Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

A Performance Study of Applications of AI in Toxic Comments Classification

K. Pavithra[1], Dr. S. Rathi[2]

[1]PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, India

[2]Department of Computer Science and Engineering, Government College of Technology, Coimbatore, India

Email: [1]pavi.71772177104@gct.ac.in, [2]rathi@gct.ac.in

Abstract

Social media websites and tweeting apps have seen a sharp rise in popularity in the recent years. One can express their opinions and sentiments about things, people, and events through these platforms. Arguments frequently start on social media platforms during discussions and debates and involve the usage of toxic comments, which are unpleasant, disrespectful, and hurtful statements. According to many, social networking sites must be able to identify these harmful comments. This research analyses several deep learning and machine learning methods like Convolutional Neural Network, Long Short -Term Memory, Support Vector Machine, Random Forest, and Naive Bayes for toxic comments classification along with the study that examines the effects of many word embedding methods including Word2Vector, Bag of Words, Global Vectors, Bidirectional Encoder Representations from Transformers, and Embeddings from Language Model on the classification of toxic comments and also the future scope of the research.

Keywords: Toxic Comments, Machine Learning, Deep Learning, Word Embedding

1. Introduction

The use of social media platforms and microblogging websites for interpersonal and group communication has grown rapidly. Through these platforms, users can express their feelings through comments and criticism also takes place by sharing their thoughts, ideas, and opinions. As of January 2023, India had 692 million internet users. At the beginning of 2023,

DOI: https://doi.org/10.36548/jaicn.2023.2.002
Received: 28.03.2023, received in revised form: 28.04.2023, accepted: 13.05.2023, published: 23.05.2023

48.7 percent of the country's population was using the internet in India [1]. Similarly, 5.16 billion individuals utilise the internet globally, which is the equivalent of 64.4 percent of the world's population as of 2023 [2]. These people have a common space on social media platforms where they may express their viewpoints and engage in discussion. However, issues develop when social media platforms are used as the venue for arguments and disputes that involve the use of toxic comments. Online comments are fraught with dangers like toxicity, fake news, cyberbullying and online harassment. Unfortunately, users now experience a variety of psychological issues including sadness, dissatisfaction, and even suicidal thoughts as a result of these toxic comments. Toxic comments also harm the reputation of social media sites. Many people have been detained by the authorities in the past few years as a result of abusive or defamatory remarks on personal pages. To avoid the problems stated above and to keep online arguments stable, toxic comments classification is important.

For text classification, the flow of the methodology is text pre-processing, feature engineering and the final step is classification. In order to feed information into a model for further analysis and learning, text must first be transformed into a clear and consistent format which is called as text pre-processing. Lower case conversion, punctuations removal, numbers removal, stop words removal, stemming, tokenization, etc. are a few of the pre-processing techniques. Feature engineering transforms the raw data to features to improve performance of the model.

Need For Feature Engineering:

Better Features Give Greater Flexibility: One can choose the simpler models due to the features' versatility. It is because, simpler models are quicker to run, simpler to comprehend, and easier to maintain, all of which are always desired.

Better Features Give Simpler Models: After feature engineering, choosing the best model with the most optimised parameters does not need much effort. If strong features are present, the entire set of data can be more accurately represented and used to define the given issue.

Better Features Result In Better Outcomes: In machine learning, the results will be the same regardless of the data supplied. So, better features must be applied in order to get better results.

After feature engineering, the text classifier model is fed with training data, which includes feature vectors for each text sample, after the data has been vectorized. The model will be able to generate accurate predictions with enough training samples. There are several deep learning and machine learning algorithms available for classification according to the needs and applications. Figure 1 shows some of the most popular text classification algorithms.

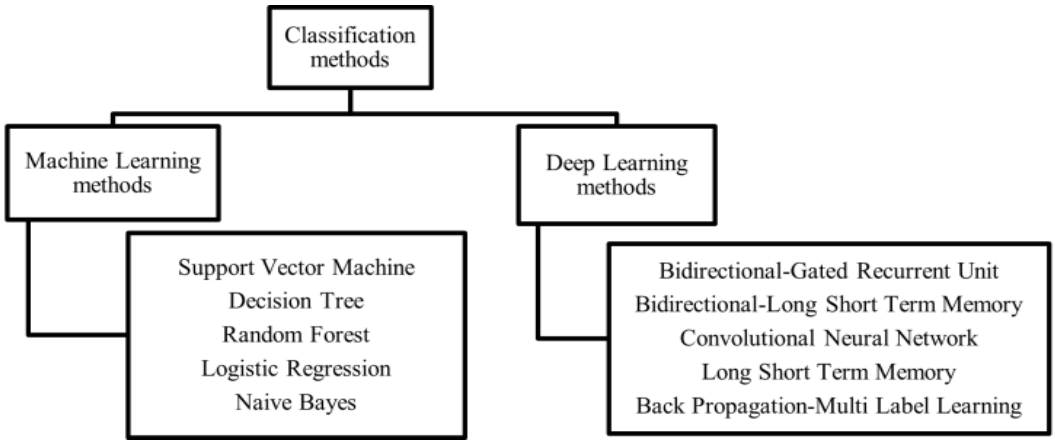

Figure 1. Types of Classification methods

The organization of the manuscript is structured as follows. Sections 2 and 3 provide descriptions about the working of machine learning models and the role of word embedding in classification along with the working of deep learning models. The literature review of the toxic comments' classification is described in section 4. The comparative analysis of the current methodologies is presented in section 5. The survey is concluded in section 6 and offers a suggestion for a future implementation.

2. Machine Learning

It focuses on developing statistical models and algorithms that enable computers to learn, predict the future and make decisions without explicit programming. Large datasets are used to train algorithms to find relationships and patterns, and these patterns are subsequently used to predict or decide on fresh data. Machine Learning methods use manual feature extraction, which is subsequently used by Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest and Logistic Regression algorithms for classification. Figure 2 shows the working of machine learning models.

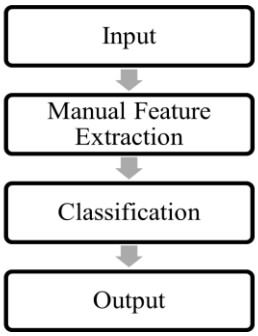

Figure 2. Working of Machine Learning models

Text analysis makes use of the word embedding. Toxic comments' classification is categorised under text analysis. Therefore, in this study, various word embedding techniques and their impacts on classification are analysed. A word embedding in natural language processing is a representation of a word that bridges the human understanding of language to that of a machine. The representation, which is frequently a real-valued vector, forecasts the similarity of the meanings of words that are close to one another in the vector space. There are several static and contextual word embedding techniques available to represent a word. Figure 3 shows some of the most popular word embedding techniques.

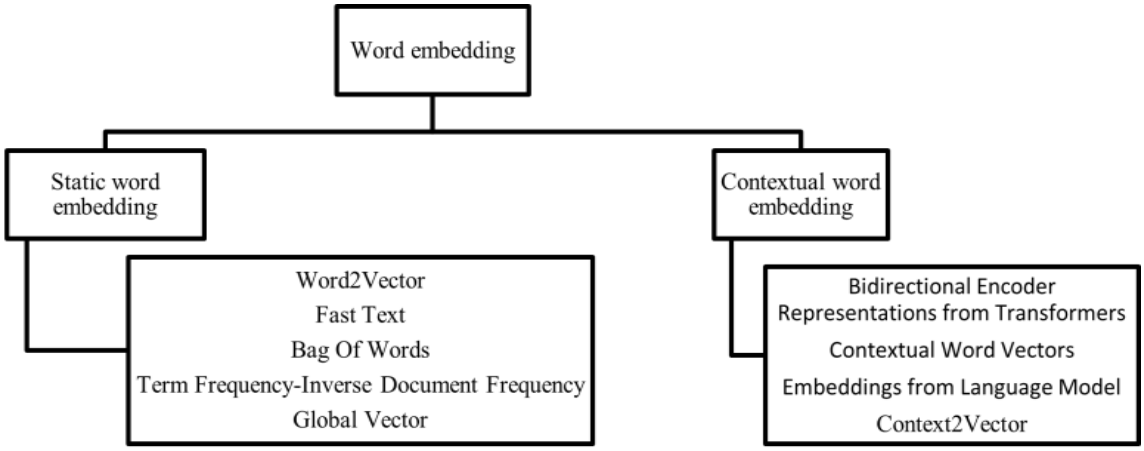

Figure 3. Types of Word embedding techniques

Each word type is mapped to a single vector via a static word embedding function. The vocabulary size is significantly less dimensional and frequently dense in these vectors. By giving each word a representation based on its context, contextual embeddings capture word usage in a variety of contexts and encode cross-linguistic knowledge. Table 1 summarizes the characteristics of some of the word embedding techniques.

Table 1. Characteristics of Word Embedding Techniques

Word Embedding Techniques	Characteristics
TF-IDF	Term Frequency – Inverse Document Frequency (TF-IDF) is a statistical approach to evaluating word importance with respect to the corpus of text. It doesn't record word associations with semantic meaning.
Word2Vector	Continuous Bag of Words (CBOW) and Skip-gram architectures based on neural networks are superior at capturing semantic information.
GloVe	Worldwide word to word co-occurrence -based matrix factorization is Global Vectors (GloVe). It resolves Word2Vec's local context issues.
BERT	High-quality contextual information can be captured using a transformer-based attention mechanism, Bidirectional Encoder Representations from Transformers (BERT).
ELMo	Embeddings from Language Model (ELMo) includes positioning embedding, which generates various vectors for the same word depending on its location and context inside a sentence.

3. Deep Learning

Deep Learning uses neural networks with numerous layers to examine complex patterns and connections in data. It is motivated by the structure and operation of the human brain and is effective at many different tasks including image recognition, natural language processing and speech recognition. Large volumes of data and algorithms that can learn are employed to train deep learning models over time, which enable them to analyse more data and increase their accuracy. This enables them to learn and adjust to new circumstances and makes them well-suited to complex, real world problems.

Since deep learning is an improved form of machine learning, it does not require the manual feature extraction for every issue but instead it attempts to learn high-level features independently from the data. Those are then utilized by deep learning methods such as Bidirectional Gated Recurrent Unit (BiGRU), Bidirectional Long Short -Term Memory (BiLSTM), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for classification. Figure 4 depicts the working of deep learning models.

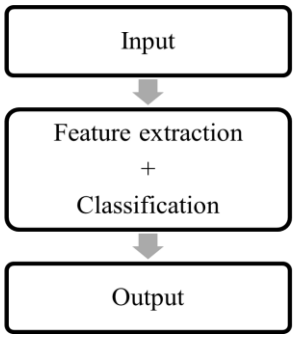

Figure 4. Working of Deep Learning models

4. Literature Survey

The past studies on deep and machine learning methods for toxic comments classification have been provided and analysed in this section along with various word embedding strategies used by the earlier researchers. It covers the dataset that was taken into account, the classification strategies that were employed, and the effectiveness of each methodology.

4.1 Deep Learning Methods

Bjorn Gamback et al., [3] used Hate speech detection dataset [21] which contains four labels such as “racism, sexism, both (racism and sexism) and non-hate-speech”. CNN was used for classification along with different word embedding techniques like Word2Vector, Character n-grams as well as a combination of Word2Vector and Character n-grams. Among that, CNN performed better with Word2Vec word embedding technique. Classification was done by 10-fold cross-validation with a F1 score of 78.3%.

Mai Ibrahim et al., [4] proposed an ensemble of classification models consisting of Bi-LSTM, CNN and Bi-GRU. Wikipedia dataset [20] which contains six labels such as toxic, obscene, threat, severe toxic, identity, hate, and insult, was used. The solution begins with identifying the toxic comment and its types. The ensemble strategy attained 87.2% for toxicity categories prediction and 82.8% F1 score for categorizing.

Mujahed A. Saif et al., [5] presented four deep learning models. Google Jigsaw’s Toxic Comment Classification Dataset [19] which contains six labels such as toxic, obscene, threat, severe toxic, identity, hate, and insult, was used. The results showed that integrating two deep learning offers the best accuracy of 98.2%, which is the most effective one.

Spiros V. Georgakopoulos et al., [6] performed toxic comments classification by using deep learning and machine learning. Wikipedia dataset [20] was used for the research and the final result showed that CNN outperformed when compared to all other machine learning algorithms. An accuracy of 91.2% was achieved for CNN.

Hafiz Hassaan Saeed et al., [7] accurately classified the overlapping toxic comments using a Deep Neural Network architecture called Bi-GRU. Furthermore, it was demonstrated that the suggested classification framework is capable of managing text pre-processing (such as stop word removal, feature engineering, etc.) intrinsically without the need for any time-consuming laborious text pre-processing. Google Jigsaw's Toxic Comment Classification Dataset [19] was used. The best recall, precision, and F1 score were all achieved by CNN. Moreover, Bi-LSTM demonstrated very strong precision, recall, and F1 scores. The greatest F-1 score, and best precision and recall were both achieved by Bi-GRU. Overall, it appeared that Bi-GRU outperformed all other suggested architectural designs.

A.N.M. JuBaer et al., [8] performed toxic comments classification for bangla language. Their own dataset was created from different Facebook page comments. Several methods were compared, and among that Back Propagation- Multi Label Learning (BP-MLL) performed well. The accuracy for Binary Relevance Method with Multinomial Naive Bayes (NB) Classifier is 52.30%, for Binary Relevance with SVM Classifier is 30.76%, for Binary Relevance Method with Gaussian NB classifier is 49.23%, for Classifier Chain with Multinomial NB Classifier is 52.30%, for Label Powerset with Multinomial NB Classifier is 58.46%, for Multi Label KNN Classifier is 58.46% and for Back Propagation Multi Label Learning classifier is 60% which is highest among others.

Sara Zaheri et al., [9] proposed LSTM for toxic comments classification and compared it with well-known Naive Bayes algorithm. Google Jigsaw's Toxic Comment Classification dataset [19] was used, and an accuracy of 73% was attained when Rectified Linear Unit (ReLU) was used as an activation layer for LSTM.

Muhammad Umer et al., [10] used three datasets to gauge the success of the suggested strategy: Hate speech detection dataset [21], Women's e-commerce clothing reviews [25] and US Airline twitter dataset [24]. All machine learning classifiers performed well with Word2Vector word embedding technique. Among all, Stochastic Gradient Descent with W2V

attained 88.5% accuracy which was still poor when compared to proposed CNN-LSTM deep learning method which achieved an accuracy of 92%.

4.2 Machine Learning Methods

Shervin Malmasi et al., [11] employed Linear SVM classifier for the categorization of toxic remarks. Hate speech detection dataset [21] which contains three labels such as Hate, Offensive, and Ok comments, was used. The best performance was achieved by SVM with character 4-gram model with 78% accuracy.

Ricardo Martins et al., [12] presented three different machine learning algorithms such as SVM, Naïve Bayes (NB) and Random Forest (RF) for classification. Their own dataset was created from Davidson's Hate Speech Detection dataset and therefore, the new dataset included fourteen labels, including the classification, the original text message, eight fundamental emotions (anger, joy, sadness, disgust, fear, trust, and surprise). Random Forest with an accuracy of 80.56% outperformed when compared to other classifiers and used ten-fold cross validation for results validation.

S. Rahamat Basha et al., [13] analyzed different feature selection techniques such as Chi Square, Information Gain, Mutual Information, Odds Ratio, NGL Coefficient, and GSS Coefficient, and their impacts on toxic comments classification. A self-made corpus was used for training and Reuters-21578 corpus [22] was used for testing. Classification was done by using two machine learning algorithms like NB and KNN along with all the feature selection techniques. Among that, Naïve Bayes with Information Gain performed better than other combinations and attained an accuracy of 89.1%.

Shaik Rahamat Basha et al., [14] examined the accuracy, average precision, recall, and precision with and without the use of certain feature selection strategies such as Document Frequency, Chi Square, Information Gain, Mutual Information, Odds Ratio, NGL Coefficient, GSS Coefficient, Relevancy Score, and Multi-Set of Feature, and their impacts on two classifiers like Naïve Bayes and K-Nearest Neighbor for toxic comments classification. A self-made corpus was used for training and Reuters-21578 corpus [22] was used for testing. When compared with KNN, Naïve Bayes classifier performed well with all feature reduction techniques.

Saqib Alam et al., [15] employed various text pre-processing approaches and investigated their effects on accuracy for classification. The accuracy of three machine learning algorithms was calculated both before and after the pre-processing steps were used. Findings showed that using the pre-processing steps considerably increased the algorithm's accuracy. Twitter dataset [23] was used for training and testing. Before pre-processing, Naïve Bayes attained an accuracy of 83.69%. Whereas after pre-processing, NB achieved an accuracy of 91.81% which shows that, after implementing text pre-processing processes, accuracy has greatly increased.

Furqan Rustam et al., [16] proposed an ensemble method called Voting Classifier (VC) in order to aid sentiment analysis for US Airline companies. The final prediction was made by the VC using a soft voting process and is based on Stochastic Gradient Descent Classifier (SGDC) and Logistic Regression (LR). Based on the sentiments they contain, tweets were divided into three categories: positive, negative, and neutral. US Airline Twitter Dataset [24] was used for the research work. The proposed ensemble approach offered the highest accuracy of 79.1%.

Salvatore Carta et al., [17] used various word embedding combinations to assess the effectiveness of the classifier. Wikipedia dataset [20] which contains Udemy reviews was used, and Logistic Regression was used for classification. Based on the results, it can be concluded that using word embeddings can increase accuracy compared to baseline approaches that do not use them. The outcomes also demonstrated that, in a given domain, contextual word embeddings outperform canonical state-of-the-art word embeddings.

Hoyeon Park et al., [18] analysed the effect of different word embedding approaches on the effectiveness of sentiment analysis. Word2Vector, Bag of Words and TF-IDF were used for word embedding along with NB, SVM, RF, Gradient Boosting and XGBoost for classification. 30,000 datasets from the IMDB dataset provided by Keras [26], which has already been classed as positive and negative, were used. The performance analysis shows that TF-IDF outperformed when compared to other word embedding techniques. Among all classifiers, SVM achieved best accuracy of 89.42% when combined with TF-IDF.

5. Comparative Analysis

Comparative analysis involves comparing methodologies against one another to identify

their similarities and differences. In order to prepare for a comparative analysis, considerable study is essential. In addition to supplying proof to back up the conclusions, research may also give a viewpoint or perspective that hadn't been previously considered. Comparative analyses are crucial for improving the understanding of a subject. Table 2 summarizes the previous works on toxic comments classification and methodologies used to classify the toxic comments. It also includes the dataset and accuracy of various learning models.

Table 2. Comparative Analysis

Ref No	Year	Dataset used	Methodology used	Accuracy
[3]	2017	Hate speech detection dataset	CNN with Word2Vector as word embedding technique	78.3% F1 score
[4]	2018	Wikipedia dataset	Ensemble of three classifiers such as CNN, Bi-LSTM, Bi-GRU	82.8% F1 score
[5]	2018	Google jigsaw's toxic comment classification dataset	Ensemble of CNN and LSTM	98.2%
[6]	2018	Wikipedia dataset	CNN	91.2%
[7]	2018	Google jigsaw's toxic comment classification dataset	Bidirectional Gated Recurrent Unit	Not given
[8]	2019	Own dataset	BP-MLL	60%
[9]	2020	Google jigsaw's toxic comment classification dataset	LSTM with ReLU as an activation function	73%
[10]	2021	Hate speech detection dataset, "Women's e-commerce clothing reviews" and US Airline twitter dataset	Convolutional Neural Network – Long Short -Term Memory (CNN-LSTM)	92%
[11]	2017	Hate speech detection dataset	SVM with Character 4-gram as word embedding technique	78%
[12]	2018	Own dataset	Random Forest with tenfold cross validation	80.56%
[13]	2019	Self-made corpus for training and Reuters-21578 corpus for testing	Naïve Bayes with Information Gain as feature selection technique	89.1%
[14]	2019	Self-made corpus for training and Reuters-21578 corpus for testing	All feature reduction methods worked well with Naive Bayes	Not given

[15]	2019	Twitter dataset	Analyze the importance of pre-processing and used Naïve Bayes for classification	91.81% After pre-processing
[16]	2019	US Airline twitter dataset	Ensemble of LR and SGDC	79.1%
[17]	2019	Wikipedia dataset	LR performed well with contextual word embedding	Not given
[18]	2020	IMDB dataset	SVM with TF-IDF as word embedding technique	89.42%

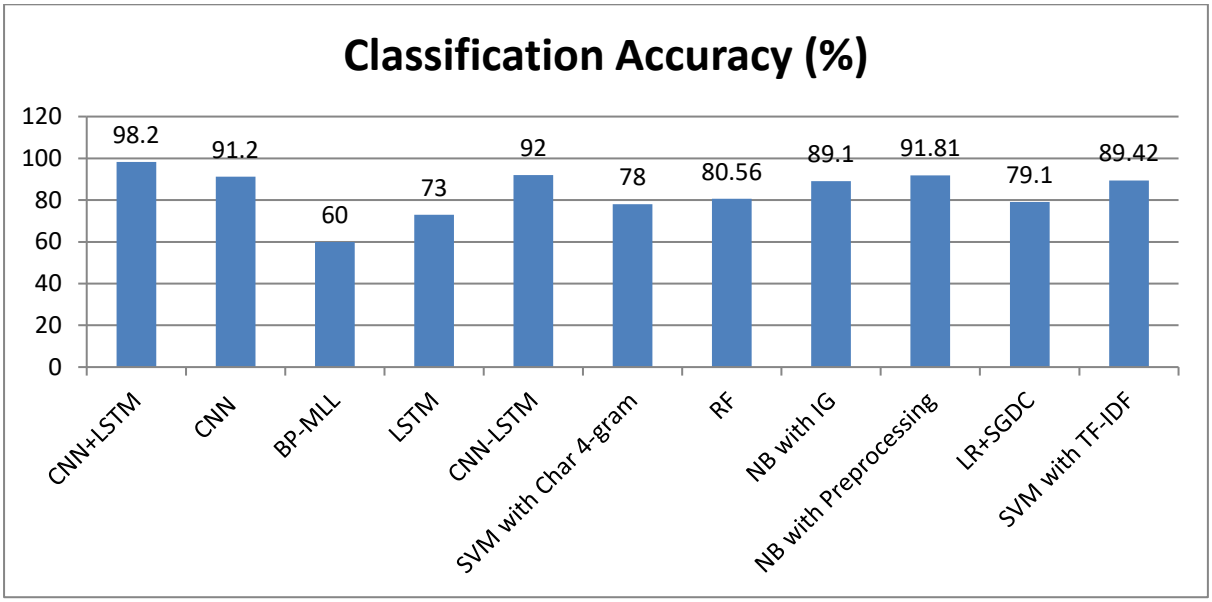

Figure 5. Performance of various classifiers

Figure 5 shows the accuracy of various classifiers for the toxic comments' classification task. From this figure, it is identified that NB with pre-processing, CNN-LSTM and CNN+LSTM give the better results. But compared to others, CNN+LSTM achieved better performance. Likewise, the outcomes showed that the accuracy of the algorithms was greatly increased by the pre-processing phases. Also, the effectiveness of sentiment analysis was examined in relation to the impact of various feature engineering techniques. The word2vector, bag of words and TF-IDF techniques were used for feature engineering along with machine learning classifiers for classification. The performance research showed that TF-IDF performed significantly better when compared to others.

6. Conclusion

In this study, a thorough review on the existing methods that are already present in the

classification of toxic comments by using various deep learning and machine learning algorithms as well as the impact of word embedding techniques on classification task has been done. From this study, it is observed that ensemble methods outperform well when compared to individual algorithms by showing better results for the classification tasks. In addition, word embedding techniques influence much when combined with the algorithms for classification. In future work, classification may be applied with ensemble methods for toxic comments classification along with static word embedding techniques for better efficiency.

References

[1] Digital 2023: India –DataReportal-Global Digital Insights, february 2023 (online). Available: https://datareportal.com/reports/digital-2023-india.

[2] Digital Around the World – DataReportal (online). Available: https://datareportal.com/global-digital-overview.

[3] B. Gamback and U. K. Sikdar, ‘‘Using convolutional neural networks to classify hate-speech,’’ in Proc. 1st Workshop Abusive Lang. Online, 2017, pp. 85–90.

[4] M. Ibrahim, M. Torki, and N. El-Makky, ‘‘Imbalanced toxic comments classification using data augmentation and deep learning,’’ in Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2018, pp. 875–878.

[5] M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, ‘‘Classification of online toxic comments using the logistic regression and neural networks models,’’ AIP Conf. Proc., vol. 2048, no. 1, 2018, Art. no. 060011.

[6] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos, ‘‘Convolutional neural networks for toxic comment classification,’’ in Proc. 10th Hellenic Conf. Artif. Intell., Jul. 2018, pp.1–6.

[7] H. H. Saeed, K. Shahzad, and F. Kamiran, ‘‘Overlapping toxic sentiment classification using deep neural architectures,’’ in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2018, pp. 1361–1366.

[8] A.N. M. Jubaer, A. Sayem, and M. A. Rahman, ‘‘Bangla toxic comment classification (machine learning and deep learning approach),’’ in Proc. 8th Int. Conf. Syst. Modeling Adv. Res. Trends (SMART), Nov. 2019, pp. 62–66.

[9] S. Zaheri, J. Leath, and D. Stroud, ‘‘Toxic comment classification,’’ SMU Data Sci. Rev., vol. 3, no. 1, p. 13, 2020.

[10] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. S. Choi, ‘‘Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model,’’ Comput. Intell., vol. 37, no. 1, pp. 409–434, Feb. 2021.

[11] S. Malmasi and M. Zampieri, ‘‘Detecting hate speech in social media,’’ 2017, arXiv:1712.06427. [Online]. Available: http://arxiv.org/abs/ 1712.06427.

[12] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, ‘‘Hate speech classification in social media using emotional analysis,’’ in Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS), Oct. 2018, pp. 61–66.

[13] S. R. Basha, J. K. Rani, J. P. Yadav, and G. R. Kumar, ‘‘Impact of feature selection techniques in text classification: An experimental study,’’ J. Mech. Continua Math. Sci., no. 3, pp. 39–51, 2019.

[14] S. R. Basha and J. K. Rani, ‘‘A comparative approach of dimensionality reduction techniques in text classification,’’ Eng., Technol. Appl. Sci. Res., vol. 9, no. 6, pp. 4974–4979, Dec. 2019.

[15] S. Alam and N. Yao, ‘‘The impact of pre-processing steps on the accuracy of machine learning algorithms in sentiment analysis,’’ Comput. Math. Org. Theory, vol. 25, no. 3, pp. 319–335, Sep. 2019.

[16] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, ‘‘Tweets classification on the base of sentiments for US airline companies,’’ Entropy, vol. 21, no. 11, p. 1078, Nov. 2019.

[17] S. Carta, A. Corriga, R. Mulas, D. Recupero, and R. Saia, ‘‘A supervised multi-class multi-label word embeddings approach for toxic comment classification,’’ in Proc. KDIR, 2019, pp. 105–112.

[18] Hoyeon Park and Kyoung-jae Kim, “Impact of Word Embedding Methods on Performance of Sentiment Analysis with Machine Learning Techniques,” in Journal of The Korea Society of Computer and Information Vol. 25 No. 8, pp. 181-188, August 2020.

[19] Google jigsaw's toxic comment classification dataset: Toxic Comment Classification Challenge. Accessed: May 5, 2020. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-commentclassification-challenge.

[20] Wikipedia dataset: Wikipedia talk page edit dataset. [Online]. Available: https://www.kaggle.com/datasets/jigsaw-team/wikipedia-talk-labels-personal-attacks.

[21] Hate speech detection dataset: [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset.

[22] Reuters 21578 corpus: [Online]. Available: https://www.kaggle.com/datasets/feyzazkefe/reuters21578-sgm.

[23] Twitter dataset: [Online]. Available:https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset.

[24] US Airline twitter dataset: [Online]. Available: https://www.kaggle.com/datasets/vedaangchopra/twitter-us-airline-sentiment-dataset.

[25] Women's e-commerce clothing reviews: [Online]. Available: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

[26] IMDB dataset provided by Keras: [Online]. Available: https://keras.io/api/datasets/imdb