International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

An Effective Approach to Hate Speech Detection on Social Media

M. Kumara Swamy[1], U. Padma Jyothi[2]

[1]M.Tech Student, Department of CSE, Vishnu Institute of Technology, Bhimavaram, India
[2]Assistant Professor, Department of CSE, Vishnu Institute of Technology, Bhimavaram, India

---

Abstract - *Social Network Systems are a great way for online users to stay in touch and exchange information regarding their everyday interests and activities, as well as publish and access documents, images, and videos. Unfortunately, these are the prime place for harmful information to spread. While SNSs facilitate communication and information sharing, they are sometimes used to launch problematic campaigns against certain organisations and individuals. Cyberbullying, hate speech to self-harm, and sexual predatory behaviour are only a few of the serious consequences of large-scale internet offensives. With the rise of social media and its unfortunate usage for hate speech, automatic hate speech identification has become a critical challenge. In this work, we offer a method for detecting hate speech on Twitter based on the automatic collection of unigrams and patterns from the training set. These patterns and unigrams are then employed as features in a machine learning method, among other things. Our experiments on a test set of 2010 tweets reveal that our method detects if a tweet is offensive or not (binary classification) with an accuracy of 87.4 percent, and hateful, offensive, or clean with an accuracy of 78.4 percent (ternary classification).*

Key Words: ***Hate Speech, Twitter, Social Media, Hate Speech Detection, Machine Learning.***

1. INTRODUCTION

Hate speech is defined as any communication act that expresses hatred toward a person or a group based on a trait such as race, ethnicity, gender, sexual orientation, nationality, religion, or another feature [34]. The number of hostile actions is rising as a result of the huge rise in user-generated web content, particularly on social media networks where anybody may make a comment freely and without any restrictions. People may rapidly express their opinions, including hate speech, via social media technology, which subsequently spreads widely and becomes viral if the issues addressed are 'interesting'. It has the potential to cause conflict amongst social groupings. According to the National Police Criminal Investigation Agency of Indonesia's data from 2015, there were 143 cybercrimes in the form of hate speech in Indonesia. In 2016, this number grew to 199. However, this information only pertains to hate speech that has been criminalized and reported to the authorities. Obviously, there are many more hate statements on numerous social media platforms.

Twitter [27] is a prominent social networking platform in Indonesia. Twitter and other social media and microblogging online services allow users to view and analyse user tweets in near real time. Because Twitter users are more inclined to convey their emotions about an event by publishing a tweet, it provides a natural source of data for hate speech analysis [5]. This research can aid in the early detection of hate speech, preventing it from spreading widespread. It's also beneficial for content screening and detecting illegal activity early on [3]. The manual method of identifying nasty tweets is inefficient and unscalable. As a result, an automated method for detecting hate speech in tweets written in Indonesian is required.

Hate speech detection has been proposed in the past, especially for English [35, 13, and 4]. The dataset is from Twitter, and the majority of them employed machine learning techniques. Meanwhile, research on hate speech detection in Indonesian is still uncommon. [26, 2] are the only works on hate speech identification in Indonesian language that we are aware. These papers give Twitter datasets for hate speech identification in Indonesian. To solve this challenge, these works also employed a machine learning method. Hate speech identification is essentially a text classification challenge for us. In this paper, we look at how to determine if a tweet is hated speech or not. For text categorization, Nave Bayes (NB) [9], K-Nearest Neighbors (KNN) [28], Maximum Entropy (ME) [8, Random Forest (RF) [36], or Support Vector Machines (SVM) [1] are often used bag of words features and machine learning approaches.

Internet users are drawn to online social networks (OSN) and microblogging websites more than any other type of website. Twitter, Facebook, and Instagram are becoming increasingly popular among people of many origins, cultures, and interests. Their contents are rapidly expanding, making them a fascinating example of so-called big data. Big data has piqued the interest of researchers who are interested in automating the analysis of people's ideas and the structure/distribution of users in networks, among other things while these websites provide a forum for people to discuss and express their ideas, the sheer volume of postings, comments, and messages make it nearly impossible to maintain control over the content.

Furthermore, because of the diversity of backgrounds, customs, and beliefs, many people use angry and abusive language when conversing with persons from other backgrounds. According to King et al. [23], 481 anti-Islamic hate crimes were committed in the year following 9/11, with 58 per cent of them occurring within two weeks of the tragedy. However, as OSN has grown in popularity, more conflicts have arisen as a result of each major event

nonetheless, while content filtering is a divisive issue, with supporters and opponents [21], such languages continue to exist in OSN. It is even more easily shared among young and elderly individuals than other "cleaner" talks. Burnap et al. [22] suggested that gathering and analyzing temporal data allows decision-makers to investigate the escalation of hate crimes after " trigger" events for these reasons.

However, because hate crimes are frequently unreported to the police, there is a scarcity of "official" information concerning them. To deal with the noise and unreliability of data, we suggest an effective method for detecting both offensive messages and hate speeches on Twitter in this paper. To detect emotive traits, we use writing patterns and unigrams. The remainder of this work is organized as follows: In Section 2, we provide our motives and describe some of our findings; in Section 3, we present our findings and discuss some of our findings; and in Section 4, we discuss some of our findings and conclusions the work that is associated with it. In Section 3, we explicitly state our research goal and outline our proposed hate speech detection approach, including how features are retrieved. We go over our experimental findings in Section 4 and discuss them. The final section of this paper wraps up the discussion and suggests some possible research topics.

The main contribution of this paper is as follows:

1) We propose a pattern-based approach to detect hate speech on Twitter: patterns are extracted pragmatically from the training set and we define a set of parameters to optimize the collection of patterns.
2) In addition to patterns, we propose an approach that collects, also in a pragmatic way, words and expressions showing hate and offence, and use them with patterns, along with other sentiment-based features to detect hate speech.
3) The proposed sets of unigrams and patterns can be used as already-built dictionaries for future works related to hate speech detection.
4) We classify tweets into three different classes (instead of only two) where we make a distinction between tweets showing hate, and those being just offensive.

2. LITERATURE SURVEY

The analysis of subjective language on OSN has been extensively researched and used in a variety of disciplines, including sentiment analysis [12, 25 and 30], sarcasm detection [15, 7], and rumour identification [22]. However, in comparison to the aforementioned issues, hate speech detection has received far less effort. Some of these studies, including those by Warner et al. [31] and Djuric et al. [19], focused on sentences on the internet. In the job of binary classification, the first effort achieved a classification accuracy of 94 per cent with an F1 score of 63.75 per cent, while the second work achieved an accuracy of 80 per cent.

Gitari et al. [20] gathered statements from the most well-known "hate sites" in the United States. They divided the statements into three main categories: "highly hateful (SH)", "weakly hateful (WH)", and "non-hateful (NH)". They utilized semantic and grammatical pattern characteristics, ran the classification on a test set, and came up with an F1-score of 65.12 per cent.

To conduct the classification job into two groups, Nobata et al. [6] employed lexical features, n-gram features, linguistic features, syntactic features, pre-trained features, "word2vec" features, and "comment2vec" features, with an accuracy of 90%.

Other research, on the other hand, focused on the identification of hostile phrases on Twitter. Kwok et al. [11] focused on detecting racist tweets directed towards black individuals. They employed unigram features, which resulted in a binary classification accuracy of 76 per cent. Focusing on hate speech directed at a certain gender, ethnic group, race, or other group links the gathered unigrams to that group. As a result, the constructed unigram dictionary cannot be utilised to detect hate speech directed at other groups as effectively. To differentiate hate speech from clean speech, Burnap et al. [22] employed typed dependencies (i.e., the relationship between words) and bag of words (BoW) characteristics.

3. PROPOSED SYSTEM

1) The system proposes a pattern-based approach to detecting hate speech on Twitter: patterns are extracted from the training set pragmatically, and we establish a set of parameters to optimise the pattern collection.
2) In addition to patterns, we present a method for collecting, in a pragmatic way, hateful and offensive phrases and expressions, and combining them with Patterns and other sentiment-based features to detect hate speech.
3) The proposed unigrams and patterns can be used as pre-built dictionaries in future hate speech detection research.
4) The method divides tweets into three categories (rather than just two), allowing us to distinguish between hateful tweets and those that are just offensive.

Advantages
The following Approaches have been covered by the system.

1) A quick strategy using tweets that are neutral, non-offensive, and do not contain hate speech.
2) An effective strategy that includes offensive tweets but does not include hate or segregate / racist statements.
3) The tactic, which comprises provocative tweets that present hate, racist, and segregator words and attitudes.

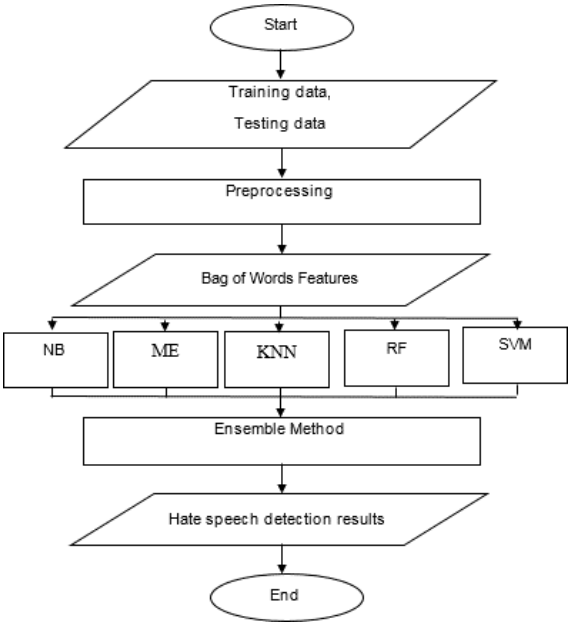

Fig. 1 Hate Speech Detection Flowchart

4. METHODOLOGY

Hate speech is a sort of objectionable language in which the speaker bases his viewpoint on a segregate, racist, or extremist background, as well as stereotypes. Hate speech, according to Merriam-Webster1, is "speech expressing hatred of a specific group of people." It is defined as a "speech designed to insult, offend, or threaten a person because of some attribute (such as race, religion, sexual orientation, national origin, or disability)" from a legal standpoint. As a result, hate speech is regarded as a global issue against which many nations and organizations have taken a position. With the spread of the internet and the growth of online social networks, this problem has gotten even worse, because people's interactions have become more indirect, and people's speech tends to be more aggressive when they feel physically safer, not to mention that many hate groups see the internet as an "unprecedented means of recruiting communication" [21].

Hate speech on the internet and social media not only creates friction between groups of people but may also harm businesses or lead to major real-life confrontations. Hate speech is prohibited on platforms like Facebook, YouTube, and Twitter for these reasons. Controlling and filtering all of the material, on the other hand, is always challenging. As a result, hate speech has been the focus of various studies in the field of research, intending to automatically detect it. The majority of these hate speech detection studies aim to create dictionaries of hate terms and phrases [1] or to categorise hate speech into two categories: "hate" and "non-hate" [31]. However, it is seldom easy to tell if a phrase contains hatred or not, especially if they hate speech is hidden under sarcasm or if there are no explicit terms indicating hate, racism, or stereotyping.

As a result, we suggest a variety of characteristics in this paper, including writing patterns and hate speech unigrams. We utilize these characteristics in combination to classify texts gathered from Twitter (i.e., tweets) into three categories: "Clean," "Offensive," and "Hateful."

4.1. DATA EXPLORATION AND ANALYSIS

4.1.1. DATA

We've gathered and integrated three distinct data sets for this project:

- A first publicly accessible data set on Crowdflower2: this data set comprises over 14, 000 tweets that have been carefully categorized into one of three categories: "Hateful," "Offensive," or "Clean." Three individuals manually annotated all of the tweets in this data collection.
- A second data set, also publicly available on Crowdflower3, which was previously utilized in [29] and has been manually annotated into one of three classes: "Hateful," "Offensive," and "Neither," the latter referring to the previously described "Clean" class.
- A third data set, which was utilized in the study [33] and was released on github4: The tweets in this data collection are divided into three categories: "Sexism," "Racism," and "Neither." The first two ("Sexism," "Racism"), which connect to particular types of hate speech, have been added to the class "Hateful," but the tweets from the class "Neither" have been removed since it is unclear whether they are clean or offensive (several tweets were manually checked, and they have been identified as belonging to both classes).

As previously indicated, the three data sets were merged to form a larger data set, which we divided as described later in this section.

The data set is divided into three subgroups to conduct the classification task:

- **A training set:** There are 21 000 tweets in this collection, evenly split among the three groups (i.e., "Clean," "Offensive," and "Hateful"): each class includes 7 000 tweets. In the rest of this document, this set will be referred to as the "training set."
- **A test set:** There are 2 010 tweets in this set, with 670 tweets in each class. This set will be known as the "test set" and will be utilized to improve our suggested method.
- **A validation set:** There are 2 010 tweets in this set, with 670 tweets in each class. This collection will be known as the "validation set" and will be used to assess our proposed method.

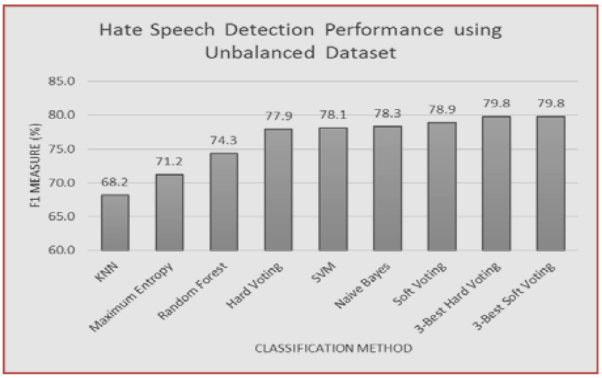

Fig. 2 Hate Speech Detection Performance using Unbalanced Dataset

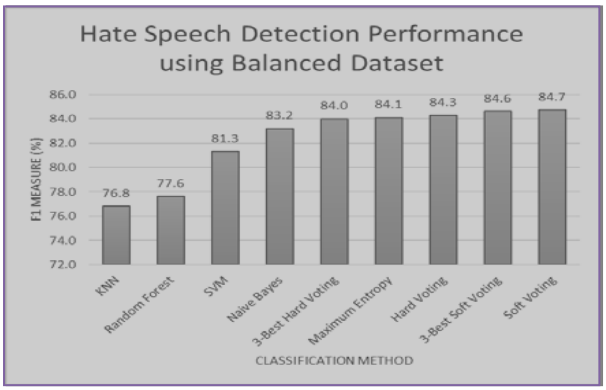

Fig. 3 Hate Speech Detection Performance using Balanced Dataset

5. RESULTS AND ANALYSIS

We move on to our last experiments after extracting features and optimizing settings. The toolkit Weka [22] is used to classify the data. Weka offers a wide range of classifiers that are divided into categories based on the algorithm type (e.g., decision tree-based, rule-based, etc.). We merged the tweets from the two classifications "hateful" and "offensive" into a single "offensive" class (since hateful tweets are indeed offensive and aggressive). This is done in order to make the categorization a binary process. We have 14 000 tweets for class "offensive" and 7 000 tweets for class "clean" in the training set. The number of tweets in the test set for the class "offensive" is 2,680, whereas the number for the class "clean" is 1 340. Run the classification with these sets.

While the binary classification mentioned in the preceding paragraph is essential since it enables for the automatic detection of hostile, aggressive, and hateful remarks with a precision of 93.2 percent, it is a more difficult process.

Table 1: Classification Performances on the Validation Set

Class	TP Rate	FP Rate	Prec.	Recall	F1
Sentiment-based Features					
Hateful	0.337	0.205	0.451	0.337	0.386
Offensive	0.394	0.182	0.520	0.394	0.448
Clean	0.664	0.415	0.445	0.664	0.533
Overall	0.465	0.267	0.472	0.465	0.456
Semantic Features					
Hateful	0.233	0.232	0.334	0.233	0.274
Offensive	0.634	0.467	0.404	0.634	0.494
Clean	0.300	0.217	0.409	0.300	0.346
Overall	0.389	0.305	0.382	0.389	0.371
Unigram Features					
Hateful	0.636	0.073	0.813	0.636	0.714
Offensive	0.652	0.050	0.867	0.652	0.744
Clean	0.924	0.271	0.630	0.924	0.749
Overall	0.737	0.131	0.770	0.737	0.736
Pattern Features					
Hateful	0.328	0.114	0.590	0.328	0.422
Offensive	0.721	0.053	0.872	0.721	0.789
Clean	0.845	0.386	0.523	0.845	0.646
Overall	0.631	0.184	0.661	0.631	0.619
All features combined					
Hateful	0.699	0.104	0.770	0.699	0.732
Offensive	0.763	0.048	0.889	0.763	0.821
Clean	0.891	0.172	0.722	0.891	0.798
Overall	**0.784**	**0.108**	**0.793**	**0.784**	**0.784**

Even though comparing patterns is difficult (since patterns do not have a clear relationship to a specific class), we feel that the same problem exists, and the patterns retrieved from both classes are highly similar and connected to one another.

TABLE 2: Classification Confusion Matrix of the Validation Set

Class	Classified as		
	Hateful	Offensive	Clean
Hateful	**468**	48	154
Offensive	83	**511**	76
Clean	57	16	**597**

6. CONCLUSIONS

We proposed a new strategy for detecting hate speech on Twitter in this paper. Our proposed method classifies tweets into hateful, offensive, and clean categories by automatically detecting hate speech patterns and the most common unigrams, as well as emotive and semantic aspects. For the binary classification of tweets into offensive and non-offensive, our suggested method achieves an accuracy of 87.4 The ternary classification of tweets into hateful, offensive, and clean had an accuracy of 78.4 per cent, and the ternary classification of tweets into hateful, offensive and clean had an accuracy of 78.4 per cent. We will strive to construct a richer dictionary of hate speech patterns in the future, which may be used in conjunction with a unigram dictionary to detect hostile and offensive online messages. We'll conduct a quantitative investigation of the prevalence of hate speech across different genders.

REFERENCES

[1] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, "Offensive Language Detection Using Multi-Level Classification," Advances in Artificial Intelligence, vol. 6085, pp. 16–27, Springer, Ottawa, Canada, June 2010.

[2] Alfina I, Mulia R, Fanany MI, Ekanata Y. "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study". In Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on 2017. IEEE.

[3] Badjatiya P, Gupta S, Gupta M, Varma V, "Deep learning for hate speech detection in tweets". In Proceedings of the 26th International Conference on World Wide Web Companion, 2017 Apr 3 (pp. 759-760). International World Wide Web Conferences Steering Committee.

[4] Barbosa L, Feng J, "Robust sentiment detection on twitter from biased and noisy data". In Proceedings of the 23rd International Conference on Computational Linguistics: Posters 2010 Aug 23 (pp. 36-44). Association for Computational Linguistics.

[5] Burnap P, Williams ML, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making". Policy & Internet. 2015 Jun 1; 7(2):223-42.

[6] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, "Abusive Language Detection in Online User Content," in Proc. WWW'16, pp. 145–153, Apr. 2016.

[7] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in Twitter and Amazon," In Proc.14th Conf. on Computational Natural Language Learning, pp. 107–116, July 2010.

[8] El-Halees AM, "Arabic text classification using maximum entropy". IUG Journal of Natural Studies. 2015 Dec 5;15(1).

[9] Fauzi MA, Arifin AZ, Gosaria SC, "Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model". Indonesian Journal of Electrical Engineering and Computer Science. 2017 Dec 1;8(3).

[10] G. I. Webb, "Decision Tree Grafting from the All-tests-but-one Partition," in Proc. IJCAI'99, pp. 702–707, CA, USA, Aug. 1999.

[11] I. Kwok, and Y. Wang, "Locate the Hate-Detecting Tweets against Blacks," in Proc. AAAI'13, pp. 1621–1622, July 2013.

[12] J. M. Soler, F. Cuartero, and M. Roblizo, "Twitter as a tool for predicting elections results," in Proc. IEEE/ACM ASONAM, pp. 1194–1200, Aug. 2012.

[13] Kwok I, Wang Y, Locate the Hate: Detecting Tweets against Blacks. In AAAI 2013 Jul 14.

[14] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part of-speech tagging for all: Overcoming sparse and noisy data," in Proc. Int. Conf. RANLP, pp. 198–206, Sept. 2013.

[15] M.BouaziziandT.Ohtsuki, "A pattern-based approach for sarcasm detection on Twitter," IEEE Access, Vol. 4, pp. 5477–5488, 2016.

[16] M. Bouazizi and T. Ohtsuki, "Sentiment Analysis: from Binary to Multi-Class Classification - A Pattern-Based Approach for MultiClass Sentiment Analysis in Twitter," in Proc. IEEE ICC, pp. 1–6, May 2016.

[17] M. Bouazizi and T. Ohtsuki, "Sentiment analysis in Twitter: from classification to quantification of sentiments within tweets," IEEE Globecom, Dec. 2016, to be published.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten "The WEKA data mining software: An update'', SIGKDD Explor. Newsk., vol. 11, no. 1, pp. 10–18, June 2009.

[19] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," in Proc. WWW'15 Companion, pp. 29–30, May 2015.

[20] Njagi Dennis Gitari, Z. Zuping, Hanyurwimfura Damien, and Jun Long, "A Lexicon - based Approach for Hate Speech Detection," in, pp., Apr. 2015.

[21] PeterJ.Breckheimer, "A Haven for Hate: The Foreign and Domestic Implications of Protecting Internet Hate Speech Under the First Amendment," in South California Law Review, vol. 75, no. 6, Sep. 2002.

[22] P. Burnap, and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," in Policy and Internet pp. 223–242, June 2015.

[23] R.D. King and G.M. Sutton, "High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending", in Criminology pp. 871–894, 2013.

[24] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," in Proc. 8th Asia Paci□c Finance Assoc. Annu. Conf., vol. 35, pp. 43, July 2001.

[25] S. Homoceanu, M. Loster, C. Lofi, and W-T. Balke, "Will I like it? Providing product overviews based on opinion excerpts," in Proc. IEEE CEC, pp. 26–33, Sept. 2011.

[26] S. H. Pratiwi, "Detection of Hate Speech against Religion on Tweet in the Indonesian Language Using Naïve Bayes Algorithm and Support Vector Machine", B.Sc. Tesis, Universitas Indonesia, Indonesia, 2016.

[27] Sitorus AP, Murfi H, Nurrohmah S, Akbar A, "Sensing Trending Topics in Twitter for Greater Jakarta Area", International Journal of Electrical and Computer Engineering (IJECE). 2017 Feb 1; 7(1):330-6.

[28] Suharno CF, Fauzi MA, Perdana RS, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-Square". Systemic: Information System and Informatics Journal. 2017 Dec 7;3(1):25-32.

[29] T. Davidson, D. Warmsley, M. Macy, and I. Weber "Automated Hate Speech Detection and the Problem of Offensive Language," in Proc. ICWSM'17, May. 2017.

[30] U. R. Hodeghatta, "Sentiment analysis of Hollywood movies on Twitter," in Proc. IEEE/ACM ASONAM, pp. 1401–1404, Aug. 2013.

[31] W. Warner and J. Hirschberg "Detecting hate speech on the World Wide Web," in Proc. Second Workshop Language social media, pp. 19– 26, June 2012.

[32] Z. Zhao, P. Resnick and Q. Mei, "Enquiring Minds: early detection of Rumors in Social Media from Enquiry Posts," in Proc. Int. Conf. World Wide Web, pp. 1395–1405, May 2015.

[33] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proc. NAACL'16 Student Research Workshop, pp. 88–93, June. 2016.

[34] Warner W, Hirschberg J, "Detecting hate speech on the World Wide Web". In Proceedings of the Second Workshop on Language in social media 2012 Jun 7 (pp. 19-26). Association for Computational Linguistics.

[35] Waseem Z, Hovy D, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". InSRW@ HLT-NAACL 2016 Jun 12 (pp. 88-93).

[36] Wu Q, Ye Y, Zhang H, Ng MK, Ho SS, "ForesTexter: an efficient random forest algorithm for imbalanced text categorization". Knowledge-Based Systems. 2014 Sep 30; 67:105-16.