Detecting and Preventing Cyber Bullying Comments on social media Using Deep Learning

M.Goudhaman[1], M.S.Jothi[2]
[1]*Assistant Professor, Jeppiaar Engineering College*
[2]*M.E Scholar Jeppiaar Engineering College*

Abstract—**Everyone has the right to express themselves freely. However, this right is abused under the pretext of freedom of expression to discriminate verbally or physically, or to hurt people. Such intolerance is called hate speech. Hate speech is defined as language that expresses hostility towards an individual or group based on characteristics such as race, religion, ethnicity, gender, national origin, disability, or sexual orientation. It can take the form of statements, sentences, actions, or depictions that single out a person for belonging to a particular group. Both offline and online, hate speech has grown in importance in recent years. Hate content is increasingly proliferated on social media and other online platforms, ultimately leading to hate crimes. Humanity has greatly benefited from the use and information sharing of social media platforms. Nevertheless, this caused many problems, including the spread of hate speech. To address this growing problem on social media platforms, recent research has combined various machine learning and deep-his learning approaches with text-his mining techniques to automatically generate hate speech on real-time datasets. detected at Therefore, the purpose of this research is to review different hate speech detection algorithms and predict the best ones for social media datasets. Additionally, hate speech detection for real-time social settings is now enabled via mobile phone notifications.**

Index Terms— **Social Media, Hate Speech, Deep learning, Text mining.**

I. INTRODUCTION

Social media is a trendy and, most importantly, easy way for people to communicate with others online and openly share their thoughts and opinions. It is become an essential element of daily life. It's a stage where people are more susceptible to abuse or harassment from others who exhibit *hate in a variety of ways, including sexism, racism, politics*, and other types. These social media platforms are increasingly being used for cyber tyranny, online annoyance, and blackmail. We can now easily interact with a variety of societies or organizations that interest us thanks to *social networking sites (SNS)*. Due to the advancement of several technologies, including high-speed internet and portable gadgets, these websites have reached a sizeable portion of the population. In these networks, handlers predominately have ages under thirty. Researchers have conducted considerable research in a variety of subjects by utilising the enormous volumes of data present on different social networking websites. Popular academic discipline called sentiment analysis makes extensive use of data from social media.

Figure 1 shows the various social media platforms.

Fig 1: Social media types

II. RELATED WORK

P. Fortuna and S. Nunes, et al.,[1] Analyzed the challenges of recognizing hate speech, which is labeled in a range of circumstances and platforms and offers a consistent definition. Particularly in online communities and virtual media systems, this area has clear potential to have a positive social impact. The improvement and systematization of shared assets, as well as suggestions, annotated datasets in several languages, and algorithms, are all necessary for the evolution of automated hate speech recognition. Hate speech is defined as language that denigrates or offends people, or incites violence or hatred toward businesses, based on particular characteristics such as physical appearance, faith, descent, national or ethnic origin, sexual orientation, gender identification, or other characteristics. It can take many different linguistic forms, including humor or diffuse bureaucracy.

A. Tolba, Z. Al-Makhadmeh, and others [2] examined 1500 samples to determine whether combining device learning techniques with NLP was advantageous. The automated method was developed to aid in the improvement of hate speech identification and prediction on social networking platforms. Additionally, compared to the conventional method, this one was proven to be faster and more accurate at identifying hate speech. This is due to the killer herbal language processing optimizing ensemble deep learning algorithm (KNLPEDNN) being used to accurately forecast hate and non-hate messages by analyzing Twitter comments and Twitter responses. The suggested method classified comments from beyond-the-facts evaluation, which significantly reduced the misclassification charge, and used massive amounts of Tweets as statistics during the self-learning system.

R. Cao, R. K.-W. Lee, and T.-A. Hoang, et.al,[3] developed DeepHate, a single deep learning model that makes use of various textual representations to detect hate speech on social media. And perform fantastic experiments on three real-world datasets that are accessible to the general public. The test results show that DeepHate routinely outperforms cutting-edge methods in the hate speech identification challenge. The DeepHate version is then empirically investigated, and behavior offers insights into the standout features that helped in identifying hate speech in social media. The prominent feature evaluation enhances our suggested model's capacity to explain.

Z. Waseem and D. Hovy, et.al,[4] provide a dataset with 16k tweets annotated with hate speech. Also take into account which of the variables we employ yields the most accurate identification outcomes. We examine the functions that improve the identification of hate speech in our corpus and discover that, despite expected variations in geographic and phrase-duration distribution, they rarely outperform character-degree functions in terms of overall performance. The one exception to this rule is gender. Additionally, he offered a set of standards for distinguishing racist and sexist utterances that were wholly based on significant race theory. These can be used to gather more data and deal with the issue of a small but incredibly prevalent hateful group. Even if the issue is still far from being resolved, we have found that a man or woman n-gram-based approach offers a strong foundation. With the exception of gender, demographic data only makes a little difference, although this is because to a lack of coverage. We want to enhance area and gender type to update future data and tests.

T. Davidson, D. Warmsley, et.al,[5] categorized tweets as hate speech, profanity, or neither. We teach a model to differentiate between those categories, and then we look at the outcomes to teach it how we will differentiate between them. The results point up a number of significant barriers to effective categorization and suggest that fine-grained tags can help in the detection of hate speech in a publication. We draw the conclusion that future depictions of the use of hate speech must more accurately take context and heterogeneity into consideration. Additionally, they gathered tweets containing important phrases from a crowd sourced dictionary of hate speech. We divide a pattern of these tweets using crowd sourcing into three categories: those that contain hate speech, those that only contain objectionable language, and those that do not. We train a multi-elegance classifier to be able to distinguish between these different classes. When we can reliably identify hate speech from other unacceptable words and when this distinction is more challenging are both shown by an analysis of the expectations and errors. We found that while chauvinist tweets are more likely to be labeled as offensive, racial and homophobic tweets are more likely to be labeled as hate speech. Additionally, it is more challenging to categorize tweets devoid of overt hate speech.

P. Badjatiya, S. Gupta, et.al,[6] Among the classifiers evaluated were Deep Neural Networks, Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs), and Random Forest (DNNs). Project-specific embedding discovered utilizing three deep learning architectures—Feed Neural Networks, Convolutional Neural Networks (CNNs), and Long Short-Term Memory Networks—specifies the feature areas of these classifiers in turn (LSTMs). As baselines, we investigate common spaces like char n-grams, TF-IDF vectors, and Bag of Words vectors (BoWV). This task is quite difficult because to the complexity of the botanical language constructs. We conduct extensive experiments using a variety of deep learning architectures to investigate semantic phrase embeddings that can manage this complexity.

M. O. Ibrohim and I. Budi,et.al,…[7] constructed an Indonesian Twitter dataset with the purpose of recognizing offensive language and hate speech, as well as identifying the goal, category, and severity of such speech. The target, category, and level of hate speech are detected using device learning processes with Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifiers and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) as information transformation techniques. This research discusses multi-label written content grouping for abusive language and hate speech detection in Indonesian Twitter. The function extractions we used were those for term frequency, orthography, and lexicon. Our research's findings show that the RFDT classifier uses LP while it's in style since the transformation strategy offers superior accuracy with fast computation times.

I. Alfina, R. Mulia, et.al,…[8] generated a new dataset for the identification of hate speech in Indonesian, which covers all forms of hate speech, including those motivated by racial, religious, and ethnic animosity. We also conducted preliminary study to determine the best features and device learning rules combinations. The assignment is to look for offensive words in Indonesian. This topic hasn't received much research, as far as I can tell. A dataset for religious hate speech has been produced as a result of the most fundamental study we could find, but the quality of this dataset is insufficient. A new dataset that contained hate speech in general, such as hatred of religion, race, ethnicity, and gender, was something the researchers intended to build. Additionally, we used the system learning approach to conduct a preliminary investigation. Machine learning has been the most popular approach for classifying texts up until this point.

M. O. Ibrohim and I. Budi, et.al,[9] Introduced a new Twitter dataset to track offensive Indonesian language. In order to defend the abusive phrasing and writing styles in Indonesian social media, tests for spotting abusive language were also presented. In this study, we develop a new dataset and investigate the use of harsh language in Indonesian. The test results reveal that using our dataset, NB consistently performs better than SVM and RFDT at classifying abusive language. Phrase unigram and phrase n-gram combinations perform better for capability extractions than alternative features like NB, SVM, and RFDT. The test results also demonstrate that it is more challenging to classify a tweet into one of three categories—non-abusive, abusive but not offensive, or offensive—than to simply decide whether it contains abusive language. The classifier we used struggled to determine if this tweet contained abusive language that was no longer offensive or offensive language itself.

J. Salminen, M. Hopf,et.al,[10] Undertaken the creation of a mobile platform-based online hate classifier. This model is made available to researchers and practitioners for similar use and refinement. It uses cutting-edge language functions, such as Bidirectional Encoder Representations from Transformers (BERT) (see "BERT" phase), to detect nasty feedback across numerous social media networks. Then, numerous experiments with various feature representations and classification techniques (Logistic Regression, Nave Bayes, Support Vector Machines, XGBoost, and Neural Networks) were conducted (Bag-of-Words, TF-IDF, Word2Vec, BERT, and their aggregate). Despite the fact that all models seem to outperform the keyword-based baseline classifier, XGBoost performs brilliantly (F1=0.92). BERT talents had the most influence on the forecasts, according the feature significance analysis. The results point to the applicability of the high-quality version because the platform-specific effects from Twitter and Wikipedia are comparable to their respective supply papers. Make code widely accessible so that it can be used in actual software systems and improved by online hate researchers.

III. EXISTING METHODOLOGIES

Research on text classification in social media has significantly increased during the past ten years. A particularly helpful feature of this effort is identifying and preventing the use of various forms of abusive language in blogs, microblogs, and social networks. In this study, we examine how to distinguish hate speech from common vulgarity on social media. To create lexical baselines for this study, we want to use supervised category methods using a recently available dataset that has been annotated for this purpose.

Figure 2 illustrates the fundamental steps for detecting hate speech.

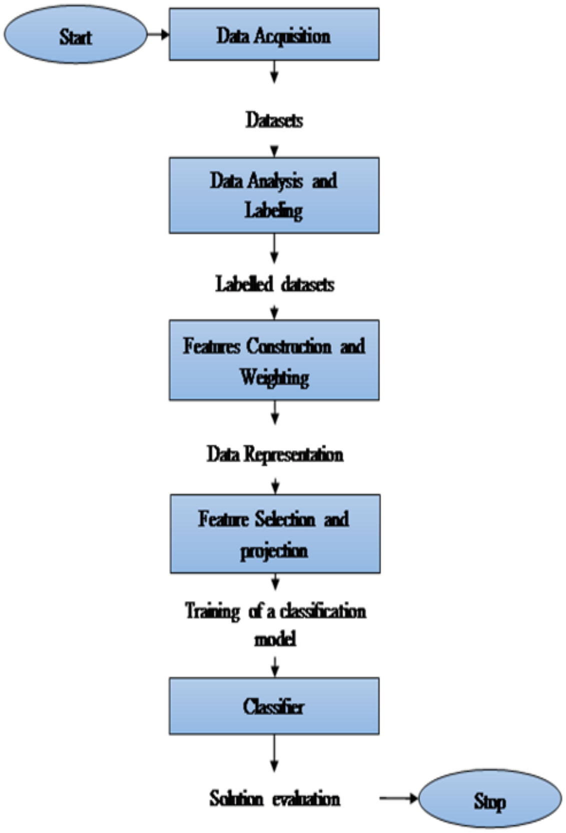

Start
Data Acquisition
Datasets
Data Analysis and Labeling
Labelled datasets
Features Construction and Weighting
Data Representation
Feature Selection and projection
Training of a classification model
Classifier
Solution evaluation
Stop

Fig 2: Steps for detecting hate comments

The majority of social media companies have put in place specific standards to restrict hate speech, but enforcing these regulations requires a lot of human work because each file needs to be reviewed. Recently, certain platforms, like Facebook, increased the number of content moderators. To expedite the reviewing process or allocate human resources to roles that necessitate a comprehensive human evaluation, automatic technology and methods may be deployed. In this section, we examine the process of automatically identifying hate speech in text.

3.1 KEYWORD-BASED APPROACHES

The use of a key-word-based technique is a fundamental strategy for identifying hate speech. Using an ontology or dictionary, it is possible to identify text that contains potentially harmful keywords. For instance, Hatebase keeps a database of derogatory terms for several businesses in 95 languages. Such well-kept artefacts are valuable because terminology evolves with time. However, employing a degrading slur isn't always sufficient to qualify as hate speech, as we discovered throughout our research into hate speech standards. Techniques based on keywords are rapid and easy to understand. They do, however, encounter significant challenges. The most frequent racial slurs could be detected by a highly specific device with low recall, where recall is the percentage of relevant data from the entire population, and precision is the percentage of applicable data from the set discovered. To put it another way, a system that relies heavily on key phrases might not be able to recognize nasty content that doesn't contain these phrases. Include words that aren't often offensive, such as "trash," "swine," and many others, however, would lead to an excessive number of false alerts, enhancing awareness at the expense of accuracy.

MACHINE LEARNING CLASSIFIERS

A classifier that can identify hate speech using labels provided by content reviewers is created by a machine learning model utilizing samples of textual data that has been tagged. Several ideas were put out and shown to work well in the afterlife. We address an improvement to the open-source structures employed in the current investigation in this paper.

i) Content preprocessing and function selection.

To identify or categorize user-generated content, text qualities that signal hate should be retrieved. Individual phrases or sentences that clearly serve a purpose (n-grams, i.e., series of n consecutive phrases). By removing morphological disparities from the root, stemming words can improve function matching. Functionalities can also be extracted during metaphor processing. The bag-of-words hypothesis is a common one in categorizing literary material. According to this method, a submission is represented as a collection of n-grams or phrases that are not necessarily ordered. Although this presumption blatantly ignores an essential component of languages, it has nonetheless shown to be helpful in a number of circumstances. The TF-IDF is one method for allocating weights to the phrases that are more significant in this context. for an outline of modern information retrieval. In addition to distributional characteristics, phrase embedding—which involves giving a vector to a phrase—is frequently used with deep learning techniques in textual content mining and natural language processing, including word2vec. Several deep learning architectures, such as recurrent and transformer neural networks, which imitate the ordering of the words by processing over a succession of word embedding, challenge the bag-of-words hypothesis.

ii) Hate speech detection methods and standards.
Naive Bayes, Support Vector Machines, and Logistic Regression are three text categorization models. Nave Bayes models classify chances immediately under the premise that the features do not interact. The linear classifiers SVMs and Logistic Regression foretell lessons based on a mixture of ranks for each attribute.

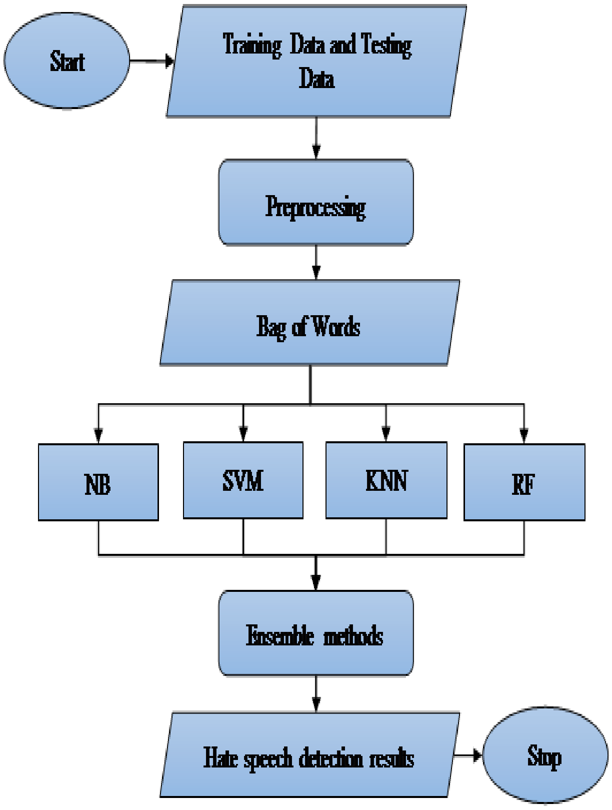


Fig 3: Existing methods for Hate speech detection

IV. PROPOSED METHODOLOGIES

Utilizing social networking platforms is the best approach to meet new people. As social networking sites have gained popularity, people have discovered a dishonest and unlawful way to use them. The most pervasive and harmful abuses of online social media are the expression of hate and harassment. Examples of hate speech include antagonism, bullying, coercion, harassment, racism, insults, provocations, and sexism. These are some of the biggest internet dangers to social networking sites. Deep learning-based algorithms are used to categorize the data and decide if the comments are hostile or commonplace.

- Feed-forward networks consider script to be a collection of words.
- Word relationships and text structures can be captured using RNN-based representations, which treat text as a collection of words.
- CNN-based models are trained to identify textual patterns, such as key phrases, for Term Count (TC).
- Recently, capsule networks have been utilized in TC to alleviate the issue of information loss brought on by CNN pooling operations.
- The construction of DL models can benefit from the attention mechanism, which is active in classifying related terms in text.
- Memory-augmented networks allow models to read and write to datasets by combining neural networks with an external memory.
- Syntactic and semantic parse trees are among the core graph structures of natural language that graph neural networks are intended to capture.

Finally, we can discuss numerous methods for text classification in social media datasets, including machine learning and deep learning techniques. The suggested framework is depicted in figure below.

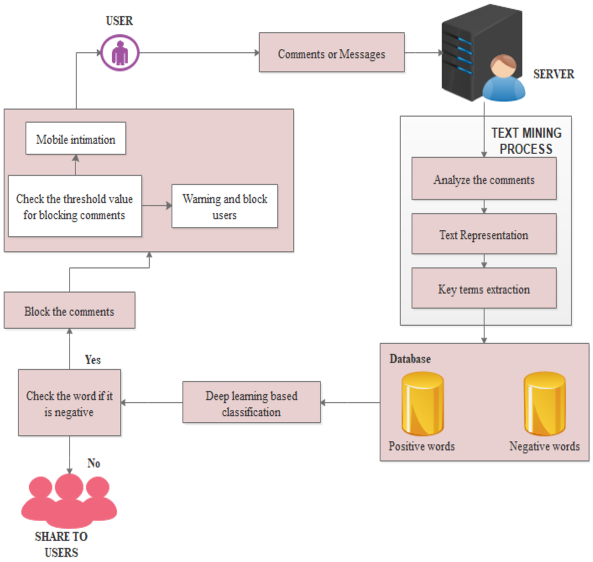


Fig 4: Proposed Work

The majority of efforts in creating a powerful deep learning classifier are concentrated on the extraction and selection of a set of characterizing and discriminating features. The following are the steps of the text mining algorithm:

- Treat text-based review phrases as tokens.
- Analyze unigrams, bigrams, and n-grams
- Remove stop words, analyze stemming words, and remove special characters
- Finally, extract key phrases
- Analyze extended words that can be substituted with right words

Here, a database of classified keywords is established and utilized to screen the words for any offensive words. The communication will be sent to the Blacklists, which will remove any offensive language, if it contains any vulgar keywords. Finally, a message free of profanity will be posted on the user's wall as a result of the content-based filtering technique. The suggested deep learning classifier is as follows:

Step 1: Initialize the neural network model
Step 2: Specify the layer type as convolution, max pooling, fully connected layers
Step 3: Activate the layers
Step 4: Specify the inputs and neurons
Step 5: Construct key terms as positive and negative
Step 6: Match with testing keywords
Step 7: Label as “positive” and “negative”

```
        function INITCNNMODEL ($\theta$, [$n$1–5])

layerType = [convolution, max-pooling, fully-connected, fully-connected];
layerActivation = [tanh(2), max(),softmax()]
model = new Model();
for$i$=1 to 4 do
layer = new Layer();
layer.type = layerType[$i$];
layer.inputSize = $ni$
layer.neurons = new Neuron [$ni$+1];
layer.params = $\theta i$;
model.addLayer(layer);
end for
return model;
end function
```

Blacklists are used by a system to automatically reject unwanted messages based on the relationships and traits of the message authors as well as the message content. To assist users with Filtering Rules(FRs) specification, the set of features assessed during the classification process has been expanded. Additionally, a distinct semantics for filtering rules has been developed to better fit the domain under consideration.

V. EXPERIMENTAL RESULTS

With the help of ASP.NET as the front end and SQL SERVER as the back end, we can build the social network in this chapter. The F-measure parameter can be used to examine the system performance.

Precision, Recall, and F-measure are used to assess the system's performance.

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

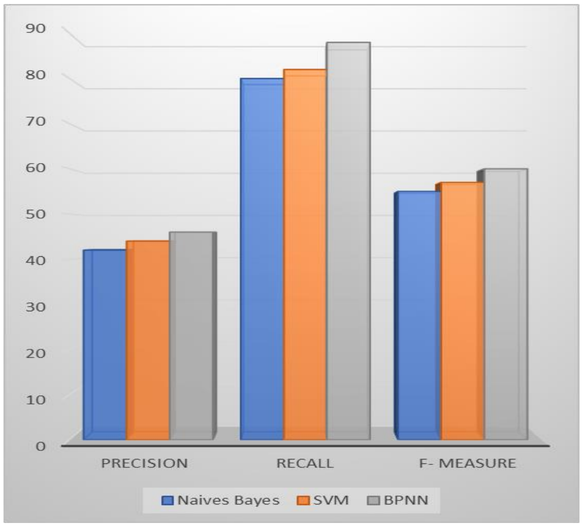
Recall =$\frac{TP}{TP+FN}$

F measure = 2* $\frac{Precision*Recall}{Precision+Recall}$

The performance evaluation result is shown in following table 1 and shows in fig 3.

Table 1: Performance Table

Algorithm/ Performance measures	Precision	Recall	F- measure
Naives Bayes	42	80	55
SVM	44	82	57
BPNN	46	88	60



(a)

Fig 5: Performance chart

According to the calculations above, the suggested neural network technique offers higher F-measure values than the Naives Bayes and SVM algorithms currently in use.

VI. CONCLUSION

In this study, we may examine the current deep learning machine learning models. We could draw the conclusion that a number of issues can be resolved using deep learning models. In this paper, deep learning and machine learning approaches to text classification were examined and contrasted. When learning long-term relationships in this work, we found that different variants of BPNN perform well in sequential learning tasks and address the issues of disappearance and explosion of weights in conventional text classification algorithms. The batch size and hidden size of BPNN models can also have an impact on their performance.

ACKNOWLEDGMENT

I would wish to express my special thanks of gratitude to my supervisors Mr. M.Goudhaman and Mrs. Jospine Jeya who supported me to do this needful project for our society.

I might not forget the support from my family and friends who helped me lots in finalizing this project within the limited time-frame.

REFERENCES

[1] P. Fortuna and S. Nunes, ‘‘A survey on automatic detection of hate speech in text,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.

[2] Z. Al-Makhadmeh and A. Tolba, ‘‘Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach,’’ Computing, vol. 102, no. 2, pp. 501–522, Feb. 2020.

[3] R. Cao, R. K.-W. Lee, and T.-A. Hoang, ‘‘DeepHate: Hate speech detection via multi-faceted text representations,’’ in Proc. 12th ACM Conf. Web Sci., Southampton, U.K., Jul. 2020, pp. 11–20.

[4] Z. Waseem and D. Hovy, ‘‘Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,’’ in Proc. NAACL Student Res. Workshop, San Diego, CA, USA, Jun. 2016, pp. 88–93.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensive language,’’ in Proc. ICWSM, Montreal, QC, Canada, May 2017, pp. 15–18.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ‘‘Deep learning for hate speech detection in tweets,’’ in Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion), Perth, WA, Australia, Apr. 2017, pp. 759–760.

[7] M. O. Ibrohim and I. Budi, ‘‘Multi-label hate speech and abusive language detection in Indonesian Twitter,’’ in Proc. 3rd Workshop Abusive Lang. Online, Florence, Italy, Aug. 2019, pp. 46–57.

[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, ‘‘Hate speech detection in the Indonesian language: A dataset and preliminary study,’’ in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Jakarta, Indonesia, Oct. 2017, pp. 233–238

[9] M. O. Ibrohim and I. Budi, ‘‘A dataset and preliminaries study for abusive language detection in Indonesian social media,’’ ProcediaComput. Sci., vol. 135, pp. 222–229, Jan. 2018.

[10] J. Salminen, M. Hopf, S. A. Chowdhury, S.-G. Jung, H. Almerekhi, and B. J. Jansen, ‘‘Developing an online hate classifier for multiple social media platforms,’’ Hum.-centric Comput. Inf. Sci., vol. 10, no. 1, pp. 1–34, Dec. 2020

[11] A. Jha and R. Mamidi, ‘‘When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data,’’ in Proc. 2nd Workshop NLP Comput. Social Sci., Vancouver, BC, Canada, Aug. 2017, pp. 7–16.

[12] S. Yuan, X. Wu, and Y. Xiang, ‘‘A two phase deep learning model for identifying discrimination from tweets,’’ in Proc. EDBT, Bordeaux, France, Mar. 2016, pp. 696–697.

[13] M. Mozafari, R. Farahbakhsh, and N. Crespi, ‘‘Hate speech detection and racial bias mitigation in social media based on BERT model,’’ PLoS ONE, vol. 15, no. 8, pp. 1–26, Aug. 2020.

[14] 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.

[15] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, ‘‘Detection of abusive language: The problem of

biased datasets,'' in Proc. HLT-NAACL, Minneapolis, MN, USA, Jun. 2019, pp. 602–608.

[16] M. Mozafari, R. Farahbakhsh, and N. Crespi, ''Hate speech detection and racial bias mitigation in social media based on BERT model,'' PLoS ONE, vol. 15, no. 8, pp. 1–26, Aug. 2020.

[17] 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.

[18] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, ''Detection of abusive language: The problem of biased datasets,'' in Proc. HLT-NAACL, Minneapolis, MN, USA, Jun. 2019, pp. 602–608.

[19] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. Sung, and R. Kurzweil, ''Universal sentence encoder,'' in Proc. EMNLP, Brussels, Belgium, Mar. 2018, pp. 169–174.

[20] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, ''SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter,'' in Proc. 13th Int. Workshop Semantic Eval., Minneapolis, MN, USA, Jun. 2019, pp. 54–63.