IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

An Improve Framework for hate speech detection using Machine Learning Approach

J. Palimote[1], F. Gaage[2]

Department of Computer Science, Kenule Beeson Saro-wiwa Polytechnic, Bori, River State, Nigeria[1,2]

Abstract: Hate Speech is any correspondence that decries an individual or a gathering based on some trademark, for example, race, identity, sex, sexual direction, ethnicity, religion, or other trademark. Harmful language (e.g., scorn discourse, damaging discourse, or other hostile discourse) principally targets individuals from minority gatherings and can catalyze genuine savagery towards them. The paper proposes an improve framework for hate speech detection using machine learning approach. This system uses a twitter dataset that contains tweeted messages of both hate speech, offensive language, and also messages that is neither hate speech nor offensive language. The dataset was downloaded from kaggle.com, the dataset contains a total of 24,784 twitted messages. The dataset is made up of 8 columns which we later reduced it to two columns by means of feature_extraction. The reduced columns are the tweet columns which contain the twitted messages and the class columns which contains 0,1 and 2, where 0 is classified as hate speech, 1 is classified as offensive language and 2 is classified as neither hate speech or offensive language. we trained our model using support vector machine and random forest classifier and had an accuracy of 95% and 99%. We then deployed our model to web using python flask for easy evaluation and testing. Our experimental results show that our proposed system had better performance in terms of classifying text as hate speech.

Keywords: Hate Speech, Offensive Language, Random Forest Classifier, Support Vector Machine, Machine Learning

1. INTRODUCTION

These days, the quantity of web-based media clients is expanding quickly. Facebook as the market chief, on June 2017 had 2 billion month to month dynamic users1, which is in excess of a fourth of human populace on earth. This shows that online media has become a significant correspondence medium today. Web-based media innovation empowers the message to be sent immediately, become far and wide and even popular if the theme draws in open consideration. Shockingly, this likewise implies that disdain discourse can likewise spread effectively and rapidly that it can prompt clashes between bunches in the public arena. Disdain discourse, particularly concerning race and religion, turned into the most detailed type of online wrongdoing in 2016, as indicated by Indonesian police2. The cop in Indonesia guaranteed at any rate 5 cases detailed every day, which implies there are around 150 each month3. The police additionally said that taking care of digital criminal isn't simple that offices and HR are required. This makes programmed scorn discourse identification is important to be created for the Indonesian language with the goal that the police can identify the spread of disdain discourse rapidly. [1].

Hate Speech is any correspondence that decries an individual or a gathering based on some trademark, for example, race, identity, sex, sexual direction, ethnicity, religion, or other trademark [2].

Harmful language (e.g., scorn discourse, damaging discourse, or other hostile discourse) principally targets individuals from minority gatherings and can catalyze genuine savagery towards them [3]. Web-based media stages are under expanding strain to react, however mechanized evacuation of such substance chances further smothering as of now minimized voices.

With ongoing flood for information, there has been a huge degree in computerized text examination in the area of computational phonetics. Prevalence of assessment rich online assets like audit discussions and microblogging destinations has urged clients to communicate and pass on their musings the whole way across the world continuously. This frequently brings about clients posting hostile and injurious substance internet utilizing derisive discourse. These might be guided towards an individual or network to show their dispute. Recognizing disdain discourse is along these lines significant for legislators and online media stages to debilitate occurrence of any illegitimate exercises. Past examination identified with this undertaking has essentially been centered around monolingual writings [4]. Because of their enormous scope accessibility. Be that as it may, in multilingual social orders like India, utilization of code-blended dialects (among which Hindi-English is generally noticeable) is very normal for passing on feelings on the web. This paper proposes a machine learning approach for hate speech and offensive words detection.

2. RELATED WORKS

Afina et.al, [1] make another dataset that covers Hate Speech by and large, including scorn for religion, race, nationality, and sex. What's more, they likewise led a starter study utilizing machine-learning approach. Machine learning so far is

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

the most habitually utilized methodology in arranging text. they analyzed the presentation of a few highlights and AI calculations for disdain discourse location. Highlights that separated were word n-gram with n=1 and n=2, character n-gram with n=3 and n=4, and negative opinion. The arrangement was performed utilizing Naïve Bayes, Support Vector Machine, Bayesian Logistic Regression, Random Forest, and Decision Tree. A F-proportion of 93.5% was accomplished when utilizing word n-gram include with Random Forest Decision Tree calculation. Results likewise show that word n-gram highlight beat character n-gram.

Watanabe et.al. [5] proposed a way to deal with recognize Hate articulations on Twitter. Their methodology depends on unigrams and examples that are naturally gathered from the preparation set. These examples and unigrams are later utilized, among others, as highlights to prepare an AI calculation. Their trial result on a test set made out of 2010 tweets show that our methodology arrives at a precision equivalent to 87.4% on recognizing if a tweet is hostile (double arrangement), and an accuracy equivalent to 78.4% on distinguishing whether a tweet is contemptuous, hostile, or clean (ternary grouping).

Sap et.al. [6] explore how annotators' lack of care toward contrasts in tongue can prompt racial predisposition in programmed hate speech detection models, conceivably intensifying mischief against minority populaces. They initially reveal surprising relationships between's surface markers of African American English (AAE) and evaluations of poisonousness in a few generally utilized hate speech datasets. At that point, they show that models prepared on these corpora gain and spread these inclinations, with the end goal that AAE tweets a lot without anyone else recognized African Americans are up to multiple times bound to be named as hostile contrasted with others. At last, they propose vernacular and race preparing as approaches to diminish the racial predisposition in comment, indicating that when annotators are made unequivocally mindful of an AAE tweet's tongue they are essentially less inclined to mark the tweet as hostile.

Schmidt and Wiegand [7] gives a short, thorough and organized review of programmed hate speech detection, and blueprints the current methodologies in an efficient way, zeroing in on element extraction specifically. It is primarily focused on Natural Language Processing analysts who are new to the field of hate speech detection and need to educate themselves about the state regarding the craftsmanship.

Bohra [8] break down the issue of hate speech recognition in code-blended messages and present a Hindi-English code-blended dataset comprising of tweets posted online on Twitter. The tweets are explained with the language at word level and the class they have a place with (Hate Speech or Normal Speech). They additionally propose a directed characterization framework for identifying hate speech in the content utilizing different character level, word level, and vocabulary based highlights.

Raghavi e.t al., [9] built up a Question Classification framework for Hindi-English code-blended language utilizing word level assets. The shared assignments have been additionally coordinated on arranging code-blended cross-content inquiry and on data recovery of Hindi English code-blended tweets where the errand was to recover the top k tweets from a corpus for a given question comprising of Hind-English terms where the Hindi expressions are written in Roman transcribed structure.

Del Vigna et al., [10] tended to the issue of Hate speech for Italian language. They constructed their clarified corpus utilizing remarks recovered from the Facebook public pages of Italian papers, government officials, specialists, and gatherings. They led two distinctive grouping tests: the first thinking about three unique classifications of disdain (Strong Hate, Weak Hate and No Hate) and the second thinking about just two classifications, No Hate constantly, where the last class was acquired by blending the Strong Hate and Weak Hate classes. In the two investigations they had the option to accomplish the best correctness's of 64.61% and 72.95% individually.

3. METHODOLOGY

Here, we discuss the architectural components and the processes involved in building and developing our model using machine-learning approach. This processes are:

1. Data Collection

We utilized Twitter data as the wellspring of the dataset and gathering the tweets utilizing Twitter Streaming API7.The tweets were identified with a political occasion, the Jakarta Governor Election 2017. This political decision was a likely wellspring of disdain discourse information since one of its competitors came from a minority bunch in Indonesia, regarding religion and race, while another applicant was a lady that possibly set off scorn discourse identified with sex [1].

2. Pre-processing

We adopted the preprocessing method used by [11] with little modification. There are six steps in the preprocessing stage, i.e. 1) retweet removal; 2) text cleansing; 3) lowercasing; 4) spell correction; 5) negation handling; and 6) stop word removal. The only step in that we did not carry out was hashtag handling. We also converted the texts to arrays using CountVectorizer function.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

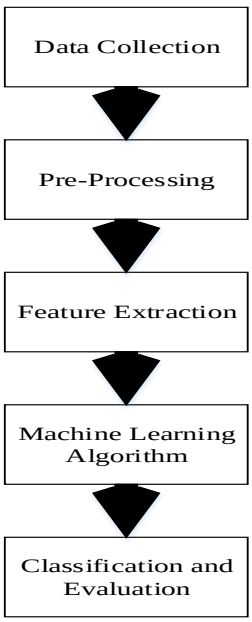

Figure 1: Architecture of the proposed system design

3. Feature Extraction

We used the bag of words (BOW) model [12] in representing the text. In general, we utilized 3 classes of features: word n-gram, character n-gram, and negative sentiment. For word n-gram, we implemented only for n=1 (word unigram) and n=2 (word bigram). For character n-gram, we implemented only for n=3 (character trigram) and n=4 (character quadragram). The usage of character n-gram was based on [13]. For the negative sentiment feature, we adopted the method used by [14] that used sentiment dictionary created by [15] as the basis in counting the number of words in a tweet that has negative sentiment. Thus, we used five features: word unigram, word bigram, character trigram, character quadragram, and negative sentiment.

4. Machine Learning Algorithm

We adopted two machine learning algorithms, which are support vector machine and Random Forex Classifier in building/training our hate speech model.

5. Classification and Evaluation

We used supervised learning approach in detecting hate speech in the Indonesian language. We would compare the performance of four algorithms: Support Vector Machine, and Random Forex Classifier using our dataset. We conducted an experiment by exporting the trained model to web for evaluation and performance of our proposed model.

4. RESULT AND DISCUSSION

This system uses a twitter dataset that contains tweeted messages of both hate speech, offensive language, and also messages that is neither hate speech nor offensive language. The dataset was downloaded from kaggle.com, the dataset contains a total of 24,784 twitted messages. The dataset is made up of 8 columns which we later reduced it to two columns by means of feature_extraction. The reduced columns are the tweet columns which contain the twitted messages and the class columns which contains 0,1 and 2, where 0 is classified as hate speech, 1 is classified as offensive language and 2 is classified as neither hate speech or offensive language. The original and reduced data can be seen in figure 2 and figure 3, while figure 4 shows the histogram of the dataset of both hate speech, offensive language and messages that is neither hate speech nor offensive language. We then apply pre-processing by converting the tweet column from words to binary arrays using CountVectorizer function. We then split the dataset into a training test and a testing test.

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Figure 2: Original dataset which contains a total of 8 columns.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

We trained two machine learning model using support vector classifier, which a training accuracy of about 95.18% and Random Forest Classifier which had an accuracy of about 99.98% which can be seen in figure 5. Figure 6,7, and 8 depicts the graphical interface of the propose system, which is being deployed to web for evaluation and testing.

	tweet	class
0	!!! RT @mayasolovely: As a woman you shouldn't...	2
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	1
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	1
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	1
4	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	1

Figure 3 The reduced dataset in which will applied feature_extraction in removing unwanted columns.

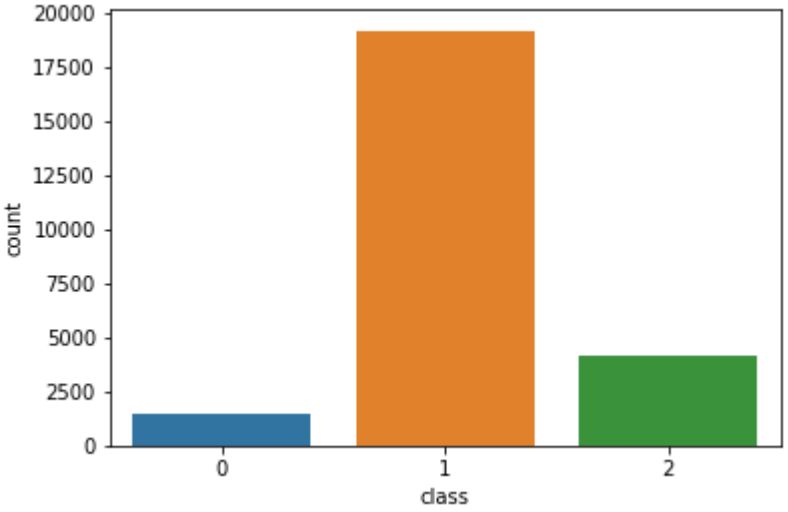

Figure 4 A count plot of the dataset showing the total number of hate speech, offensive language and messages that is neither offensive nor hate speech. The histogram shows that the total number of hate speech is about 19000 while that of offensive language are about 2200 and messages that is neither offensive nor hate speech are about 4800.

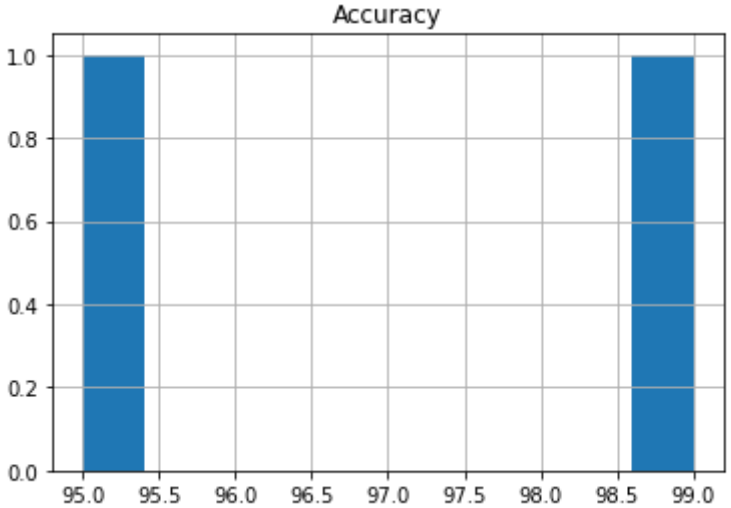

Figure 5: Accuracy of support vector classifier and Random forest classifier, support vector classifier had an accuracy of about 95% while random forest classifier had an accuracy of about 99%

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

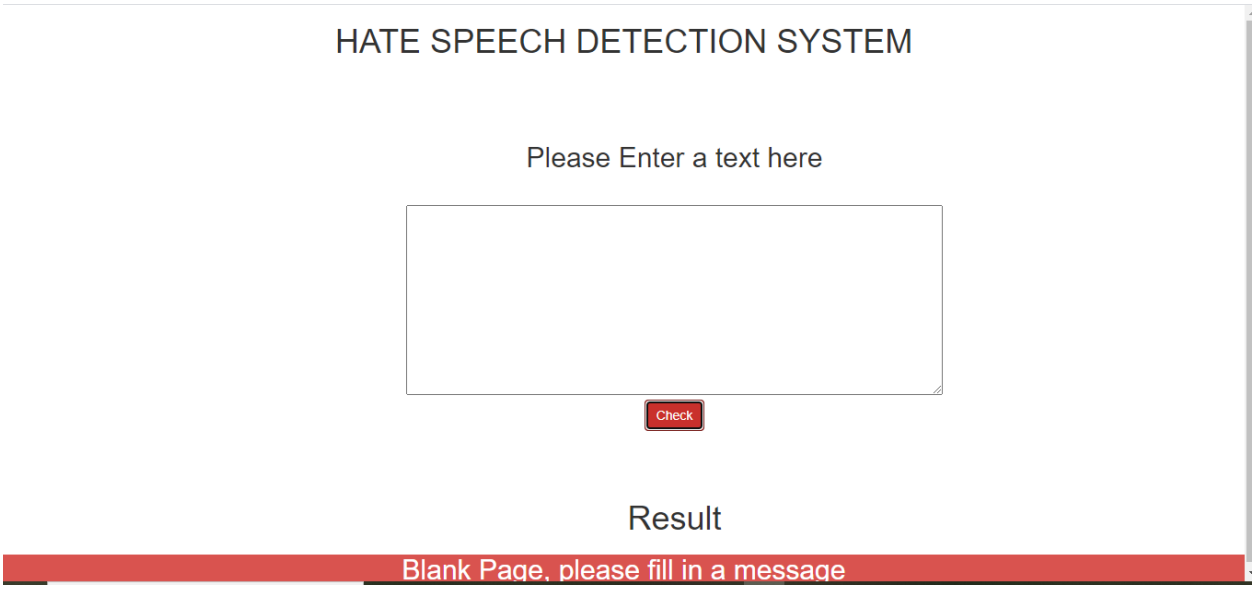

Figure 6: Home Module

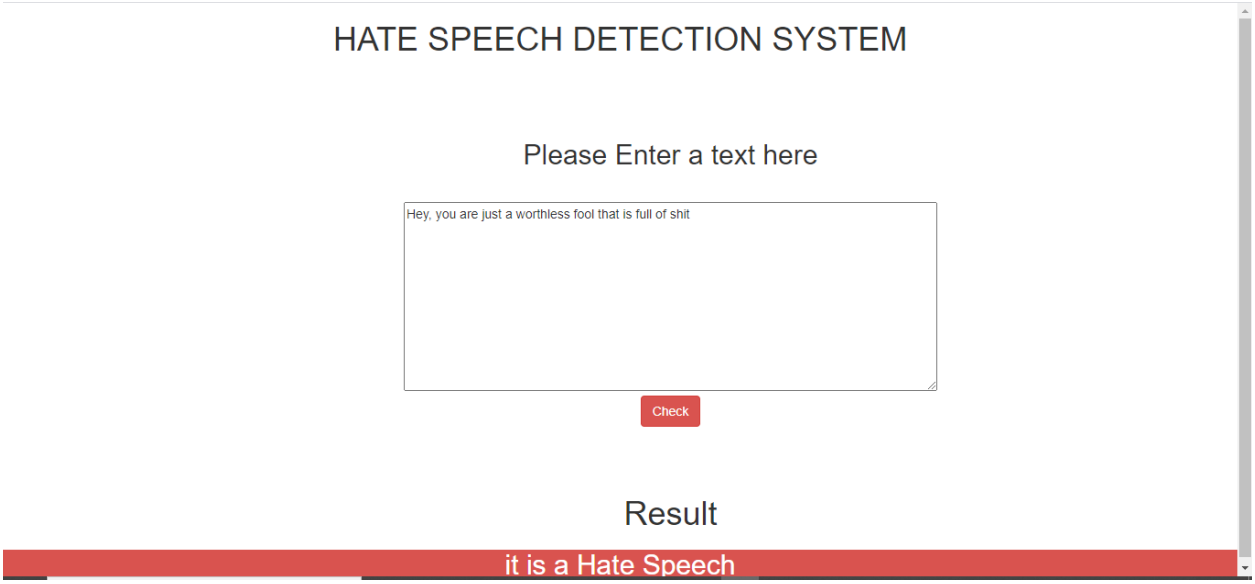

Figure 7: Evaluation result on web of the propose system where it classifies a text as hate speech.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10332

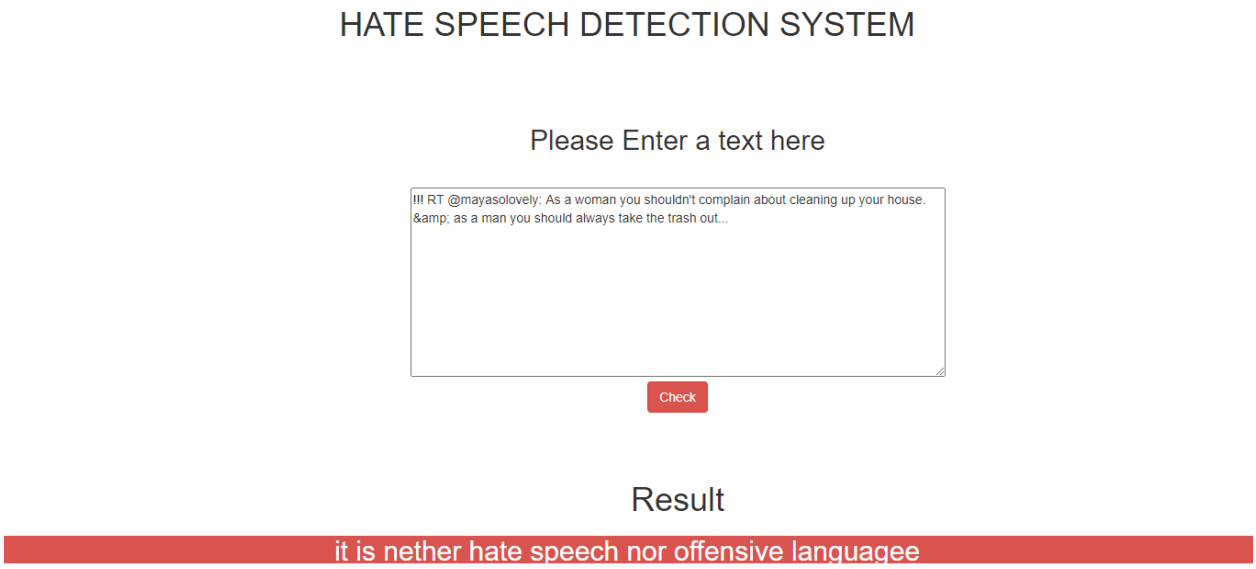

Figure 8: Evaluation result on web of the propose system where it classifies a text as neither hate speech nor offensive.

5. CONCLUSION

Hate Speech is any correspondence that decries an individual or a gathering based on some trademark, for example, race, identity, sex, sexual direction, ethnicity, religion, or other trademark. Harmful language (e.g., scorn discourse, damaging discourse, or other hostile discourse) principally targets individuals from minority gatherings and can catalyze genuine savagery towards them. We proposed a system to detect hate speech using machine learning approach, we trained our model using support vector machine and random forest classifier and had an accuracy of 95% and 99%. We then deployed our model to web using python flask for easy evaluation and testing. Our experimental results show that our proposed system had better performance in terms of classifying text as hate speech.

REFERENCES

[1]. I. Alfina, R. Mulia, M.I. Fanany, Y. Ekanata "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study", International Conference on Advanced Computer Science and Information Systems, pp.233-238, 2017.

[2] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media, no. Lsm, pp. 19–26, 2012.

[3]. Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, and Council on Communications and Media. 2011. The impact of social media on children, adolescents, and families. Pediatrics, 127(4):800–804.

[4]. Anna Schmidt and Michael Wiegand "A survey on hate speech detection using natural language processing". In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10 2017.

[5]. H. Watanabe, M. Bouazizi , 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.

[6] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith "The Risk of Racial Bias in Hate Speech Detection", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678 2019.

[7]. A.Schmidt and M. Wiegand "A Survey on Hate Speech Detection using Natural Language Processing", Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media , pp.1–10, 2017.

[8]. A. Bohra , D. Vijay , V. Singh, S. S. Akhtar, M. Shrivastava "A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection", Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 36–41 New Orleans, Louisiana, 2018.

[9] Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. Answer ka type kya he?: Learning to classify questions in code-mixed language. In Proceedings of the 24th International Conference on World Wide Web, pages 853–858. ACM.

[10] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy*.

[11]. I. Alfina, D. Sigmawaty, F. Nurhidayati, and A. N. Hidayanto, "Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain," in Proceedings of the 9th International Conference on Machine Learning and Computing, 2017, pp. 43–47.

[12]. Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," Int. J. Mach. Learn. Cybern., vol. 1, no. 1–4, pp. 43–52, 2010.

[13] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work., pp. 88–93, 2016.

[14]. 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

[15] C. Vania, M. Ibrahim, and M. Adriani, "Sentiment Lexicon Generation for an Under-Resourced Language," Int. J. Comput., vol. 5, no. 1, pp. 59–72, 2014.