Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 8286-8294

Identification of Islamophobia Sentiment Analysis on Twitter Using Text Mining Language Detection

Fachrul Kurniawan [1]*, Badruddin [2], Aji Prasetya Wibawa [3]

*[1] *Department of Informatics Engineering, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia.*
[2] Department of Islamic Law, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia.
[3] Department of Electrical Engineering, State University of Malang, Indonesia.
*Email: [1] * fachrulk@ti.uin-malang.ac.id*

Abstract

This research aims to discover the issue of sentiment analysis related to Islamophobia in social media, especially Twitter. The text mining and language detection approach was used to perform the data collection and selection. By identifying a text's polarity, sentiment analysis is a technique for extracting information from a person's attitude about an issue or occurrence. The grouping is made to discuss whether the reader is positive or negative. Moreover, the machine learning approaches also performed using long-short term memory (LSTM) and support vector machine (SVM) to classify the data. From the pre-processing stage, the drop duplication procedure creates 4339 from the preceding 10997, and the result language detection is 31 languages. Although the data comes from the world's largest Muslim country, the problem is not limited to it, as evidenced by using text mining tools to identify languages.

Keywords: social media, Islamophobia, language, text mining

I. INTRODUCTION

Social media utilization to have dialogues has expanded in recent years [1], [2]. This corresponds to the exponential growth of the internet, which continues to increase year after year. The concerns that arise in social media dialogues are diverse, ranging from positive to bad. Islamophobia is one of the essential concerns to consider. This problem develops due to religious conflict and the rapid advancement of digital technology [3]–[5].

Hatred of religion stems from violence committed solely to protect one's truth and is the result of advances in communication technology, which allows all information to be disseminated fast and without verification. Extremely prejudiced words are thought to influence communication amongst fellow believers. As the world's largest Muslim country, Indonesia is keenly interested in Islamophobia, which can have political and security implications [6], [7].

The number of social media comments, particularly on Twitter, is enormous. Positive and negative remarks are both acceptable. Islamophobia is one of the most often discussed harsh comments. Islamophobia is an outpouring of anti-Islamic sentiment manifested in various ways, including demonstrations, the passage of legislation forbidding the display of Islamic symbols, and the dissemination of unfavourable views on social media [8], [9].

The process of community communication through social media will be fascinating to conduct an in-depth study of the classification of conversation content and community patterns that lead to sentences about Islamophobia [10]. Sentiment analysis is a technique or method used to identify how expressed using text and how that sentiment can be classified as positive or negative. The results of the prototype system achieve high precision (75-95% depending on the data) in finding sentiments on web pages and news articles [11].

Data on social media conversations with hashtags related to hate speech that causes fear of specific ideas, categorized by the city where the conversation took place, will aid in understanding society's social model and character. Activities involving the interaction between cutting-edge internet-based technology and society's participation in the growth of numerous disciplines that have not been fully realized, particularly in the area of ethics, have had exceptional social consequences. This study will take a quantitative approach, which stresses in-depth data analysis.

According to the problem description above, we will describe pre-processing in text mining to parse the composition of the sorts of language that frequently arise to determine which communities from which locations typically debate Islamophobia in this article. This identification is critical as the initial step toward lowering the social attitude associated with hating speech in community debates. It is intended that a community will emerge to participate in the forums identified to better comprehend the genuine teachings of religion and lessen the hate speech community on social media.

II. RELATED WORK

Islamophobia has become a hot topic in various circles, posing a challenge to individuals all around the world. This has an impact on many Muslims, particularly those who are minorities in specific locations. Western intellectuals invented "Islamophobia" to denote anti-Islamic sentiment and prejudice [12]. Islamophobia is a shorthand term for fear or mockery of the Islamic religion and the hostility of most or all Muslims [13].

When connected to the internet, social media is a tool that can be used to interact or communicate online [14]–[16]. Social media is frequently used to form social relationships with others and share personal activities or real-life experiences. One element fueling social media's development is its high mobility and ease of use. In real-time, the community of social media users provides data in a wide range of unstructured formats and languages, as well as thoughts and attitudes [17]. One of the topics in data mining known as text mining is the availability of vast and unstructured data. This strategy is well-organized.

Social media platforms are very diverse in type and type, allowing people to choose the community they want. One of the platforms is Twitter, which provides several facilities for its users to interpret, convey, and share posts of up to 280 characters, better known as tweets. The platform is accessible via mobile devices, instant messaging, and website interface generating 326 million monthly active users [18]. By linking hashtags, users can quickly share any information when they want to search for information. The Twitter social network is included in the speedy category in terms of information exchange due to its easy use and high mobility [19], [20].

The pre-processing stage involves determining the data's quality before it is processed using specialized algorithms to categorize, classify, or visualize as required. This stage is also critical because it determines the data quality to be used. Some of the processes are governed by rules that the researchers specify. Imperfect data, data interference, and inconsistent data can all be avoided by pre-processing. Pre-processing is critical in sentiment analysis, particularly in social media, where informal and unstructured words or sentences abound and a lot of noise [21]–[23].

The following are the pre-processing stages often carried out: case folding, punctuation removal, tokenizing, and stop words removal. Case folding is a way to convert data in the same font size into lowercase letters. Punctuation removal is the stage to remove punctuation marks, numbers, links, and others. There are punctuation marks in some conversation data, such as periods (.), commas (,), and a link. It is not necessary, so it needs to be removed. The process of dividing sentences into words and forming word vectors is a tokenizing process. Eliminating irrelevant words reduces the repetition of words to give rise to unconnected opinions [24].

III. THE PURPOSE METHODS

The overall scheme research activity in this study includes five pre-processing features and a selection and visualization detecting language. Figure 1 presents the scheme activity of this research.

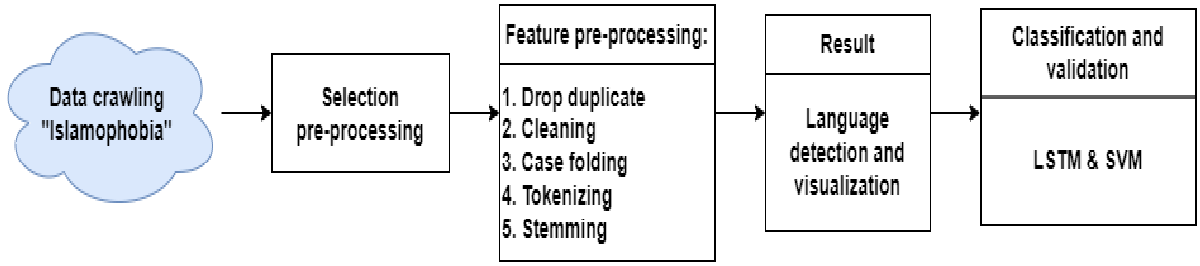


Figure 1. *The scheme activity of the proposed research*

IV. DATA CRAWLING

In this study, the data came from social media, which became a place of reference for community conversations and especially in religious or belief communities. The data comes from retrieval that comes from the Islamophobic hashtag on a social media platform called Twitter. By using the crawling technique carried out through the language.

In this case, we are using Tweepy and need to sign up for a developer account first. Tweepy is an easy-to-use Python library for accessing the Twitter API. This option is open to the public, so anyone can collect the data text from Twitter. We also need to create a project and answer all the questions. In the implementation, it required to install of the tweepy module and import the required modules.

Furthermore, in the implement utility functions to get tweets, there are three main functions: authentication function, client function in order to interact with the Twitter API, and final function that collects tweets and creates a data frame containing specific information about a given Twitter account. Moreover, the information collected includes (1) the date the tweet was created, (2) the author of the tweet, (3) the screen name corresponding to the name on the Twitter account, (4) the number of likes obtained by that tweet, and (5) number of retweets from the tweet

V. SELECTION AND LANGUAGE DETECTION

The pre-processing selection process starts by identifying or determining features by considering the interests of the desired research target. The selection uses five stages: drop duplicate, cleaning, case folding, tokenizing, and stemming. The determination of the features used is the researcher's choice according to the level of importance of the data sorting process.

This study resulted in detecting the number of languages that often appear and are used in conversations which detected the number of languages that are most widely used. This language detection will use a python-based algorithm, which produces several frequently used languages and visualizations in graphical form. Figure 2 presents the main stages in this language detection step.

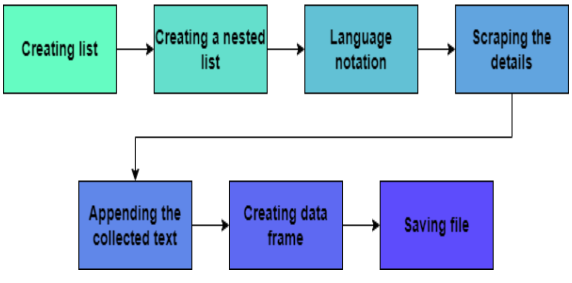


Figure 2. *Language detection stages*

We utilize the Language Detection dataset, which contains textual information for 33 languages. In the first phase, a list for each language is compiled. The list served as a repository for textual data for each language. Then, a nested list must be created to contain a second member that references the language name. This step will be used to store a sub-list along with other data types. The following step is to construct a notation for each language in the Twitter search. The symbol represents a language code, such as "en" for English, "fr" for French, "id" for Indonesian, etc.

In addition, the following step is scraping Twitter for "Islamophobia"-related information. This phase begins by obtaining information from Twitter using the Islamophobia keyword. The gathered text will be analyzed for language detection using natural language processing based on the dataset. Next, the discovered data will be added to the list of detected data based on the language. In addition, the data will be merged within the data frame for each language. And the last step is to save the data to a prepared.csv file.

VI. CLASSIFICATION AND VALIDATION

Moreover, a machine learning method was used to validate and classify the data using long-short term memory (LSTM) and support vector machine (SVM). LSTM using three gates including input gates, forget gates, and output gates. In LSTM, a single gate component controls the information entered the memory, which is tasked with resolving the gradient loss and division problem. Repeated connections provide the network with state or memory, enabling it to take use of ordered observations. Internal memory signifies network output that is dependent on the most recent context in the input queue, as opposed to input that is displayed as a network. While SVM operates by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not linearly separable. A divider is identified between the categories, and then the data are processed so that the divider can be depicted as a hyperplane.

VII. RESULT AND DISCUSSION

The data used in this study are social media users' opinions with the hashtag Islamophobia in 2021. The data taken is only in the form of tweets in all languages. Furthermore, the following is the dataset that has been collected, shown in Table 1.

Table 1. *Dataset crawling Islamophobia in Indonesian*

No.	Tweet
1	rintangan terbesar tangani islamofobia adalah perbedaan definisi dan konteks https://t.co/5c4chyv9zd
2	islamofobia di Eropa diduga disuburkan oleh golongan elite ini. https://t.co/h5yquetakc
3	beberapa muslim inggris mengalami rasialisme, islamofobia, dan pengucilan. https://t.co/7id5wncsge
4	islam agama masa depan, agama rahmat bagi semesta alam dan agama sejalan dgn peradaban dunia yg manusiawi, muhammad saw adalah role model /reladan karena salah kaprah itu ia pun bersemangat untuk mengampanyekan islamofobia melalui akunakun media sosial - https://t.co/bzetbkxerc

The experimental process was carried out, starting with dropping duplicates and deleting data detected as word duplication. The data used is 10997 tweet data, and after it is done, drop duplicate data to 4339 data. Then do the deletion of duplicate data first. The data used is 10,997 tweet data; after dropping duplicates, it becomes 4339. After the data has been determined, the next step is the language detection process. This process is carried out for labeling and calculating the number of languages used in tweets. This process found that the most widely used language was English with 3854 tweets, followed by Indonesian with 143 tweets and 31 tweets from other languages.

The cleaning process carried out in the pre-processing is deleting emoticons/characters, URLs, mentions, hashtags, numbers, excess spaces, retweets, punctuation marks, and enter. It is the process of deleting symbols and characters contained in the data used in this study. The next step is tokenizing, which is the process of making tweet sentences per word. The last process is stemming, which is removing affixes, both prefixes, infixes, and suffixes. Moreover, the following is shown in Table 2 and 3, the selection process for the pre-processing feature.

The selection process of the pre-processing feature shows a very significant result where all the data, which initially amounted to 10997, turned into 4339 data. Due to make the experiment more specific the language used in this research was Indonesian. Moreover, feature selection is selected and adjusted to make the data better and more accurate.

The pre-processing resulting in text classification-ready data for the LSTM and SVM classification algorithms. Each algorithm's classification evaluation produces numbers for accuracy, precision, recall, and f1-score using a confusion matrix. Each algorithm has its own unique testing scenario. Based on the number of layers and the number of neurons, Table 4 lists 21 possible LSTM implementations. As shown in Table 5, SVM has three possible configurations dependent on the kernel type used.

Table 4. *LSTM testing scenario*

Scenario	Layer	Neuron
1-7	1	(5), (10), (25), (50), (100), (150), (500)
8-14	2	(5, 5), (10, 10), (25, 25), (50, 50), (100, 100), (150, 150), (500, 500)
15-21	3	(5, 5, 5), (10, 10, 10), (25, 25, 25), (50, 50, 50), (100, 100, 100), (150, 150, 150), (500, 500, 500)

Table 5. *SVM testing scenario*

Scenario	Kernel
1	RBF
2	Linear
3	Polynomial

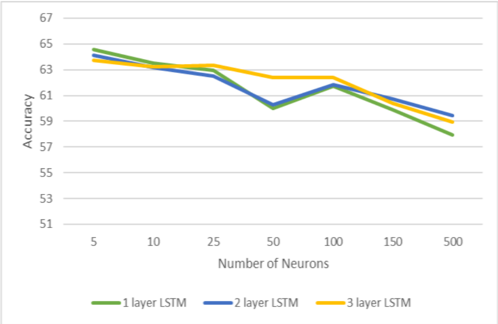
Five classifications of each situation using each algorithm were performed. Table 6 displays the results of the classification evaluation for the LSTM method, whereas Table 7 displays the results for the SVM algorithm.

Table 6. *Evaluation of LSTM classification*

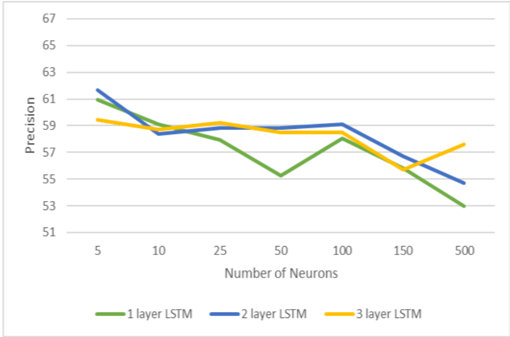
Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
1	64.57	60.96	60.14	60.42
2	63.49	59.09	62.39	60.59
3	62.98	57.96	65.63	61.47
4	60.00	55.26	60.56	57.63
5	61.71	58.04	54.79	56.28
6	59.87	55.80	54.09	54.85
7	57.91	52.96	60.42	56.38
8	64.13	61.66	55.49	57.55
9	63.17	58.39	63.94	60.97
10	62.48	58.82	60.11	59.61
11	60.26	58.82	53.10	56.38
12	61.84	59.11	51.97	54.90
13	60.70	56.72	55.35	55.82
14	59.43	54.72	56.26	55.45
15	63.75	59.44	64.22	64.24
16	63.24	58.72	65.18	60.51
17	63.36	59.22	60.14	59.66
18	62.41	58.47	58.45	58.30
19	62.41	58.48	58.31	58.27
20	60.38	55.72	59.72	57.24
21	58.92	57.59	53.20	55.18

Based on Table 6, it is known that scenario 1 has the most accurate evaluation findings, at 64.57%. The parameter for scenario 1 is 3 LSTM layers with 5 neurons. The scenario with the lowest accuracy is 7, with a value of 57.91%. The parameter for Scenario 7 is 1 LSTM layer with 500 neurons. Figure 3 comprehensively compares all scenarios in terms of accuracy, precision, recall, and F1 score.

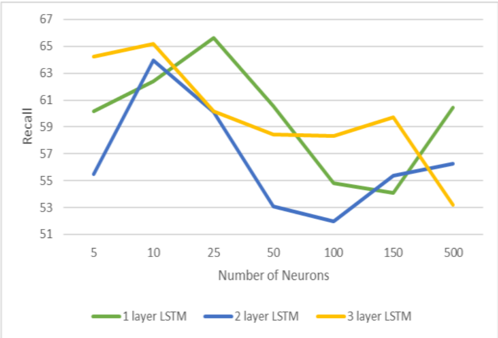
Journal of Positive School Psychology
2022, Vol. 6, No. 5, 8286-8294



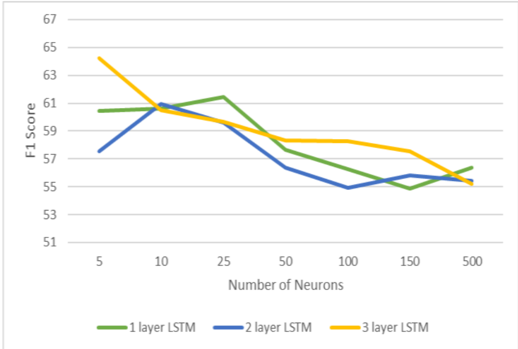
(a)



(b)



(c)



(d)

Figure 3. *Accuracy (a), Precision (b), Recall (c), F1 score (d)*

Based on the overall results shown in Figure 3, it was determined that the optimal number of neurons was five, paired with one LSTM layer. Regarding 2-layer and 3-layer LSTM, this is different. In two layers of LSTM, the number of best-performing neurons is five. In 3 layers of LSTM, the optimal results are equivalent to those of 5 neurons. For 1 layer, 2 layer, and 3 layer LSTM, the addition of 5 neurons diminishes accuracy and precision values. Therefore, 5 neurons are inadequate for text classification.

LSTM with three layers can achieve greater accuracy than LSTM with one layer and two layers, according to the study's findings. The usage of 500 neurons, does not improve accuracy; rather, it decreases it. 100 neurons produce the best performance on one LSTM layer, 25 neurons produce the highest accuracy on two LSTM layers, and 50 neurons produce the best accuracy on three LSTM layers. As long as the number of neurons utilized in text classification is not excessive, the resulting accuracy will be pretty high.

Table 7. *Evaluation of SVM classification*

Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
1	61.90	66.66	68.13	67.39
2	62.53	68.18	65.93	67.03
3	61.58	61.25	91.20	73.28

According to evaluation results, each kernel's SVM has distinct accuracy, precision, recall, and F1-score. The accuracy of the Linear kernel is 62.53%, which is higher than the accuracy of the other kernels, but the accuracy of the Polynomial kernel and the RBF kernel differ by only 0.32%. Regarding precision, the Linear kernel remains better to all others. In contrast, the Polynomial kernel has a higher value than other kernels due to a considerable difference of approximately 23.07% to 25.27%. In terms of the F1-score, the Polynomial kernel is superior to other kernels.

From the experimental results, some observations about the proposed approach are shown as follows. First, selecting the pre-processing feature can be carried out effectively where the time required to determine the pre-processing technique can be. Second, data taken from specific social media requires pre-processing selection according to interests. The selection of pre-processing features improves text mining-based data structuring performance to become more ready to be processed in any form. However, the performance of feature selection needs to be compared between each stage to avoid inefficiency in the process of each stage. In the end, crawling data originating from conversation forums on social media requires processing using a selection of pre-processing features because each platform requires different treatment. The behavior outlined in the discussion of Islamophobia by social media users does not necessarily come from the majority of countries that use religion as one of the parameters in every community activity.

VIII. CONCLUSION

This research shows that social media, especially Twitter, has become a commonly used platform for conversations about the issue of Islamophobia. This paper proposes selecting the pre-processing feature to control the gross crawling data and then measure it in a limited and short manner. The selection of features used in crawling data processing can optimize results that can be used to perform limited analysis. Research using pre-processing features with the hashtag Islamophobia identified 31 types of languages from various countries. Moreover, based on the results of the research that has been done, it can be concluded that the long-short term memory (LSTM) algorithm and support vector machine (SVM) can be used to classify the issue of Islamophobia. This is indicated by the resulting accuracy value of 73.797% for LSTM and 60.22% for SVM with Polynomial kernel. Identification uses text mining techniques with very maximum feature selection to find out the target desired by the researcher in a limited and fast manner. In the future, additional features will be produced in pre-processing in text mining to create high-quality data.

ACKNOWLEDGMENTS

This work was supported in part by the research institute and community service LP2M Universitas Islam Maulana Malik Ibrahim Malang 2021.

REFERENCES

1. M. Eriksson, "Lessons for Crisis Communication on Social Media: A Systematic Review of What Research Tells the Practice," Int. J. Strateg. Commun., vol. 12, no. 5, pp. 526–551, Oct. 2018, doi: 10.1080/1553118X.2018.1510405.
2. Y. Wang and Y. Yang, "Dialogic communication on social media: How organizations use Twitter to build dialogic relationships with their publics," Comput. Human Behav., vol. 104, p. 106183, Mar. 2020, doi: 10.1016/j.chb.2019.106183.
3. S. Eckert, J. Metzger-Riftkin, S. Kolhoff, and S. O'Shay-Wallace, "A hyper differential counterpublic : Muslim social media users and Islamophobia during the 2016 US presidential election," New Media Soc., vol. 23, no. 1, pp. 78–98, Jan. 2021, doi: 10.1177/1461444819892283.
4. B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," J. Inf. Technol. Polit., vol. 17, no. 1, pp. 66–78, Jan. 2020, doi: 10.1080/19331681.2019.1702607.
5. U. M. Hashmi, R. A. Rashid, and M. K. Ahmad, "The representation of Islam within social media: a systematic review," Information, Commun. Soc., vol. 24, no. 13, pp. 1962–1981, Oct. 2021, doi: 10.1080/1369118X.2020.1847165.
6. Syaza Farhana Mohamad Shukri, "The Perception of Indonesian Youths toward Islamophobia: An Exploratory Study," Islam. Stud. J., vol. 5, no. 1, p. 61, 2019, doi: 10.13169/islastudj.5.1.0061.
7. Z. Syarif, S. A. Mughni, and A. Hannan, "Post-truth and Islamophobia narration

in the contemporary Indonesian political constellation," Indones. J. Islam Muslim Soc., vol. 10, no. 2, pp. 199–225, Dec. 2020, doi: 10.18326/ijims.v10i2.199-225.

8. C. Allen, Islamophobia. London: Routledge, 2016.
9. S. Sayyid, "A Measure of Islamophobia," Islam. Stud. J., vol. 2, no. 1, p. 10, 2014, doi: 10.13169/islastudj.2.1.0010.
10. S. Rani and P. Kumar, "Deep Learning Based Sentiment Analysis Using Convolution Neural Network," Arab. J. Sci. Eng., vol. 44, no. 4, pp. 3305–3314, Apr. 2019, doi: 10.1007/s13369-018-3500-z.
11. T. Nasukawa and J. Yi, "Sentiment analysis," in Proceedings of the international conference on Knowledge capture - K-CAP '03, 2003, p. 70, doi: 10.1145/945645.945658.
12. M. J. Sirgy, M. Y. Kim, M. Joshanloo, and M. Bosnjak, "Is Subjective Ill-Being Related to Islamophobia in Germany? In Search for Moderators," J. Happiness Stud., vol. 20, no. 8, pp. 2655–2675, Dec. 2019, doi: 10.1007/s10902-018-0063-3.
13. J. M. Roose and B. S. Turner, "Islamophobia, Science and the Advocacy Concept," Society, vol. 56, no. 3, pp. 210–221, Jun. 2019, doi: 10.1007/s12115-019-00357-6.
14. S. Sendari, I. A. E. Zaeni, D. C. Lestari, and H. P. Hariyadi, "Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm," Knowl. Eng. Data Sci., vol. 3, no. 1, pp. 50–59, Aug. 2020, doi: 10.17977/um018v3i12020p50-59.
15. Y.-T. Huang and S.-F. Su, "Motives for Instagram Use and Topics of Interest among Young Adults," Futur. Internet, vol. 10, no. 8, p. 77, 2018, doi: 10.3390/fi10080077.
16. T. D. Baruah, "Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study," Int. J. Sci. Res. Publ., vol. 2, no. 1, pp. 1–10, 2012, doi: ISSN 2250-3153.
17. M. Hitesh, V. Vaibhav, Y. . A. Kalki, S. H. Kamtam, and S. Kumari, "Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model," in 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Sep. 2019, pp. 146–151, doi: 10.1109/ICCT46177.2019.8969049.
18. S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," IEEE Access, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.
19. Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," IEEE Access, vol. 6, pp. 23253–23260, 2018, doi: 10.1109/ACCESS.2017.2776930.
20. U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis," Futur. Gener. Comput. Syst., vol. 113, pp. 58–69, Dec. 2020, doi: 10.1016/j.future.2020.06.050.
21. J. Camacho-Collados and M. T. Pilehvar, "On the Role of Text Pre-processing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.01780.
22. S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," in 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Oct. 2019, pp. 1–8, doi: 10.1109/KSE.2019.8919368.

23. S. Alam and N. Yao, "The impact of pre-processing steps on the accuracy of machine learning algorithms in sentiment analysis," Comput. Math. Organ. Theory, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.
24. K. Moolthaisong and W. Songpan, "Emotion Analysis and Classification of Movie Reviews Using Data Mining," in 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Jul. 2020, pp. 89–92, doi: 10.1109/DATABIA50434.2020.9190363.