VOL 4 (2020) NO 4

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

# A Systematic Literature Review of Different Machine Learning Methods on Hate Speech Detection

Calvin Erico Rudy Salim[#], Derwin Suhartono[*]

[#] *Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
*E-mail: calvin.salim@binus.ac.id*

[*] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
*E-mail: dsuhartono@binus.edu*

***Abstract*— Hate speech is one of the most challenging problem internets is facing today. The most common practice to deal with online suspects of hate speech is by manually reporting the comment or the post which at the back end is reviewed by a person. This has a lot of limitations. it requires a lot of time as human intervention is required. Many countries have made laws so that companies have to deal with this type of content within a time frame. This systematic literature review examines hate speech detection problem and will be used to do an experimental approach on detecting hate speech and abusive language. This work also provides an overview of previous research, including methods, algorithms, and main features used. We observe 31,633 papers of current research about hate speech detection from online databases, after applying inclusion and exclusion criteria the result is 1,929 papers and then returned 15 papers after the full text analysis. These papers are for answering the research questions of this systematic literature review. We use two research questions in this literature review which will be the foundation of the next experimental research. Correctly classifying a piece of text as an actual hate speech requires a lot of correctly labelled data. Most common challenges are different languages, out of vocabulary words, long range dependencies and many more.**

***Keywords*— natural language processing; hate-speech; artificial intelligence.**

## I. INTRODUCTION

Online communities are growing at a tremendous rate. Users are uploading tons of data in forms of videos, images, text posts. This has resulted in a very common problem of online hate speech [1]. Various platforms have to deal with it on a daily basis. It is important to ensure that everyone is treated in a decent way. They have to protect their users from threats, hate speech and insults which may have a negative impact on the user [2]. It is equally important to keep these discussions on the right track without harming the dignity of a fellow user. Previous research had been done to prevent hate speech and abusive language from happening in social media, but it's very common to see it in English language rather than Indonesian language.

Another approach which can be used is using automatic detection and removing this type of content using Artificial Neural Network. This itself has various challenges such as out of vocabulary words which cannot be identified and multiple language or mixed language used within a single sentence. Lack of correctly labelled data to train our model is also a problem. Use of offensive words has become a part of life of young generation nowadays [3]. Independently these words can be described as hate speech but in the real context, they are used in a general way. All these things make it difficult to really classify something as hate speech. It is also very difficult to track all the racial and abusive words used for a particular group. These words change with time and keeping an active track of blacklisted words is a hard task. All these problems and challenges are found in our literature review, and will be used for improving our upcoming experimental research model in Indonesian language.

Sometimes the hate speech is so fluent and grammatically correct that it becomes very hard to spot such phrases as the whole phrase is equally contributing to it [4]. Hate speech is not just confined to a single line in a sentence. Many a times we have to consider the other sentences to conclude the actual meaning hidden behind the simple words. It requires a lot of knowledge to resolve and find the actual meaning of these types of sentences [5]. These types of comments are mostly used as it is hard to classify them and less chances of getting removed. People come up with very creative ways to insult others.

Other systematic literature review such as [6] use soft computing techniques and [7] use benchmark corpora. In this research, we focus on using the word "hate speech" which according to [8] is a term that hits a particular community or individual that makes them suffer, while the opposition doesn't care. And we use the term "abusive language" which according to [9] is speech that contains harsh words or phrases that are conveyed to the interlocutor (individual or group), both orally and in writing. Based on [10], users are freely express their expression and bad words in social media. Because of this, our proposed approach is by using the research questions in this literature review to find best method to classify hate speech and abusive language. We also ask which types of dataset is used in other research to find out which one is the best type for our upcoming research.

## II. Material and Method

The Systematic Literature Review was driven by the following research questions :

1. What are the most effective machine learning method for hate speech detection?
2. What types of dataset that are the most widely used?

We choose these research questions because we want to know what is the most effective machine learning method for hate speech detection. That way, we can implement the most effective method in our experiment to get the best result. We also ask the second research question is because we also need to observe and investigate what kind of dataset is best for our future research. The articles and research papers that are being searched this study are found from online database libraries such as IEEE Xplore, Springer Link, Science Direct, and ACM Digital Library. All papers were from 2015 until present, written in English, and met a certain set of criteria that will be included in this paper. The search keywords or string format used in the search was :

- (“hate speech” OR abusive OR offensive)
- (neural network) AND (machine learning)

The first search string used is “find at least one in contexts or titles” from advanced search feature of digital libraries. It is for finding all papers that related to hate speech detection. Second search string used in “find exactly the same” from advanced search feature. This search string used to filter every research paper that are not related with neural network and machine learning.

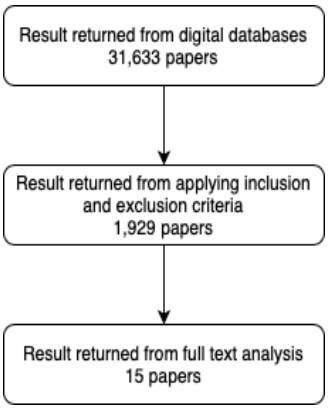


Fig. 1. Flow of paper selection process.

From figure 1, we can see that our keywords returned 31,633 papers from all the databases. After we apply some filter inclusion and exclusion criteria, we got 1,929 returned papers. And after from all those results, we only pick the papers with full access and are open for public. So in the end, we got 15 papers as our result of finding.

## III. Result and Discussion

Each paper that had been analysed to fulfil all questions mentioned before. 15 papers were selected from previous process. The papers are listed below in Table I.

From paper [11], it described about Neural Language model that use similarity of words concept. Words which are close to each other in a sentence are more dependent. They used Continuous Bag Of Words (CBOW) model as a component of paragraph2vec. It is based on surrounding words. It tries to predict the most important word in sentence around which the whole sentence is based. This model tries to find the words which are in neighborhood of this type of words and takes the central word to guess whether it is hate speech or not. This approach is very good as it requires less training and is gives very promising results.

Paper [12] is about using deep learning as a method to detect hate speech on Twitter. With the current increase of interaction on social networks, there has been also increase of hateful activities. This paper experimented with multiple classifiers such as Random Forest, Logistic Regression, Gradient Boosted Decision Trees (GBDT), SVMs and Deep Neural Networks (DNN). There are 3 deep learning architecture that is being used for this research. Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and FastText. All of these networks are fine-tuned using labeled data with back-propagation. The research found out that it significantly outperform the existing methods which is char n-grams, word Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words Vector (BoWV). Embeddings learned from DNN models combined with GBDT led to best accuracy values.

Paper [13] talked about detecting hate speech in social networks with seven model approached. Social networking has changed the way people interact online, that’s why it allows malicious entities to influence opinions by posting hateful material. The dataset contains about 300k data from different sources of Wikipedia and Twitter. Methods used are Gated Recurrent Units (GRU), LSTM, and Logistic Regression. Turns out that the simple method using Logistic Regression with word n-gram is the most effective than the more complex model using DNN, GRU or CNN. The research suggest that future work should focus more on the dataset rather than the model.

Paper [14] approached a model called Hate2Vec, a method for offensive comment detection based on word and comment embeddings. The research use multiple dataset, with one from another research contains 16k of English tweets annotated for hate speech with three labels and another dataset from Kaggle that contains 6k of English tweets with two labels. The research compared other method which is SVM, and the result showed that Hate2Vec has an average F1-Score of 0.93 while SVM was 0.80 on average

TABLE I
LIST OF SELECTED PAPERS

| Source | Publication Year | Title |
|---|---|---|
| ACM Digital Library | 2015 | Hate Speech Detection with Comment Embeddings [11] |
| ACM Digital Library | 2017 | Deep Learning for Hate Speech Detection in Tweets [12] |
| ACM Digital Library | 2018 | All You Need is "Love": Evading Hate Speech Detection [13] |
| ACM Digital Library | 2018 | A Classifier Ensemble for Offensive Text Detection [14] |
| ACM Digital Library | 2019 | Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach [15] |
| ACM Digital Library | 2020 | Towards Automatic Detection and Explanation of Hate Speech and Offensive Language [16] |
| ACM Digital Library | 2020 | A Multilingual Evaluation for Online Hate Speech Detection [17] |
| IEEE Xplore | 2018 | Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection [18] |
| IEEE Xplore | 2020 | Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets [19] |
| IEEE Xplore | 2020 | Automatic Detection of Offensive Language for Urdu and Roman Urdu [20] |
| Science Direct | 2018 | A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media [21] |
| Science Direct | 2020 | Detection of Hate Speech Text in Hindi-English Code-mixed Data [22] |
| Springer Link | 2016 | Us and them: identifying cyber hate on Twitter across multiple protected characteristics [23] |
| Springer Link | 2018 | Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments [24] |
| Springer Link | 2020 | Developing an online hate classifier for multiple social media platforms [25] |

Paper [15] talked about hate speech detection in long text documents using machine learning approach. What makes this paper different is the dataset of documents are in Indonesian language. It also shows that there is an increasing number of political practices that use hate speech in 2017 worldwide. The data was collected in Facebook posts that mainly talks about politics. The data was annotated by 3 annotators consist of females and males aged between 20 – 26 years old with different domicile and background. The result shows that combining SVM with TF-IDF, offensive words, positive words, char quad-gram and word unigram have the best performance with F1-score of 85%.

Paper [16] is about a novel system called HateDefender. It is based on deep LSTM neural network[11]. The average accuracy is 90.82% on hate speech detection and 89.10% on offensive language detection. It uses a dataset from another research authored by Davidson et al. [26] consisting of tweets labeled as hate speech, offensive language or neither. The model outperform baseline model explained in the research paper.

Paper [17] is about detecting hate speech in social networks. Hate speech increased of attacks targeting spesific groups of users based on their religion, ethnicity, or social status. The research use 3 different language dataset; 16k of English tweets containing 32% of hate speech, 4k of Italian tweets containing 32% of hate speech, and 5k of German tweets containing 34% of hate speech. Methods used in this research were LSTM, GRU and Bidirectional LSTM (BiLSTM) with FastText embedding. The result showed the highest score for Italian dataset is by using unigram with average F1-Score of 0.801. However the German and English dataset achieve the highest score is by not using unigram with average F1-Score of 0.785 for English and 0.718 for German.

Paper [18] is about a pragmatic approach to collect hateful and offensive language for hate speech detection. The dataset is combined from Crowdflower and dataset used from another research [27] that has been manually annotated. The training set contains 21,000 tweets and the test plus validation set contains 2,010 tweets with each class has 670 tweets. The results shows that this research paper approach reaches an accuracy equal to 87.4% on detecting a tweet is offensive or not with binary classification, and another accuracy equal to 78.4% on detecting a tweet is hateful, clean or offensive with ternary classification.

Paper [19] is about detecting hate speech in South African tweets using machine learning approach. Twitter is the most used social media in South Africa. A total of 21,350 tweets was collected between the period of May 5, 2019 until May 13, 2019 using Twitter Achiver. Non-English tweets were removed except code-mixed English tweets with Sesotho, Isizulu and Afrikaans words. The machine learning method used in this research were SVM, Random Forest, Gradient Boosting and Logistic Regression. Optimized SVM with character n-gram achieved the best true positive rate of 0.894 for hate speech with overall accuracy of 0.646, while it recorded very low true positive rate of 0.069 for offensive speech.

Paper [20] is about offensive language detection in Urdu and Roman Urdu language. The Roman Urdu dataset that was being used is publicly available on Github, contains 147,000 user comments. However there is no standard Urdu dataset that can be used for offensive language detection, instead the authors manually built and collected

the user comments from Youtube video about political, entertainment, religion and sports that is uploaded in India and Pakistan. 3 graduate students and local speakers are assigned to annotate the dataset. The methods used are LogitBoost, which is based on AdaBoost procedure that trains the model on weighted samples. And another method is simple regression function (SimpleLogistic). Other techniques also used in this research such as SVM, Naive Bayes, Hoeffding Tree and K-Nearest Neighbor The results were LogitBoost achieved F1-score of 99.2% on Roman Urdu dataset using character tri-gram and 94.9% on Urdu dataset. SimpleLogistic achieved F1-score of 95.8% on Urdu dataset using also character tri-gram and 98.3% on Roman Urdu dataset.

Paper [21] is about detecting hate speech in Indonesian tweets. Indonesia is one of the countries that use social media for many purposes. The dataset was crawled and filtered manually from Twitter's API and was annotated by 20 volunteers. It is used with 100% annotators agreement, so tweets that has a different label removed from the dataset. That gives 2,016 total tweets. The method that was used in this research were Naive Bayes, SVM and Random Forest Decision Tree (RFDT). The result shows that Naive Bayes is better than SVM and RDFT with F1-Score of 86.43% using word unigram feature extraction.

Paper [22] is about detecting hate speech in Hindi-English code-mixed data. Around 44% of the Indian population speak Hindi, so the usage of Hindi-English language is very high in Twitter and Facebook. The dataset used is from a combination of 3 research paper, so there was a total of 10,000 texts and it's divided equally to 2 class namely hate and non-hate. The methods that was used are SVM, SVM-Radial Basis Function(SVM-RBF) and Random Forest. The result shows that SVM-RBF combined with FastText gives F1-Score of 85.81%, higher compared than SVM-RBF combined with word2vec that gives F1-score of 75.11%.

Paper [23] talked about identifying cyber hate speech on Twitter. It uses CrowdFlower to annotate the tweets that contains hate speech about disability, race, sexual orientation, or none. The classification was done using SVM and RFDT combined with BoWV. The result showed that overall F1-Score is 0.96.

Paper [24] approached a method for hate speech detection in digital microenvironments. The combination of the people (i.e., accounts), who say things (i.e., tweets) to other people (i.e., other accounts), is the definition of digital microenvironments in cyberspace. Dataset contains 9,488 annotated tweets. The method for classification uses RFDT. This research focus more on the metadata rather than text variables.

Paper [25] is about making an online hate speech classifier for multiple social media platforms. Around 22% of adult have experienced offensive name-calling. The dataset used in this research were from other research. There are 4 social media dataset that were used such as Youtube, Wikipedia, Twitter and Reddit. A total of 197,566 comments were collected and 80% of the comments were labelled as non-hateful and the rest 20% were labelled as hateful. The result showed that XGBoost was the best classifier combined with BERT as the best feature representation with the F1-Score of 0.916.

## IV. Research Questions

### A. *What are the most effective machine learning method for hate speech detection?*

TABLE II
LIST OF METHODS FOR HATE SPEECH DETECTION

| Method | Study Id |
|---|---|
| Paragraph2Vec | 11 |
| Term Frequency – Inverse Document Frequency | 11, 12, 15 |
| Long Short-Term Memory | 13, 12, 17, 16 |
| Gradient Boosted Decision Trees | 12, 19 |
| FastText | 12, 17, 22 |
| Convolutional Neural Network | 13, 12, 17 |
| Random Forest Decision Tree | 15, 19, 20, 21, 22, 23, 24 |
| Support Vector Machine | 14, 15, 17, 19, 20, 21, 22, 23, 25 |
| Logistic Regression | 13, 14, 15, 19, 20, 25 |
| J48 Graft | 18, 20 |
| Naive Bayes | 17, 21, 25 |
| XGBoost | 25 |

As we can see from Table II, the most used machine learning method for hate speech detection is Support Vector Machine. But while the most used method is SVM, the best result is achieved by using LSTM model. However, this might be due to difference of dataset used for training to solve different types of issue.

### B. *Which types of dataset that are the most widely used?*

TABLE III
LIST OF DATASET TYPES FOR HATE SPEECH DETECTION

| Type | Study Id |
|---|---|
| 2 label dataset | 11, 15, 17, 20, 22, 24, 25 |
| 3 label dataset | 13, 12, 14, 16, 18, 19, 23, 21 |
| Balanced | 18, 20, 22, 25 |
| Imbalanced | 13, 11, 12, 14, 15, 16, 19, 21, 23, 24 |
| 10,000 – 20,000 data | 11, 12, 15, 17, 19, 20, 21, 22, 23, 24 |
| 20,001 – 30,000 data | 14, 16, 17, 18 |
| More than 30,000 data | 13, 25 |

Table III shows that the most used dataset contains 10,000 to 20,000 data. It also shows us that most of the research use imbalanced dataset. Ganganwar [28] stated that an unbalanced dataset could give negative result for classification. This is due to difference number of datasets between major and minor class that could make the major class have better performance than minor class. Table III also show us that the use of binary classification with 2 label dataset and 3 label dataset are the same

## V. Conclusion

From the results at section 4, it shows that there is a possibility to do a research of hate speech detection. We

decided to do an experiment for hate speech and abusive language detection using LSTM method. Today most of the research has been done on English datasets. There are many other languages for which this research needs to be carried upon. This is why our experiment will be based on Indonesian language tweets made publicly available by Ibrohim et al. [29]. The dataset contains 13,169 tweets that consist of 7,608 not hate speech and 5,561 hate speech and will be split to train-test-validate of 60%-20%-20%. We have tried the experimental process. It's using LSTM with 50 hidden layers. The model also uses Adam for optimizer, binary crossentropy for the loss function, accuracy for the metrics, and sigmoid as activation function. Table 4 shows the last 5 trained data of 10 epoch and 1000 batch size.

TABLE IV
LIST OF EPOCHS

| Epochs | Accuracy |
|---|---|
| Epoch 6/10 | 0.7449 |
| Epoch 7/10 | 0.7424 |
| Epoch 8/10 | 0.7665 |
| Epoch 9/10 | 0.7783 |
| Epoch 10/10 | 0.7845 |

The experiment will be tuned or try different data pre-processing method and feature extraction method to get maximum results of accuracy.

Hate speech content is growing online daily. This is a growing problem which needs to be addressed quickly. There are various online forums to report hate speech supervised by Humans. It is a slow process. Alternatively, using Artificial intelligence to address these issues seems promising. There has been a lot of work done in this field and a lot is going on. Previous works have come up with several revolutionary ideas to tackle the problem of online hate speech filtering. We analyzed different approaches using Machine Learning to classify the input from various platforms either hate speech or as general text.

Based on this research, Support Vector Machine is the most used machine learning method to deal with hate speech detection. Although it is the most used, it is not quite the most effective. Our research shows that Long Short-Term Memory is by far the most effective method for achieving best results. This may be caused by difference of dataset used for the study. This research also showed that imbalanced dataset is still widely used for hate speech detection. But this can be solved by data re-sampling technique to balance the dataset by deleting duplicates of major class data so the number of dataset could become more balanced. The average data used for hate speech detection in this research is around 10,000 to 20,000 data from different resources.

It is quite rare to see a research paper done for hate speech and abusive language detection in Indonesian language with LSTM method. The problem of speech without any hate word is very common in datasets. Such type of classification requires deep understanding of language and grammatical knowledge. Hence more work is required so that our model can improve semantic paradigms and distinguish between them. Ironical sentences pose a very different situation, they actually mean the opposite of the actual wordings.

## REFERENCES


[1] Pitsilis GK, Ramampiaro H, Langseth H. Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence. 2018 Dec 1;48(12):4730-42.

[2] Saleem HM, Dillon KP, Benesch S, Ruths D. A web of hate: Tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159. 2017 Sep 28.

[3] Allan J. The Harm in Hate Speech. Constitutional Commentary. 2013 Jun 22;29(1):59-80.

[4] van Aken B, Risch J, Krestel R, Löser A. Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572. 2018 Sep 20.

[5] Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing. InProceedings of the Fifth International Workshop on Natural Language Processing for Social Media 2017 Apr (pp. 1-10).

[6] Kumar A, Jaiswal A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. Concurrency and Computation: Practice and Experience. 2020 Jan 10;32(1):e5107.

[7] Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation. 2020 Sep 30:1-47.

[8] Anam MC, Hafiz M. Surat Edaran Kapolri Tentang Penanganan Ujaran Kebencian (Hate Speech) dalam Kerangka Hak Asasi Manusia. Jurnal Keamanan Nasional. 2015 Dec 28;1(3):341-64.

[9] Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. InProceedings of the 25th international conference on world wide web 2016 Apr 11 (pp. 145-153).

[10] Hayaty M, Adi S, Hartanto AD. Lexicon-Based Indonesian Local Language Abusive Words Dictionary to Detect Hate Speech in Social Media. Journal of Information Systems Engineering and Business Intelligence. 2020 Apr 27;6(1):9-17.

[11] Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N. Hate speech detection with comment embeddings. InProceedings of the 24th international conference on world wide web 2015 May 18 (pp. 29-30).

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

[13] Gröndahl T, Pajola L, Juuti M, Conti M, Asokan N. All You Need is" Love" Evading Hate Speech Detection. InProceedings of the 11th ACM Workshop on Artificial Intelligence and Security 2018 Jan 15 (pp. 2-12).

[14] Pelle R, Alcântara C, Moreira VP. A classifier ensemble for offensive text detection. InProceedings of the 24th Brazilian Symposium on Multimedia and the Web 2018 Oct 16 (pp. 237-243).

[15] Aulia N, Budi I. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. InProceedings of the 2019 5th International Conference on Computing and Artificial Intelligence 2019 Apr 19 (pp. 164-169).

[16] Dorris W, Hu R, Vishwamitra N, Luo F, Costello M. Towards Automatic Detection and Explanation of Hate Speech and Offensive Language. InProceedings of the Sixth International Workshop on Security and Privacy Analytics 2020 Mar 16 (pp. 23-29).

[17] Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT). 2020 Mar 14;20(2):1-22.

[18] Watanabe H, Bouazizi M, Ohtsuki T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access. 2018 Feb 15;6:13825-35.

[19] Oriola O, Kotzé E. Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. IEEE Access. 2020 Jan 20;8:21496-509.

[20] Akhter MP, Jiangbin Z, Naqvi IR, Abdelmajeed M, Sadiq MT. Automatic Detection of Offensive Language for Urdu and Roman Urdu. IEEE Access. 2020 May 15;8:91213-26.
[21] Ibrohim MO, Budi I. A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science. 2018 Jan 1;135:222-9.
[22] Sreelakshmi K, Premjith B, Soman KP. Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science. 2020 Jan 1;171:737-44.
[23] Burnap P, Williams ML. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data science. 2016 Dec 1;5(1):11.
[24] Miró-Llinares F, Moneva A, Esteve M. Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Science. 2018 Dec 1;7(1):15.
[25] Salminen J, Hopf M, Chowdhury SA, Jung SG, Almerekhi H, Jansen BJ. Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences. 2020 Dec 1;10(1):1.
[26] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. InEleventh international aaai conference on web and social media 2017 May 3.
[27] Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. InProceedings of the NAACL student research workshop 2016 Jun (pp. 88-93).
[28] Ganganwar V. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering. 2012 Apr;2(4):42-7.
[29] Ibrohim MO, Budi I. Multi-label hate speech and abusive language detection in Indonesian twitter. InProceedings of the Third Workshop on Abusive Language Online 2019 Aug (pp. 46-57)..