter of NB also improves accuracy. If viewed in the terms of precision NB performance is lower than LR. It is examined that the recall for non-target tweets is comparatively low 0.92. This means that 8% of actually non-targets tweets were misclassified by the NB. In addition, the target class precision is 0.92 which means that there were 8% tweets that were originally non-target have been classified as target. On the contrary, the precision for non-target class and the recall for target class is 0.93, which is considerably better.

If we view at LR performance, its performance in terms of accuracy is less than NB but its performance is better than other classifiers. After completing experiments, we analyze that Decision tree performs poorly as compared to Logistic Regulation, Naive Bayes, Random Forest Tree and SVM. NB performing better than LR because NB has a lower variance but higher bias compared to LR. Because our dataset has bias due to the vocabulary used in Target and Non Target class's tweets therefore the performance of NB

is better than LR. The cause lies in one that, all the features of a large tree size need to be included. Due to the size of the tree, the classifier requires to cross multiple nodes till it arrives on the leaf node and anticipate the target and non-target classes. That is a reason that, it increases the likelihood of errors due to long path as a result classifier accuracy reduces. As depicted in the Figure 5.3, as the size of the dataset is increasing the performance of DT and RF is decreasing. We have used different size of dataset on HTPK As that can be seen in the Figure 5.3. The purpose of using different size is to determine the effectiveness of various data size for ML algorithms. Consequently, increasing the dataset size improves the performance of HTPK.

Table 5.4: Precision, Recall, F1-Measure and Accuracy of prediction using different classifiers

Classifiers	**Target Class**			**Non-Target Class**			**Accuracy**
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**	
NB	0.92	0.93	0.93	0.93	0.92	0.92	**93.06**
DT	0.83	0.83	0.84	0.83	0.83	0.83	84.81
RF	0.94	0.79	0.86	0.81	0.95	0.87	86.13
SVM	0.91	0.87	0.89	0.87	0.91	0.89	89.76
LR	0.95	0.89	0.92	0.89	0.95	0.92	92.42

5.4 Characterizing Target and Non-Target Posts

RQ4: Can we predict the target of hate speech by their tweets content?

The basic purpose of this research is to examine that why people are becoming the target of hate speech in social media. For this purpose, we have analyzed the common users' tweet-contexts and observed that proper selection is very important while posting any contextual thing on online social media. We find that the post contains more negative words[1], promote hate speech, attack on someone personality, criticizing on celebrity personality, defend racism, contain anger and misrepresent truth due to which user be-

[1] `https://web.archive.org/web/20190718204432/https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words`, last accessed on Jul 18, 2019

comes the victim of hate speech. In Figure 5.4 we depict word-cloud containing the

Figure 5.4: Most frequent words in target posts

target posts top words. We have deliberately hidden f***, b**** and other such words in Figure5.4 because these words fall into the category of abusive words. On the other hand, the non-target posts contain more positive words [2], show respect, promote love and contains less words relevant to anger, shame and insult. Figure 5.5 depict the visualization of word cloud related to non-target posts top words. The frequency of each word in word-cloud depicts how often it occurs in the HSTD. We analyze that the general words used in both classes are just same that user use when appreciating or insulting someone. Furthermore, we have specified target and non target user with an example. Which is as follows:

- @user1 “I want to scream“ **Non-Target**
- @user1 “you die soon” **Target**

[2]https://gist.github.com/mkulakowski2/4289437

These are two tweets from one user. In one, user is becoming the victim of hate speech in reply of the tweet. And in the other one user is not becoming the victim of hate speech. The tweet in which user is the victim of hate is referring to the second person in bad words. In another user is expressing himself.

RQ5: Why common users on social media are becoming the victims of hate speech?

Hence, we conclude that the most common users on social media are becoming the victims of hate speech because of the content used in their posts. The words user use in their posts on social media reflect their personalities. Therefore, choosing the right words has the great importance.

Figure 5.5: Most frequent words in non-target posts

5.5 Summary

The purpose of this chapter is to validate the HTPK Framework proposed to predict the targets of hate speech on social media, specifically on Twitter. For this purpose, at the

beginning of the chapter the methods of state–of–the–art have applied to Hate Speech Targets Dataset (HSTD) so that it can be compared to HTPK framework. Afterwords, the HSTD is applied to HTPK and it is examined that the results obtained from HTPK are better than state–of–the–art methods. Finally, the interesting facts have discussed that common user on Twitter are often become victims of hate speech because of the content of their posts.

Chapter 6

Conclusion and Future Work

In this study, we highlight an important aspect of predicting hate speech targets on social media using machine learning. To achieve this, we have developed a new Hate Speech Target Dataset (HSTD) because we could not find a similar existing dataset. We propose a novel framework Hate–speech Targets Prediction Framework (HTPK) for predicting targets of hate speech. A diverse set of experiments is performed to measure the effectiveness of HTPK. HSTD is the first dataset that is about target posts. We performed comparative study of Naive Bayes, Decision Tree, Logistic Regression, Support Vector Machines and Random Forest Classifier on several set of features and hyper-parameters of model. We use various features to improve the performance of these models. Using PoS tags, TFIDF and N-grams separately did not produce fine results. Therefore, we combine all these features which resulted in improved results. Couples with, Among different machine learning algorithms, Naive Bayesian produces the best results along with the combination of PoS tags, TFIDF and N-grams. Our analysis provides a number of unexpected and interesting findings about targets of hate speech. In general, we have found that most common users on social media are targeted because of their own post contents. As such, their posts defend against anger, misbehavior, criticizing, racism and misrepresenting truth. Not only that, but the common users become less of a target of

hate speech based on their personality. And we have also analyzed that if the target's posts does not contain demeaning words, it is less likely to be a target of hate speech. At the beginning of our study, our main question was why people become victims of hate speech on social media? We finally found out that victims often use words in their posts that are offensive to other users. And then other users respond to these users by expressing their anger using hateful words in their posts. Consequently, the social media users should have this awareness to think of the words while posting so that the purpose of their post is not to hurt anyone. This research will be helpful in resolving issues such as hate speech and its target prediction. In future, we intend to extend our work on targets' profiles while considering their followers and followees.

Bibliography

S. Agarwal and A. Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *ArXiv*, abs/1701.04931, 2017.

I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238, Oct 2017. doi: 10.1109/ICACSIS.2017.8355039.

A. Anagnostou, I. Mollas, and G. Tsoumakas. Hatebusters: A web application for actively reporting youtube hate speech. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5796–5798. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/841. URL `https://doi.org/10.24963/ijcai.2018/841`.

N. Aulia and I. Budi. Hate speech detection on indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19, pages 164–169, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6106-4. doi: 10.1145/3330482.3330491. URL `http://doi.acm.org/10.1145/3330482.3330491`.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web*

Companion, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4914-7. doi: 10.1145/3041021.3054223. URL https://doi.org/10.1145/3041021.3054223.

A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1105. URL https://www.aclweb.org/anthology/W18-1105.

M. Bouazizi and T. Ohtsuki. Sentiment analysis in twitter: From classification to quantification of sentiments within tweets. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2016. doi: 10.1109/GLOCOM.2016.7842262.

P. Burnap and M. L. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11, Mar 2016. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0072-6. URL https://doi.org/10.1140/epjds/s13688-016-0072-6.

J. Chen, S. Yan, and K.-C. Wong. Verbal aggression detection on twitter comments: convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, Mar 2018. ISSN 1433-3058. doi: 10.1007/s00521-018-3442-0. URL https://doi.org/10.1007/s00521-018-3442-0.

N. Chetty and S. Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, 05 2018. doi: 10.1016/j.avb.2018.05.003.

M. Dadvar, R. Trieschnigg, and F. de Jong. Expert knowledge for automatic detection of bullies in social networks. In *25th Benelux Conference on Artificial Intelligence, BNAIC*

2013, pages 57–64, Netherlands, 11 2013. Delft University of Technology. ISBN not assigned.

T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017. URL http://arxiv.org/abs/1703.04009.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2742760. URL http://doi.acm.org/10.1145/2740908.2742760.

ECHR. The European Court of Human Rights. https://web.archive.org/web/20170316015131/https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b/, 2019. [Last accessed on September 8, 2020].

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257, 2018a. URL http://arxiv.org/abs/1804.04257.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257, 2018b. URL http://arxiv.org/abs/1804.04257.

M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. M. Belding. Peer to peer hate: Hate speech instigators and their targets. *CoRR*, abs/1804.04649, 2018c. URL http://arxiv.org/abs/1804.04649.

Facebook. Facebook Community Standards. https://web.archive.org/web/

20191103031011/https://www.facebook.com/communitystandards/hate_speech/, 2019. [Last accessed on September 8, 2020].

A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*, 2018.

V. Ganganwar. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2:42–47, 2012.

A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and TFIDF based approach. *CoRR*, abs/1809.08651, 2018. URL http://arxiv.org/abs/1809.08651.

GitHub. HateSpeech-Hindi-English-Code-Mixed-Social-Media. https://web.archive.org/web/20180617153028/https://github.com/deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text/, 2018. [Last accessed on September 8, 2020].

Y. Haralambous and P. Lenca. Text classification using association rules, dependency pruning and hyperonymization. In *Proceedings of the 1st International Conference on Interactions Between Data Mining and Natural Language Processing - Volume 1202*, DMNLP'14, pages 65–80, Aachen, Germany, Germany, 2014. CEUR-WS.org. URL http://dl.acm.org/citation.cfm?id=3053762.3053768.

B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering - a tertiary study. *Inf. Softw. Technol.*, 52(8):792–805, Aug. 2010. ISSN 0950-5849. doi: 10.1016/j.infsof.2010.03.006. URL https://doi.org/10.1016/j.infsof.2010.03.006.

N. Kurniasih, L. A. Abdillah, I. K. Sudarsana, I. Yogantara, I. Astawa, R. F. Nanuru, A. Miagina, J. O. Sabarua, M. Jamil, J. Tandisalla, et al. Prototype application hate speech

detection website using string matching and searching algorithm. *International Journal of Engineering & Technology*, 7(2.5):62–64, 2018.

I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1621–1622. AAAI Press, 2013. URL http://dl.acm.org/citation.cfm?id=2891460.2891697.

S. Liu and T. Forss. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1*, IC3K 2014, pages 530–537, Portugal, 2014. SCITEPRESS - Science and Technology Publications, Lda. ISBN 978-989-758-048-2. doi: 10.5220/0005170305300537. URL https://doi.org/10.5220/0005170305300537.

S. Liu and T. Forss. New classification models for detecting hate and violence web content. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01, pages 487–495, Nov 2015.

M. C. McHugh, S. L. Saperstein, and R. S. Gold. Omg u #cyberbully! an exploration of public discourse about cyberbullying on twitter. *Health education & behavior : the official publication of the Society for Public Health Education*, 46 1, 2019.

P. Mishra, H. Yannakoudakis, and E. Shutova. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5101. URL https://www.aclweb.org/anthology/W18-5101.

H. Mubarak and K. Darwish. Arabic offensive language classification on twitter. In I. We-

ber, K. M. Darwish, C. Wagner, E. Zagheni, L. Nelson, S. Aref, and F. Flöck, editors, *Social Informatics*, pages 269–276, Cham, 2019. Springer International Publishing. ISBN 978-3-030-34971-4.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883062. URL https://doi.org/10.1145/2872427.2883062.

K. Nugroho, E. Noersasongko, Purwanto, Muljono, A. Z. Fanani, Affandy, and R. S. Basuki. Improving random forest method to detect hatespeech and offensive word. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 514–518, July 2019.

J. H. Park and P. Fung. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3006. URL https://www.aclweb.org/anthology/W17-3006.

F. M. Plaza-del Arco, M. D. Molina-González, M. Martin, and L. A. Ureña-López. SINAI at SemEval-2019 task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 476–479, Minneapolis, Minnesota, USA, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/S19-2084. URL https://www.aclweb.org/anthology/S19-2084.

F. M. Plaza-del Arco, M. D. Molina-González, M. Martin, and L. A. Ureña-López. SINAI at

SemEval-2019 task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 476–479, Minneapolis, Minnesota, USA, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/S19-2084.

N. I. Pratiwi, I. Budi, and M. A. Jiwanggi. Hate speech identification using the hate codes for indonesian tweets. In *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, DSIT 2019, page 128–133, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450371414. doi: 10.1145/3352411.3352432. URL https://doi.org/10.1145/3352411.3352432.

M. Ribeiro, P. Calais, Y. dos Santos, V. Almeida, and W. Meira Jr. "like sheep among wolves": Characterizing hateful users on twitter. In *cs*, 12 2017.

M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira. Characterizing and detecting hateful users on twitter. In *ICWSM*, 2018.

A. Rodríguez, C. Argueta, and Y. Chen. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174, Feb 2019. doi: 10.1109/ICAIIC.2019.8669073.

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, and T. Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, sep 2016.

T. Y. Santosh and K. V. Aravind. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 310–313, New York, NY,

USA, 2019. Association for Computing Machinery. ISBN 9781450362078. doi: 10.1145/3297001.3297048. URL https://doi.org/10.1145/3297001.3297048.

S. Sharma, S. Agrawal, and M. Shrivastava. Degree based classification of harmful speech using twitter data, 2018a.

S. Sharma, S. Agrawal, and M. Shrivastava. Degree based classification of harmful speech using twitter data. *CoRR*, abs/1806.04197, 2018b. URL http://arxiv.org/abs/1806.04197.

S. Sharma, S. Agrawal, and M. Shrivastava. Degree based classification of harmful speech using twitter data. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 106–112, Santa Fe, New Mexico, USA, Aug. 2018c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4413.

L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. *CoRR*, abs/1603.07709, 2016. URL http://arxiv.org/abs/1603.07709.

D. Stammbach. Offensive language detection with neural networks for germeval task 2018. In *Proceedings of the GermEval 2018 Workshop*, 2019.

S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans. A dictionary-based approach to racism detection in Dutch social media. In *Proceedings of the LREC 2016 Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*. European Language Resources Association (ELRA), 2016.

Twitter. Twitter Rules and policies. https://web.archive.org/web/20191114203551/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy/, 2019. [Last accessed on September 8, 2020].

F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *ITASEC17*, pages 86—-95, Venice, Italy, 1 2017. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). URL http://ceur-ws.org/Vol-1816/paper-09.pdf.

W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390374.2390377.

Z. Waseem and D. Hovy. 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, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-2013.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. In *First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics, 01 2017. doi: 10.18653/v1/W17-3012.

H. Watanabe, M. Bouazizi, and T. Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835, 2018a. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2806394.

H. Watanabe, M. Bouazizi, and T. Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835, 2018b. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2806394.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation,*

SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, pages 75–86, 2019a. URL https://www.aclweb.org/anthology/S19-2010/.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. *CoRR*, abs/1902.09666, 2019b. URL http://arxiv.org/abs/1902.09666.

H. Zhang, D. Mahata, S. Shahid, L. Mehnaz, S. Anand, Y. Singla, R. R. Shah, and K. Uppal. Identifying offensive posts and targeted offense from twitter. *CoRR*, abs/1904.09072, 2019. URL http://arxiv.org/abs/1904.09072.

Z. Zhang and L. Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10:925–945, 2018a.

Z. Zhang and L. Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, Accepted, 10 2018b. doi: 10.3233/SW-180338.

Z. Zhang, D. Robinson, and J. Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, editors, *The Semantic Web*, pages 745–760, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93417-4.

H. Şahi, Y. Kılıç, and R. B. Sağlam. Automated detection of hate speech towards woman on twitter. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 533–536, Sep. 2018. doi: 10.1109/UBMK.2018.8566304.