American Journal of Data Mining and Knowledge Discovery
2023; 8(1): 11-17
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20230801.12
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

Review Article

Machine Learning for Text Classification on Twitter: A Literature Review

Muneer Alsurori, Ahlam Enan*, Rahuf Alwan, Wafa Algumaei, Somia Alturki, Entsar Alkahtany

Information Technology Science, Ibb University, Ibb, Yemen

Email address:

msurory@yahoo.com (Muneer Alsurori), enanahlam92@gmail.com (Ahlam Enan), alwanrahuf@gmail.com (Rahuf Alwan), wafaali555@gmail.com (Wafa Algumaei), somiaalturki36@gmail.com (Somia Alturki), Entesir2000@gmail.com (Entsar Alkahtany)

*Corresponding author

To cite this article:

Muneer Alsurori, Ahlam Enan, Rahuf Alwan, Wafa Algumaei, Somia Alturki, Entsar Alkahtany. Machine Learning for Text Classification on Twitter: A Literature Review. *American Journal of Data Mining and Knowledge Discovery*. Vol. 8, No. 1, 2023, pp. 11-17.
doi: 10.11648/j.ajdmkd.20230801.12

Received: September 17, 2023; **Accepted**: October 16, 2023; **Published**: November 9, 2023

Abstract: This literature review examines the application of machine learning (ML) techniques for text classification on Twitter. With the immense volume of data generated on social media platforms like Twitter, there is a need for automated methods to extract valuable information. ML, known for its ability to learn patterns and relationships in large datasets, has gained significant attention in this context. The purpose of this review is to explore the background and aim of ML for text classification on Twitter, the methods employed, the results obtained, and the conclusions drawn. The review begins by discussing the background and aim, emphasizing the vast amount of data available on Twitter and the need for automated techniques to extract useful information from this data. It highlights the significance of ML in addressing this challenge, particularly in tasks such as sentiment analysis, topic modeling, and spam detection, which play a crucial role in social media analysis. Next, the review provides an overview of the methods used in various studies on text classification using Twitter data. It explores the latest approaches and techniques employed in ML, including feature extraction methods like bag-of-words, n-grams, and word embeddings. It also discusses the preprocessing steps involved in preparing Twitter data for classification tasks. subsequently, the review presents the results obtained from different studies in the field. It discusses the performance metrics used to evaluate the effectiveness of ML models, highlighting measures such as accuracy, precision, recall, and F1-score. The review also discusses variations in performance across different classification tasks, providing insights into the strengths and limitations of the approaches used.

Keywords: Machine Learning, Text Classification, Twitter Data, NLP

1. Introduction

Twitter is one of the most broadly utilized social media platforms where people share their opinions, express their feelings, and exchange information on various topics. With over 330 million monthly active users, Twitter generates a massive amount of data, including text messages, images, and video. [1]. This vast amount of data has created a need to develop automated methods for data analysis to extract meaningful information from Twitter data. One such method is machine learning (ML) for text classification. ML has gained much consideration within the last few years due to its capability to automatically learn patterns and relationships in large datasets. [2]. Text classification is one of the prominent errands in Natural language processing (NLP), which aims to automatically categorize text documents into predefined categories or labels [3]. Text classification has various applications in social media analysis, such as sentiment analysis [4], topic modeling [5], user profiling [6], and spam detection [7]. The use of Twitter data in ML research has emerged as a potential domain in recent years, and the interest in Twitter.

2. Methodology

This study reviews numerous papers that were published in English between 2019 and 2023. The search results were then filtered according to their applicability and value to the field. Studies focusing on text classification on Twitter data using machine learning and natural language processing methods were the inclusion criteria for this literature review. These criteria were not met by studies that weren't published in English or that didn't match the other requirements. The chosen papers were then examined and combined to offer a summary of the state-of-the-art text categorization methods applied to Twitter data. The investigation focused on the performance indicators presented, the machine learning and natural language processing algorithms used, and the restrictions and gaps in the literature.

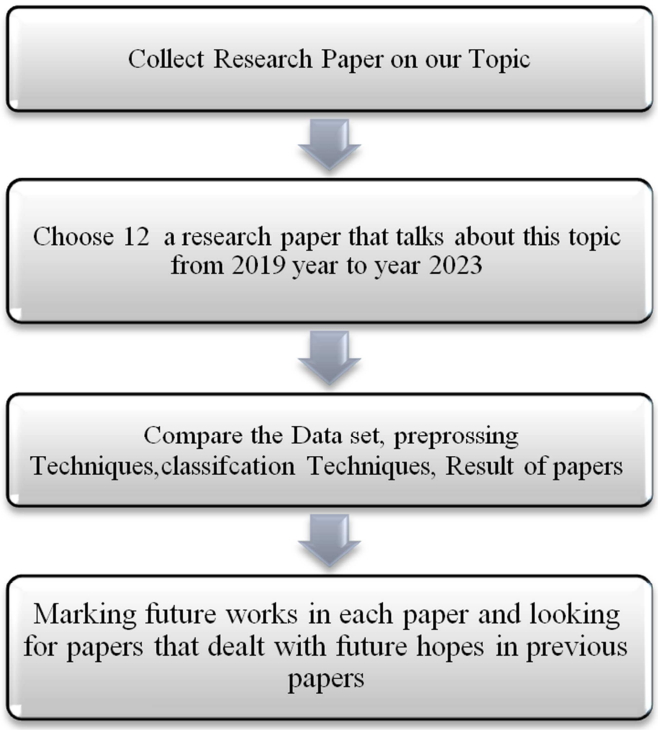

Figure 1. *The methodology.*

3. Comparison Table

A comparison table can be a useful tool in the literature review on machine learning for text classification on Twitter to enumerate and contrast the salient features, approaches, and outcomes of the chosen studies. The comparison table enables an ordered presentation of the data, allowing readers to quickly spot patterns, trends, and similarities among the examined studies. Table 1 reviews and compares a number of research papers. The table includes a comparison in terms of: dataset, preprocessing techniques, classification techniques, results, weaknesses and future work.

Table 1. *Comparison of Selected Studies on Machine Learning for Text Classification on Twitter.*

Paper	Data set	Preprocessing techniques	Classification techniques	Results	Weaknesses	Future work
[8]	The author used 1000 datasets from tweet	The author used tokenization, stemming, lemmatization, removal of stopwords, Part-of-speech (POS) tagging, labeling, named substance acknowledgment, recognition, co- reference determination, and content modelling as sack of Word and, inverse document frequency (TF IDF) Model.	support vector machine (SVM) maximum entropy naive bayes algorithm and k-nearest neighbor classifier	They found that by misusing the TF-IDF vector, the precision of assumption investigation might be altogether progressed, an accuracy of 85.25% was achieved in estimation utilizing the NLP strategy.	The author used small dataset	Apply the suggested method to another data set.

Paper	Data set	Preprocessing techniques	Classification techniques	Results	Weaknesses	Future work
[9]	The tests used data from two datasets: The first is the customer surveys about movies from the IMDB, which have been labeled by Kotizas, and the second is the Twitter tweets, counting the customer tweets about health in English in 2019 that have been gathered using the Twitter API.	the auther used Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec (W2V) modeling techniques for feature extraction..	naïve Bayes (NB), support Vector Machines (SVM) and Artificial Neural network (ANN) algorithms	Agreeing to the When the test came about, manufactured neural arrays had the best exact execution in both datasets compared to the others. What comes about with w2v on Twitter is ANN 0,87 and ANN 0,90 on IMDB datasets.	The preprocessing steps that were used are not mentioned.	include looking into Turkish tweets in addition to English. In addition, analysis of opinions will be done using data gathered from several websites and social media platforms, besides Twitter, where individuals express their opinions. Additionally, in the future, classifier models will be created using sophisticated learning algorithms, imitating conventional word insertion techniques such as Bert for Showing Content.
[10]	World wide dataset consisting of 37,373 interesting tweets from Twitter.	Evacuate and clean up undesirable commotion in the content location. For illustration, halt words, extraordinary characters, and rehashed words were expelled. At that point, the stemming for the remaining words to their unique roots has been connected as a result of this preprocessing. IF-IDF and word 2vec techniques for feature extraction	Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), and Naive Bayes (NB), and vector machine support (SVM).	The experimental results revealed the predominance of LR, which achieved a normal exactness of approximately 90.57%. Among the classifiers, calculated relapse accomplished the leading F1 scrose (0.928), SGD accomplished the finest exactness (0,968), and SVM accomplished the leading review (1.00).	This study is limited to the English language, and the size of the dataset is modest.	One development is the combination and testing of distinctive extraction, which improves the discovery rate of both the LR and SGD classifiers. We are also developing a real-time cyberbully detection stage, which will be useful for quickly identifying and avoiding the cyberbully.
[11]	Existing datasets have been used; the main one is from Stanford University's "Sentiment140," which has 1.6 million tweets, and the other one originated from Crowdflower's Information for Everyone Library, which contains 13870 sections.	The authors removed URLs, hashtags, and usernames; reduced all capitalized letters to lowercase; switched to a common dialect using dialect interpreter work; categorized as a piece of speech (POS); extracted data from HTML and XML records; used a spell checker; and tokenized tweets.	MNB LR SVM recurrent neural network (RNN) and LSTM (Long Short-Term Memory)	Recurrent Neural Network With LSTM 82% on first data set Support Vector 68.90% In second data set	The author did not specify the method by which the features were extracted	Using deep learning models to improve accuracy
[12]	The data set consists of 18,000 tweets	a) Emojis have been supplanted with significant emotional content. b) Accentuation images are expelled from tweets. c) Halt words are expelled from Tweets. 4) Stemming is performed to expel the esteemof the word from the root of the word. c) "Slang words" are transformed into words of equivalent meaning	Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbour and Logistic Regression	Logistic Regression with most noteworthy precision rate of 86.51	One of the weaknesses in this study is the size of the dataset	Execuation comparison of Classifiers on Twitter wistful investigation
[13]	The data sets that were gathered for	The content has been modified, including	k-nearest Neighbors (KNN)	Common individuals have a	Using only one classifer	They can apply another classification technique

Paper	Data set	Preprocessing techniques	Classification techniques	Results	Weaknesses	Future work
	each hashtag are focused on #Pfizer, #Moderna, and #AstraZeneca. For every hashtag, 10,000 tweets are received. based on the tweets	lowercase text, halt words, compressions, and a custom task for withdrawal replacement. Spelling checks are conducted to correct misspelled words. The emoji is replaced with "smiley" and information is tokenized, normalized, and lemmatized before being transferred to Object.		higher positive estimation of Pfizer and Moderna immuization with a rate of 47.29 and 46.16, respectively, compared to AstraZeneca immuization with a rate of 40.08.		and collect more than 30,000 tweets.
[14]	The information collected through Twitter API is 8000 tweet. then removed from duplicate and unrelated data. After data cleansing, there are 1038 relevant tweet data.	The author use stemming, stop word removal, and tokenizing, Doc2Vec	SVM and Logistic Regression.	The result of PV-DBOW with SVM, PV-DM with SVM, and calculated relapse has the highest level of precision and F1-score compared to another shoew. The finest result appears to be precision at around 87% and an F1 score at around 81%.	Small data set	In this inquiry, there are still numerous holes to be made in. One of them is how to confront the challenges as depicted within the assumption investigation segment. Future investigations ought to also ensure that they incorporate an adjusted dataset between names. In expansion, assist investigate ought to attempt to classify by the subhect, as it were by the opinion.
[15]	The author's chosen the WikiText-103 dataset for our source assignment, which comprises 28,595 preprocessed Wikipedia articles with a substance of up to 103 million words. For estimation examination, we chose the Twitter US Aircraft Assumption t dataset. The dataset contains 14,485 tweets with respect to the most-worked US carriers.	The authors isolated the content into areas based on expressions, words, images, and other critical perspectives coming about in a list of particular words for each comment. Erase halt words that have relation words and words that do not give any emotion.	The authors used a combination of two classification techniques: Widespread Dialect Demonstrate Fine-tuning (ULMFiT) and Support Vector Machines (SVM). They first finetuned a pretrained language model (ULMFiT) on their Twitter dataset to obtain tweet embeddings. They then fed the embeddings to an SVM classifier to predict the sentiment label of each tweet.	He demonstrate illustrate an precision rate of 99.78% on Twitter USCarriers, 99.71% on IMDB, and 95.78% GOP talk about.	The opinion inquiry was conducted to archival level in this case. We did not address the assumption at the viewpoint level in our investigation.	The authors suggest several directions for future work, including exploring the effectiveness of their approach on other social media platforms and languages, investigating the effects of diverse hyperparameters on the execution of their approach, and integrating other features, such as emojis and images, into their model. They also suggest applying their approach to other natural language processing tasks, such as named entity recognition and text classification.
[16]	Twitter information was extracted from utilizing its API, contains Thirty thousand dataset, which contains 11000 are related to despise discourse and are in like manner labeled to a specific course.	The tweets are cleaned up and made ready for analysis by taking out URLs, mentions, hashtags, stop words, and punctuation, as well as stemming and lemmatizing them.	The authors used several machine learning algorithms such as Decision Tree (DT), Stochastic Gradient Boosting classifier detect hate speech in the tweets.	The authors achieved an accuracy of 97% using the Decision Tree classifier and with Stochastic Gradient Boosting classifier achieved 98.04	The dataset used was limited to tweets collected during the COVID-19 pandemic, which may not be representative of hate speech in general. Also, the authors did not evaluate the	Within the future, abhor discourse may be categorized based on sexual orientation. Long Brife Term Memory (LSTM) and Convolutional Neural Network (CNN) may also be utilized before long for performing multi-class classification.

Paper	Data set	Preprocessing techniques	Classification techniques	Results	Weaknesses	Future work
					execution of the show on other datasets.	
[17]	The data collect using API which contains 28,264 tweets	The normal dialect toolkit (NLTK) tokenization is used to begin with to tokenize sentences. String substitution is utilized to normalize whitespace and extract hypertext markup dialect (HTML) labels. Each word changed into lower content some time recently.	Using hybrid method used to detect recent social issues	identifying rate 89%, 95%, 83%, 53%, and 98% for the top 5 identified crisis	study only focuses on English-language tweets, which may limit its applicability to other languages.	Due to changes in the organizational structure of Twitter, there might be an effect on users. In the future, we may work on the social crisis during the COVID-19 period by using Twitter data and finding their probable intentions. A combined model of machine learning and non-machine learning models can be applied to the further development of this work.
[18]	The dataset from twitter that collected utilizing watchord "PeduliLindungi" appeard 51740 twitter comments.	Preprocessing involves removing irrelevant information or converting it into a frame that the framework can prepare more easily. Case collapsing is the first step, followed by cleaning up the username, hashtag, url, accent, and image attributes. Tokenizing, which divides each word in a sentence into individual word units, normalization, which turns erratic words into regular words, filtering, which eliminates words that frequently appear without meaning, and stemming, which replaces suffixed words.	Naïve Bayes algorithm	The exactness gotten is 95.86%, with exactness 96.99% and recall 94.12%.	This think about restricted to Twitter media social as the data processed with a non-formal dialect that is required to paraphrase the word to urge a great result for the modeling of machine learning. And the information as it was taken from June until December 2021, the information collection can be expanded once more by taking into account the following period.	The extra information can be gathered from other social media sites like Instagram, TikTok, Facebook, or Webpage for collecting and analyzing audits. It is imperative to know more about sentiment with respect to the employment of Peduli -Lindungi application, so it will offer assistance to designers to understand their application way better, which the survey client at Android or iOS, in some cases, predispositions
[19]	The datasets collect using SNSCRAPE API which contains 11,250 tweets approximately the war between Russia and Ukraine from his Twitter account.	Prior to applying several text normalization approaches, such as stopword removal, stemming, and lowercase conversion, they eliminated any non-English tweets. Additionally, they performed sentiment analysis on the dataset using the TextBlob package, eliminating any neutral tweets.	The author perform several ML: The extra trees classifier (ETC) (LR), (DT), (SVM), (GNB), & (KNN)	The extra trees classifier (ETC) demonstrated a most noteworthy exactness of 0.84.	The datasets is limited related to the Russia-Ukraine war	To better understand the machine learning models' efficacy in sentiment analysis, investigate how they perform on diverse datasets connected to various themes and events.

4. Literature Review

In this section, the writing review of the selected papers will be presented. The researchers contributed to the field of Twitter data classification and made several comparisons of the most common methods.

The authors of this paper provide a pre-processed data system based on natural language processing (NLP) to filter tweets and analyze public opinions towards a product using sentiment analysis. They classify positive and negative tweets using the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model principles, with an accuracy of 85.25%. [8] The authors of this paper discuss sentiment analysis on social media using machine learning methods. The authors use two datasets for sentiment analysis: 4500 health-related Twitter data was collected using the Twitter API, with 1680 neutral, 1220 positive, and 1600 negative tweets. 500 positive and 500 negative opinions were collected from IMDB movie reviews. The author evaluates the

performance of popular machine learning classifier algorithms such as SVM, ANN, and NB in comparison with traditional frequency-based text representation (TF-IDF) and prediction-based text representation (W2V) methods. [9] The authors of this paper discuss the issue of cyberbullying on social media platforms, particularly Twitter, and highlight the need for detection, prevention, and mitigation strategies. This study contributed to revealing bullying without involving the victims. Seven different classifiers were compared. [10] The authors of this paper presented a comparison of several different workbooks on two different data pools. The first data set is From Stanford universitys comprising or 1.6 million tweets and the other initially came from 'Crowdflowers' information for Everyone library comprising of entries. Textblob, Sentiwordnet, MNB, LR, SVM and RNN Classifier performance comparisons were performed. An aggregated model of MNB, LR and SVM on datasets. LSTM outperformed the first dataset while SVM outperformed the other dataset. [11] The authors of this paper presented a comparison of the execution of basic classification methods such as Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbor, and Logistic Regression in analyzing tweets. the accuracy rates of these classifiers, with Logistic Regression achieving the highest accuracy rate of 86.51% and K-Nearest Neighbour performing the worst with an average accuracy rate of 50.40%. [12] The authors of this paper provided an assessment of people's opinions about vaccines from Pfizer, Moderna, and AstraZeneca. These tweets have been extracted from Twitter using the Twitter All verification token. The crude tweets were put away and prepared using NLP. The handled information was at that point classified utilizing the administered KNN classification algorithm. The calculation classifies the information into three categories: positive, negative, and unbiased. [13] The authors of this paper aim to classify the sentiment of tweets into three categories: pro, contra, and neutral, and the results show that almost all sentiments are against the development of Rinca Island. The paper uses two Doc2Vec models, the distributed model and the distributed bag of words, along with support vector machines (SVM) and logistic regression as classifiers. Each combination of the models and classifiers achieves an accuracy rate above 75%. [14]

The authors of this paper present another successful strategy for estimation investigation by combing all-inclusive dialect show fine tuning (ULMFIT) with back location proficiency and precision. The strategy presented over and over again is an approach for Twitter to discover people's groups states of mind towards certain items based on their comments. The overall results on three data sets showed that the model achieved the latest results in all data sets. [15] The authors of this paper used several machine learning algorithms such as Decision Tree, Stochastic Gradient Boosting classifier detect hate speech in the tweets. accuracy of 97% using the Decision Tree classifier and with Stochastic Gradient Boosting classifier achieved 98.04detecting hate speech related to COVID-19. [16] The authors of this paper propose a machine learning approach for social crisis detection using Twitter-based text mining. The authors use a dataset of tweets related to social crises, including natural disasters and political events, and train a logistic regression model with different features, including word embeddings and text length. The results indicate that the proposed approach achieves high accuracy in social crisis detection. The study demonstrates the potential of social media as an important source of data for crisis detection and response. [17].

The authors of this paper focus on analyzing the sentiments of societies towards the PeduliLindungi application using Twitter data. The data collection was done from June to December 2021, during a period of high COVID-19 cases and tighter movement restrictions imposed by the government. The sentiment analysis was performed using the Naïve Bayes algorithm. The results of the sentiment analysis showed that 64.69% of the sentiments were positive, indicating the pro expression from society, while 35.5% were negative, indicating the cons expression related to the performance and data security of the PeduliLindungi application. [18]. The authors of this paper provide a comprehensive assessment of administered machine learning models for sentiment analysis using Twitter information on the Russia-Ukraine war. The authors make a significant contribution to the field of sentiment analysis by providing a detailed comparison of several machine learning models and their performance on a specific dataset of 1.2 million unique English-language tweets. The additional trees classifier (ETC) model achieves the highest exactness of 0.84. [19]

5. Conclusion and Future Work

In conclusion, this literature review sheds light on the application of machine learning to Twitter text classification. The research discovered a wide range of methods and strategies for text classification, including decision trees, SVM models, Navibayes, and others. Twitter has been shown to create particular difficulties as a result of user-generated content traits, including slang, sarcasm, and misspellings. Despite these difficulties, machine learning has demonstrated its efficacy in determining sentiment, subject, and user intent on Twitter. The review also highlighted the significance of feature selection, engineering, data pre-processing, and model evaluation in attaining accurate text classification. The necessity for ongoing development and modification of machine learning techniques has also been shown to be necessary to deal with the continually changing nature of Twitter data. The effectiveness of various machine learning and deep learning algorithms and techniques for particular text categorization tasks on Twitter, such as recognizing fake news, hate speech, or political sentiments, could be explored in future studies in this field. Future research can also examine the generalizability of models developed using Twitter data to other social media sites with comparable features, including Instagram, among others. A review of research publications with a focus on the categorization of text in material written in the Arabic dialect is another task for the future.

References

[1] Statista. (2021). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2021 (in millions). Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521 (7553), 436-444.

[3] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34 (1), 1-47.

[4] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2 (1-2), 1-135.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3 (Jan), 993-1022.

[6] Cormack, G., & Lynam, T. (2007). Spam Filtering: A review. Foundations and Trends in Information Retrieval, 1 (4), 267-349.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019, July). Sentiment analysis with NLP on Twitter data. In 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2) (pp. 1-4). IEEE.

[9] Basarslan, M. S., & Kayaalp, F. (2020). Sentiment analysis with machine learning methods on social media.

[10] Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 12 (11), 187.

[11] Harjule, P., Gurjar, A., Seth, H., & Thakur, P. (2020, February). Text classification on Twitter data. In 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE) (pp. 160-164). IEEE.

[12] Wadhwa, S., & Babber, K. (2021). Performance comparison of classifiers on twitter sentimental analysis. European Journal of Engineering Science and Technology, 4 (3), 15-24.

[13] Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 23 (1), 463-470.

[14] Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Computer Science, 197, 660-667.

[15] AlBadani, B., Shi, R., & Dong, J. (2022). A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and SVM. Applied System Innovation, 5 (1).

[16] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights, 2 (2), 100120.

[17] Rahman, S., Jahan, N., Sadia, F., & Mahmud, I. (2023). Social crisis detection using Twitter based text mining-a machine learning approach. Bulletin of Electrical Engineering and Informatics, 12 (2), 1069-1077.

[18] Ellyanti, L., Ruldeviyani, Y., Pradana, L. E., & Harjanto, A. (2023). Sentiment Analysis of Twitter Users to the PeduliLindungi Using Naïve Bayes Algorithm. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7 (2), 414-421.

[19] Wadhwani, G. K., Varshney, P. K., Gupta, A., & Kumar, S. (2023). Sentiment Analysis and Comprehensive Evaluation of Supervised Machine Learning Models Using Twitter Data on Russia–Ukraine War. SN Computer Science, 4 (4), 346.