Automatic Hate Speech Detection on Social Media: A Brief Survey

Ahlam Alrehili

Computer Science Department
Taibah University
El-Madina El-Monawara, Kingdom of Saudi Arabia
amrehili@gmail.com

***Abstract*— Due to the advancement in technology and the explosion of the information age, people communicate with each other indirectly via using the online social networks (OSNs), such as Facebook Snapchat, Instagram, and Twitter. Users of OSNs can post anything without any control or constraint of the content, which leads to increase in spreading of hateful and offensive speech among users, thus resulting in an increase in crimes, murder, and terrorism. Hence, this paper provides a survey and state of the art natural language processing (NLP) technique that is used in automatic detection of the hate speech on OSNs, such as dictionaries, bag-of-words, N-gram etc.**

Keywords—hate speech, OSN, online social network, automatic detection, offensive, hateful dictionaries, bag-of-words, N-gram.

I. INTRODUCTION

In the past, people used to communicate with each other either by meeting at home or in a public place or by using the telephone or mobile phones etc. But, nowadays, with the advancement of the technology and the explosion of the information age, the communication between people has become indirect, and the world has become a global village. So, people can easily and quickly communicate with each other by using social networking sites such as Twitter, Facebook, and Snap Chat. However, each person may have one or more accounts on each social networking site and can post anything without any control or constraint.

Content posted by people may differ from a good content that carries high value for others to abusive speech. The reason behind this difference is the freedom and lack of restrictions that prevent abusive speech. Many online social networking sites are trying to mitigate the effects of non-restrictive content by implementing algorithms solution to detect and strictly prohibit any abusive speech. Despite all attempts for detecting abusive speech automatically, it is still a difficult task owing to multifacetedness of the content.

Hate speech is one of the most abusive speeches that has been recently spread on online social network sites. Hate speech refers to the language that attacks a person or group depending on a set of attributes, such as ethnic origin, religion, sex, race, national origin, sexual orientation, gender identity, and disability[1]. Hate speech was defined by Merriam Webster as a "speech expressing hatred of a particular group of people"[2]. Moreover, hate speech is defined as a "speech that is intended to insult, offend, or intimidate a person because of some attribute (national origin, sexual orientation, disability, religion, or race)[3]."

There is a huge difference between free speech and hate speech, but we can consider the hate speech as a type of free speech. Free speech supports community or individuals to express their ideas and opinions without fear of sanction, censorship, and retaliation [4]. Freedom of speech is required for democratic rights, and it ensures the autonomous enjoyment of the individual. Hence, hate speech is contrary to free speech and violates the fundamental rights of community and individuals as well as reduces the permissible limits in free speech.

In the recent years, the automatic detection of hate speech has become a hot and popular topic in the research field owing to several reasons. European Union Commission, in recent years, has contributed to reducing hate speech by incorporating different initiatives, such as designing programs aimed at fighting the hate speech. Moreover, it imposed Microsoft, Twitter, YouTube, and Facebook to sign a European Union hate speech code that necessitates to review the user post and delete any post containing hate speech in less than 24 hours [5]. On the other hand, there is no specialized tool for automatic detection of the hate speech, and there is lack of data collection, documentation and systematic monitoring of hate and violence.

A hate speech is widely spread on online social network sites (OSN), such as Twitter, Facebook, and Instagram, because it provides an open space for users to express their opinions, beliefs, and ideas without any limitations. Moreover, it gathers users from different countries and nationalities; each of them has a different background, customs, and traditions. However, OSN users tend to use hate speech for a variety of reasons, either the person is proud of his/her nationality or religion or he/she has the feeling of physical safety.

The aim of this research is to conduct a survey and state of the art technique regarding the detection of hate speech in online social networking sites. This research aims at studying the most common and famous natural language processing (NLP) techniques, which have contributed to the automatic disclosure of the hate speech.

The rest of the paper is organized as follows: section II provides a hate speech technique that is mostly used on NLP. Section III provides an overview of hate speech techniques. Section IV presents challenges and opportunities, and this paper is we conclude in section V.

II. AUTOMATIC HATE SPEECH TECHNIQUES

There are many researchers who have conducted research on the automatic detection of the hate speech in the previous literature. These researches are distinguished by a

978-1-7281-5052-9/19/$31.00 ©2019 IEEE

variety of the methodologies and approaches used for the detection of hate speech. Hence, in this section, we will highlight the most important methodologies and approaches used for hate speech automatic detection.

A. Dictionaries

Dictionaries are one of the most important techniques that are used in natural language processing. Dictionaries are based on creating a list of words that appear in a text or context and counting the number of occurrences. The occurrences of these words can be used directly to compute scores or features. In the literature, there are a lot of dictionaries for detection of hate speech, such as:

- Liu et.al[6] built dictionaries based on the content words collected from www.noswearing.com to detect hate and violence web content.
- Dictionaries based on the number of profane words in the text, such as dictionaries built by [7] consisting of 414 words, including abbreviations and acronyms.
- Lexicon based on Ortony to detect a negative effect. The Ortony lexicon contains words that have a negative effect, but each negative word, not necessarily, contains a rude comment or profanity and may be classified as a useful comment.

The dictionaries construction for the detection of hate speech has been done in different ways and in different languages. For example, in the study conducted by Tulkens et.al [8], they built a dictionary to detect racism in Dutch social media comments. They created three discourse dictionaries to classify the text into racist or non-racist category. The first dictionary was created for retrieving more neutral terms and possibly racist terms, while the second dictionary was created by automatic expansion based on using a word2vec model trained on a large corpus of general Dutch text. The third dictionary was created based on filtering out incorrect expansions manually. The aim of this study was to classify the text into racist or non-racist category. They mainly depended on using SVM as supervised machine learning classifier. They achieved precision and recall as 0.49 and 0.43 respectively.

B. Bag-of-words (BOW)

Bag of words is another NLP technique used to detect hateful and offensive speech like dictionaries. There is a difference between the bag of words and dictionaries. Dictionaries are a set of words predefined in the dictionaries, while the bag of words corpus collected from the training data. In the bag of words, after collecting a set of words, word frequencies are selected as a feature for training a classifier. The drawback of this approach is ignoring the word sequence and its semantic and syntactic content. Hence, the word used in various contexts may lead to misclassification. N-gram is adapted to overcome this limitation.

Burnap et.al [9] created several models to categorize cyber hate based on the set of characteristics, including sexual orientation, disability, and race. They used typed dependencies to show the grammatical and syntactic relationships between words. Moreover, they used a bag of words and dictionaries to improve classification for various kinds of cyber hate. Also, they depended on using three supervised learning algorithms (SVM, Random Forest, and Decision Tree) to assist the classification task. They achieved precision and recall as 0.79 and 0.59 respectively, and the F1 score is 0.68.

C. N-gram

N-gram is considered as the most commonly used technique in the automatic detection of hate speech. In sample words, N-gram technique is collected lists of sequential words with size 1,2, 3, N; previous word for an instant, the unigram collect word itself (N=1) as well as bigram collect word based on one previous word (N=2) and so on. Thus, the aim is to list all the expressions of N size and count their frequencies. N-gram is used not only to collect word, but it may be used to collect characters or syllables of a word. This approach is not susceptible to spelling changes when used in words.

Some studies indicate that when higher N values are used in N-gram features such as tri-gram and quad-gram, it leads to better performance than using lower values such as unigram and bigram [10]. On the other hand, according to a survey conducted by [11], researchers reported that using N-grams features on automatic detection of the hate speech is usually highly predictive, but when combined with others features, N-gram performs better.

Kwok et.al [12] used a supervised machine learning approach with N-gram as features for detecting racist tweets against blacks. They classified each tweet as “racist” and “non-racist”. The questionnaire methodology has been used in this research to measure the complexity of how people can identify the hate speech. They collected hundreds of tweets that contained hateful keywords and asked three students with same gender and age but different races to classify whether the tweet was offensive or not. The result showed 33% overall agreement, which is more difficult for machines to do accurately. They distinguished between racist and non-racist tweets by using Naïve Bayes classifier, which achieved an average accuracy of 76% for individual tweets and the average error rate of 24%.

On the other hand, Burnap et.al [13] proposed text classifier based on supervised machine learning to distinguish between antagonistic responses and hateful speech including religion, ethnicity, race, and more general responses. They depended on using N-gram, typed dependencies as the features and Random Forest, Decision Tree and SVM as supervised machine learning classifier. They obtained an overall F-measure of 0.95.

D. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a tool that measures the importance of the word in the document inside a corpus. It is incremented by the number of times that word appears in the document. In fact, it combined two concepts, such as term frequencies and inverse document frequency. Term frequency means the word is important when it appears in many documents, while the inverse document frequency means word appearing in many documents has less importance. Hence, TF-IDF has

given high weight or importance to a word that appears frequency in the document and is found rarely in others document.

In [10], the authors introduced an approach for the classification of a web page content into predefined labels. They depended on using sentiment analysis and applied text summarization techniques to extract sentiment indicators and the topic of web pages. Moreover, they relied on using two NLP features such as TF-IDF and N-gram and Naive Bayes as supervised machine learning classifier. Finally, they achieved overall precision and recall as 0.97 and 0.82 respectively.

E. Part-of-speech

Part of speech has a great importance in revealing the role of the word in context, thus improving the importance of the context. This feature precisely helps in defining the category of the word such as Adjectives (JJ), person singular present form (VBP), personal pronoun (PRP), and Determiners (DT). Part of speech is the feature that helps in detecting the hate speech [14]. This feature is considered as an important feature to detect the pairs of part speech that usually come together (bigram), for example VB_PRP and JJ_DT.

There are some studies based on using POS features, such as the study conducted by Vigna, et al [15]. This study aims at detecting and preventing the hateful speech on a Facebook post written in the Italian language. The hate speech is categorized into different types to distinguish between them. Then, according to the defined classification, the fakebook post is manually annotated by five distinct annotators. After that, they implemented and designed two classifiers by leveraging POS, sentiment polarity, word embedding lexicons, and morpho-syntactical features. Moreover, they depended on using two learning algorithms; Long Short-Term Memory (LSTM), and Support Vector Machines (SVM).

F. Rule-Based Approaches

Rule-based approaches are widely used in the field of text mining, which contains a set of rules created by the human and enriched by linguistic knowledge. It relies on using a dictionary or a pre-compiled list of subjectivity clues, but is not involved in the learning process. Moreover, it uses the method of processing and storing knowledge to interpret information in a meaningful way. For example, it is used to categorize tense content and antagonistic on Twitter as features based on using associational terms. The advantage of rule-based approaches is that it achieved high accuracy as compared to other method, but it, at the same time, takes a lot of time and effort to manipulate system based on rule-based approaches.

In the study conducted by Dennis et.al [16], they relied on disclosure of three kinds of hate speech in web pages, which are religion, nationality, and race. The aim of this study was to create a classifier model based on using Rule-based approach, sentiment analysis, and typed dependencies. Moreover, they depended on using Non-supervised machine learning approach.

G. Sentiment Analysis

Sentiment analysis is also known as opinion analysis, or emotion analysis. AI refers to the use of computational linguistics, text analysis, natural language processing, and biometrics to systematically quantify, extract, identify and study subjective information and affective states [17]. The goal of sentiment analysis is to identify the attitude of a writer, subject or speaker with respect to an emotional reaction to an event, interaction or document, or overall contextual polarity of some topic.

In the context of hate speech, sentiment analysis is considered as a powerful tool to detect hateful speech due to discourse that preaches hatred. The content that exists more in the sentence has a negative polarity. Sentiment analysis is usually used as an auxiliary feature in revealing hate speech.

In [18], the authors aimed at determining the speech that has racist intent on the Tumblr microblogging website. They based on using ensemble learning classifier consisting of Decision Trees, Naïve Bayes, Random Forest, One-class Classifiers, and Random Forest with Topic modeling, sentiment analysis, tone analysis, semantic analysis, and contextual metadata as features. They showed that language cues, social tendencies, personality traits of a narrative, and emotion tone are discriminatory features for identifying the racist intent. They achieved overall precision and recall as 0.73 and 0.86 respectively.

H. Template Based Strategy

Template based strategy refers to creating a corpus of words, and for every word that appears on the corpus, gathering n-word that has occurred around it. It is used as a feature for automatic detection of hate speech. For example, the research conducted by [19] was based on data provided by Yahoo, which contains a set of news that readers found it as offensive speech, and 452 URLs, which were originally collected to categorize websites that advertisers may find inappropriate. From their analysis of the data, they found that the offensive words might be hidden by a deliberate misspelling, substitution of one character of the words by another, or the word may be separated between punctuation marks, such as word “jew” and “j@e@w@”. The reason for this is to escape the automatic detection of hateful words used by some sites.

They have assumed that hate speech is often well distinguished by the Stereotypes in the text. For the classification approach, they used the template-based strategy proposed by [20], which contains set of templates as shown in figure1 ,and each template was centered around a single word. After that, they used a method called log-odds like method proposed by [20], which is called log likelihood.log. Odds are the same as log likelihood, but the difference is that log odds are added as an extra feature. Hence, they calculate log odds value as the flowing:

TABLE I. SUMMARY OF HATE SPEECH RESERCH

Years	Reference	Aim of Study	Futures	Algorithm	P	R	F
2012	[19]	Detecting hate speech on web bags.	Template-based strategies, word sense, disambiguation.	SVM.	0.68	0.6	0.63
2013	[12]	Detecting racist tweets against blacks.	N-gram.	NaiveBayes.	-	-	-
2014	[13]	distinguish between antagonistic responses and hateful speech	N-gram typed Dependencies.	Decision Tree, Random Forest, SVM.	0.89	0.69	0.95
2014	[10]	Introducing an approach for the classification of a web page content.	TF-IDF , sentiment Analysis,N-grams, topic Similarity.	Naive Bayes.	0.97	0.82	-
2015	[21]	Detecting hate and violence on the Web Content.	Dictionaries.	NaiveBayes.	52	92	-
2015	[16]	Detecting hate speech on web bags.	Rule-based approach, sentiment analysis typed dependencies.	Non-supervised.	0.65	0.64	0.65
2016	[9]	To categorize cyber hate based on the set of characteristics.	Bag of the word- typed dependencies-dictionaries.	SVM, Random Forest, and Decision Tree.	0.79	0.59	0.68
2016	[8]	To classify the text into racist or non-racist.	Dictionaries.	SVM.	0.49	0.43	0.46
2016	[18]	To determine the speech that has radicalized or racist intent on the Tumblr microblogging website.	Topic modeling, tone Analysis, contextual metadata and sentiment analysis.	Decision Trees,Naive Bayes ,Random Forest,One-class Classifiers.	0.73	0.86	-
2017	[15]	Detecting and preventing the hateful speech on a Facebook post written in the Italian language.	POS, sentiment polarity, word embedding lexicons, and morpho-syntactical features.	Long Short-Term Memory (LSTM) and Support Vector Machines (SVM) .	0.833	0.872	0.851
2018	[3]	Detecting hat espeech on twiter.	Unigrams,smentic feature sentmental and writing patterns.	SVM.	0.88	0.87	0.87

- Generate templates for each paragraph in the corpus
- Count the occurrences of the positive and negative words.
- Calculate the log-odds, which are the ratio of positive to negative occurrences.
- Discard each template that does not occur at least once for both positive and negative. This process produced 4379 features.

After calculating the feature, they applied it on SVM classifier and achieved overall precision and recall as 0.68 and 0.6 respectively.

Table 1: Example Feature Templates

unigram	"W+0:america"
template literal	"W-1:you W+0:know"
template literal	"W-1:go W+0:back W+1:to"
template part of speech	"POS-1:DT W+0:age POS+1:IN"
template Brown sub-path	"W+0:karma BRO+1:0x3fc00:0x9c00 BRO+2:0x3fc00:0x13000"
occurs in ±10 word window	"WIN10:lost W+0:war"
other labels	"RES:anti-muslim W+0:jokes"

Fig.1. Example of template features

I. Other technics

In the previous sections, we have discussed the techniques most used in the discovery of hate speech. Moreover, there are other techniques that can utilize them in revealing hate speech, such as Profanity Windows, Word Embeddings, Deep Learning, and Distance Metric. These technics can be summarized as follows:

- Profanity Windows: a technique that integrates dictionaries approaches and N-gram. The goal of profanity windows is to test whether the pronoun of the second person is followed by a profane word or not within a certain window[7].
- Word Embeddings: a set of feature learning techniques and language modeling in NLP, which converted phrases or words to the vector consisting of a real number. In the context of hate speech, Djuric[22] used paragraph2vec approach to classify user comment as clean and obvious.
- Deep Learning: it is considered as a new family of machine learning algorithms, which deals with the creation of theories and algorithms that allow the machine to learn by itself by simulating neurons in the human body. It is recently used in sentimental analysis and text classification to achieve high accuracy. Yuan et.al [23] used the deep learning approach with word2vec features for identifying the discrimination on online social network tweets. They achieved overall accuracy of 0.91.

- Distance Metric: some studies on hate speech such as [19] stated that words that denote hate speech are often hidden by a deliberate misspelling or single character substitution. Levenshtein distance is one method that can be used to solve this problem, in which the least number of modifications are needed to convert one string to other [24]. Distance metric is complementary approach to dictionaries based approaches.

III. OVERVIEW OF THE HATE SPEECH DETECTION TECHNIQUES

In the previous section, we have discussed the most widespread natural language processing (NLP) techniques used in hate speech detection. In this section, we will review these techniques to see which of these techniques have contributed significantly to detecting the problem of hate speech.

We have categorized features into three types; token frequencies, linguistic pre-processing, and content analysis. Token frequencies consist of a bag of word, N-gram, profanity windows-TF-IDF, and dictionaries as shown in Table II. Linguistic pre-processing consists of part of speech, rule-based approaches, template-based approaches, and type dependencies as shown in Table III. Finally, a content analysis, which consists of sentiment analysis, is shown in Table IV.

TABLE I. TOKEN FREQUENCIES FEATURES

Token frequencies	
Features	***Reference***
Bag of word	[9], [12], [14]
N-gram	[25],[9],[26],[14],[10],[27],[28]
Profanity windows	[7]
TF-IDF	[10]
Dictionaries	[7],[29],[30],[21].

TABLE II. LINGUISTIC PRE-PROCESSING FEATURES

Linguistic pre-processing	
Features	***Reference***
Part of speech	[10],[13], [14].
Rule-based approaches	[16].
Template based approaches	[31], [19].
Type dependencies	[13], [9],[32],[33],[16].

TABLE III. CONTENT ANALYSIS

Content analysis	
Features	***Reference***
Sentment anaylsis	[34],[26],[15],[16],[21].

Based on information provided by Table II, Table III and Table IV, we note that N-gram is the most widely used feature in the token frequencies, while Type dependencies are highly used in linguistic pre-processing features. Moreover, most research mainly depends on using sentiment analyses in the term of hate speech detection.

On other hand, we noted that some researches depended only on using one feature, such as Kwok et.al [12] used only the N-gram feature as shown in Table I. Other researches combined multiple features, such as Vigna, et al [15] used POS, word embedding, lexicons, sentiment polarity, and morpho-syntactical features.

IV. CHALLENGES AND OPPORTUNITIES

A hate speech is a complex phenomenon, and is considered as a difficult task to detect it. In this section, we will review the major challenges and difficulties explored in previous literature.

- The disclosure of hate speech requires experience and skill in social construction and knowledge in culture [35].
- The development of languages and social phenomena makes it difficult to trace all racial and minority insults [27].
- The language develops very quickly among young people, who use it frequently [35].
- Regarding the concept of the offensive nature of the hate speech, the offensive language may be very elaborate and grammatically correct, crossing the boundaries of the sentence. The use of sarcasm is also common [27].
- The automatic discovery of hate speech is more than searching about keyword[27].

In general, our research regarding the previous studies has found some notes that we would like to highlight as follows:

- Rare to find open source algorithms and platform. Most of the research describes methods, algorithms and features extracted. In fact, open source code helps in developing research faster and eliminating the hate speech problems.

- Most research lacks a clear definition of the data set, and there are no data sets that have been commonly adopted. The clear definition of dataset is considered as an important step in facilitating the comparison of different researches.

- Most hate speech detection research is done in the English language, and very small amount of research has been done in another language, such as Italian, Dutch, and German. Hence, researchers must be directed to detect hate speech in other languages, such as Arabic, France, and Spanish.

V. CONCLUSION

In this paper, we have conducted a systematic literature review of hate speech detection techniques and approaches. We have focused on eight techniques that are widely used for automatic hate detection. These techniques include dictionaries, Bag of the word, N-gram, TF-IDF, sentiment analyses, part of speech, rule-based approach, and template-based approach.

REFERENCES

[1] "Hate speech | Define Hate speech at Dictionary.com." [Online].

Available: https://www.dictionary.com/browse/hate-speech. [Accessed: 22-Nov-2018].

[2] "Hate Speech | Definition of Hate Speech by Merriam-Webster." [Online]. Available: https://www.merriam-webster.com/dictionary/hate speech. [Accessed: 22-Nov-2018].

[3] 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*, vol. 6, pp. 13825–13835, 2018.

[4] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggress. Violent Behav.*, vol. 40, no. April, pp. 108–118, 2018.

[5] "Facebook, YouTube, Twitter and Microsoft sign EU hate speech code | Technology | The Guardian." [Online]. Available: https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code. [Accessed: 23-Nov-2018].

[6] A. Fred *et al.*, *IC3K 2015 : 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management : proceedings : 12-14 November 2015, Lisbon, Portugal*. .

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context."

[8] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A Dictionary-based Approach to Racism Detection in Dutch Social Media."

[9] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, p. 11, 2016.

[10] S. Liu and T. Forss, "Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2014, pp. 530–537.

[11] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," *Proc. Fifth Int. Work. Nat. Lang. Process. Soc. Media*, no. 2012, pp. 1–10, 2017.

[12] I. Kwok and Y. Wang, "Locate the Hate: Detecting Tweets against Blacks," 2013.

[13] P. Burnap and M. L. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making."

[14] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, 2004, p. 468.

[15] F. Del Vigna, A. Cimino, F. Dell'orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook."

[16] N. Dennis Gitari, Z. Zuping, H. Damien, and J. Long, "A Lexicon-based Approach for Hate Speech Detection," *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, 2015.

[17] "Sentiment analysis." [Online]. Available: https://en.wikipedia.org/wiki/Sentiment_analysis. [Accessed: 25-Nov-2018].

[18] R. A. Stevenson, J. A. Mikels, and T. W. James, "Characterization of the affective norms for English words by discrete emotional categories.," *Behav. Res. Methods*, vol. 39, no. 4, pp. 1020–4, Nov. 2007.

[19] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," 2012.

[20] D. Yarowsky, "DECISION LISTS FOR LEXICAL AMBIGUITY RESOLUTION: Application to Accent Restoration in Spanish and French."

[21] A. Fred *et al.*, *IC3K 2015 : 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management : proceedings : 12-14 November 2015, Lisbon, Portugal*. .

[22] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings."

[23] S. Yuan, X. Wu, and Y. Xiang, "A Two Phase Deep Learning Model for Identifying Discrimination from Tweets."

[24] B. S. Nandhini and J. I. Sheeba, "Cyberbullying Detection and Classification Using Information Retrieval Algorithm," in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET '15*, 2015, pp. 1–5.

[25] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," Jun. 2017.

[26] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language *."

[27] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content."

[28] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter."

[29] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of Textual Cyberbullying.," *Soc. Mob. Web*, pp. 11–17, 2011.

[30] W. Maloba, "Use of regular expressions for multilingual detection of Hate speech in Kenya," 2013.

[31] D. M. W. Powers and Ailab, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," vol. 2, no. 1, pp. 37–63, 2011.

[32] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015.

[33] M.-C. De Marneffe and C. D. Manning, "Stanford typed dependencies manual," 2008.

[34] S. Agarwal and A. Sureka, "Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website."

[35] E. Raisi and B. Huang, "Cyberbullying Identification Using Participant-Vocabulary Consistency," Jun. 2016.