IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204354 |

# Toxic Comment Classification Using Machine Learning Methods

**Dr.J Stanly Jayaprakash, S.Aravindhan, M.Veeran, P.Hariprasath, Sfawn Ahmed**

Professor, Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India

Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India

**ABSTRACT:** Online Conversation media serves as a means for individuals to engage, cooperate, and exchange ideas; however, it is also considered a platform that facilitates the spread of hateful and offensive comments, which could significantly impact one's emotional and mental health. The rapid growth of online communication makes it impractical to manually identify and filter out hateful tweets. Consequently, there is a pressing need for a method or strategy to eliminate toxic and abusive comments and ensure the safety and cleanliness of social media platforms. Utilizing LSTM, Character-level CNN, Word-level CNN, and Hybrid model (LSTM + CNN) in this toxicity analysis is to classify comments and identify the different types of toxic classes by means of a comparative analysis of various models. The neural network models utilized for this analysis take in comments extracted from online platforms, including both toxic and non-toxic comments. The results of this study can contribute towards the development of a web interface that enables the identification of toxic and hateful comments within a given sentence or phrase, and categorizes them into their respective toxicity classes**.**

**KEYWORDS:** Machine learning, Neural Network.

## I. INTRODUCTION

Diverse communities such as education, social, business, political, etc. show a great interest in internet based collaborative environments with a desire to share valuable contents with others, to grow and nourish relationships, to get the word out about brands and causes they like or support, and many more. Online communication media is being used in ways that shape politics, business, world culture, education, careers and innovation. Flood of information is produced everyday through the global internet usage arising from the widespread use of online interactive communication media. While this situation contributes significantly to the quality of human life, unfortunately it involves dangers like texts with high toxicity that can cause personal attacks, online harassment and bullying behaviours that may prejudice an individual's emotional and mental well being.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204354 |

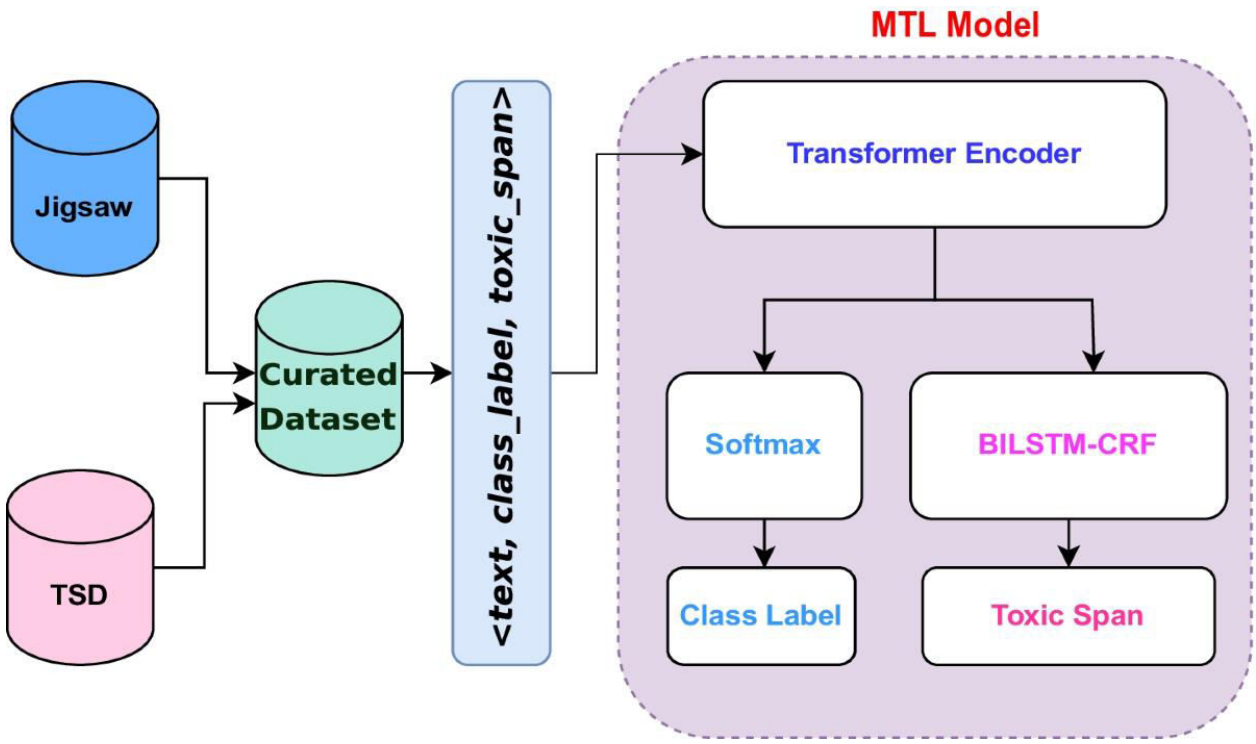


Fig 1: Multi-task learning for toxic comment classification

The threat of abuse and harassment online force many people to stop expressing themselves and give up on expressing different Every day, vast amounts of data are released by social media networks. This massive volume of data is having a big impact on the quality of human existence However, because there is so much toxicity on the Internet, it can be harmful. Because toxic comments limit people's ability to express themselves and have different opinions, as a result of the negative, there are no positive conversations on social networks. As a result, detecting and restricting antisocial behaviour in online discussion forums is a pressing requirement. Toclassify noxious comments, this study will use different methods of machine learning. To handle the problem of text categorisation, some of the methods we use are logistic entertainment, random forest, SVM classifier, multi-navigation database, and XGBoost classifier. As a result, we maintain a data set using six machine learning algorithms, evaluate their accuracy, and compare them. We choose the model with the highest accuracyandforecastthetoxicitybasedonunseendatabasedonitsaccuracylevel. In this project we are trying to explore how sentiment analysis using deep neural networks can help us identify and classify toxic texts in online communication systems and try and filter these out in the best possible way so that we can make internet communication media cleaner and safer to use.

## II. LITERATURE REVIEW

The existing system of the Toxic Comment Classification emphasizes upon the basic LSTM and CNN methods to bringdown the two levels of granularity, the first one being the word level and another being the character level which workson both binary and multi label classification tasks. Convolutional Neural Network (CNN) is a level of DeepLearning algorithm which works well on parts such as takinginput images and finding patterns in it, it works in such a way to recognize objects, classes and even categories. CNN depends upon the network architecture for learning from the data. Recurrent neural networks can learn order dependency inprediction problems using Long Short Term Memory (LSTM) networks, a complicated subfield of Deep Learning. Such learning is beneficial for numerous complicated fields, including speech recognition, machine translation, and manyothers.

In the recent times, the users on social media platforms have increased dramatically, leading to a significant increase in the amount of online content being generated. However, this surge in online activity has also led to a rise in toxicity and negativity on these platforms. Online communities are now inundated with toxic comments and hateful messages that can have serious psychological and emotional impacts on the targeted individuals. This has created a pressing need for effective methods to identify and remove toxic comments from online platforms to ensure a safe and inclusive

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204354 |

environment forall users. To deal with this problem, machine learning models have been created to automatically classify comments as toxic ornon-toxic. In this research paper, we present an approach for toxic comment classification using a Bi-directional LSTM neural network architecture. We test our approach using a publicly available dataset provided by a Kaggle competition. Our experiments show that LSTM performs excellent on text classification tasks with 98% AUC score

## III. METHODOLOGY

In this study, we classify toxic comments with four different categories such as pornography, defamation, hate speech, and radicalism by using the SVM method. We also run several scenarios with the goal is to obtain the best performance from overall scenarios. The scenario focused on comparing the preprocessing stages, which are stemming and stop words removal, the implementation of Chi Square as feature selection, and the use of SVM kernels. This research includes five main steps: preprocessing, feature extraction using TF-IDF, feature selection using Chi Square, classification using SVM, and evaluation using F 1 − Score as the metric. The flow of system shown in Figure 1 and all the detail of each step is described as follows.

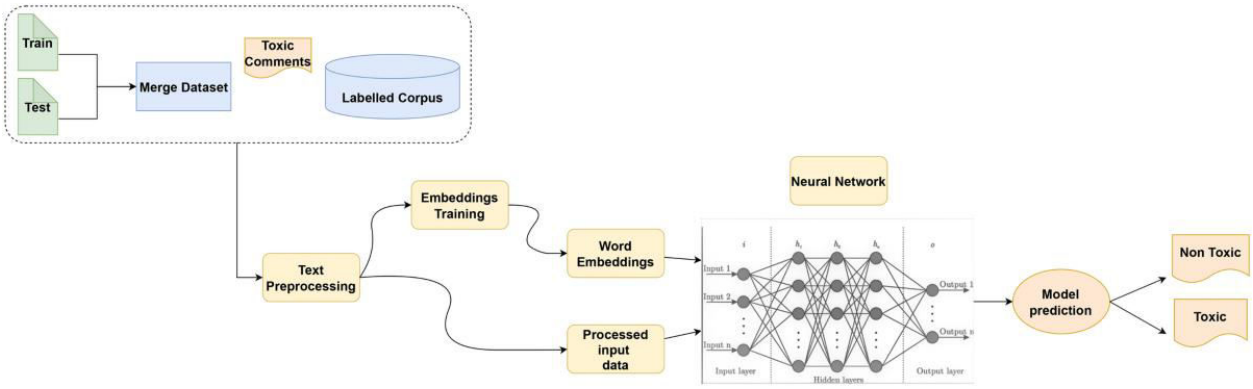


Fig 2: Deep learning for religious and continent-based toxi

Toxic comments have become a major issue on online platforms, including social media, blogs, and news websites.Such comments can cause harm to individuals and communities, and can negatively impact online discourse. Due to the scale of the problem, it is not feasible for human moderators to manually review every comment posted on these platforms. Thus, machine learning models have been developed to automatically classify comments as toxic or non-toxic which can help in having valuable discussionsonline and can also help governments make regulations. In this study, we suggest a Bi-directional LSTM model for toxic comment classification. Recurrent neural network of this type are very effective at processing sequential data, including text and every component of an input sequence hasinformation on both past and future data. Also, BidirectionalLSTM model is beneficial in some NLP tasks, such as sentence classification, translation, and entity recognition.

## IV. RESULT ANALYSIS

The process of converting unprocessed data into a format better suited for analysis is known as data preprocessing. It is an essential phase in machine learning since the accuracy of the output is closely related to the quality of the input data. Data preprocessing can involve a variety including data cleaning, data transformation, and data normalization. Some common techniques used in data preprocessing include: Data Cleaning- In this part of the project we performed different data cleaning operations such as Lower casing, removing unwanted characters, removing characters from both left and right, removing punctuations and numbers and single character removing. Regardless of how advanced our machine learning algorithm is, we cannot obtain better results from bad data.

Which is why data cleaning is a very essentialpart of machine learning. Tokenization- Tokenization is the process of breaking down streams of textual data into smaller meaningful elements called tokens. For example, a sentence can be a token of a paragraph, a word can be a token of a sentence. This turns anunstructured textual data into a numerical data structure that is suitable for machine learning. In our project we performed tokenization by first creating an object

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204354 |

instance of tokenizer class with argument num_word equal to 100000 which is nothing but the number of wordstowork with.Then we will be passing the dataset as a parameterin “fit_on_texts” then the list of words are converted into tokens using “tokenizer texts_to_sequences”.

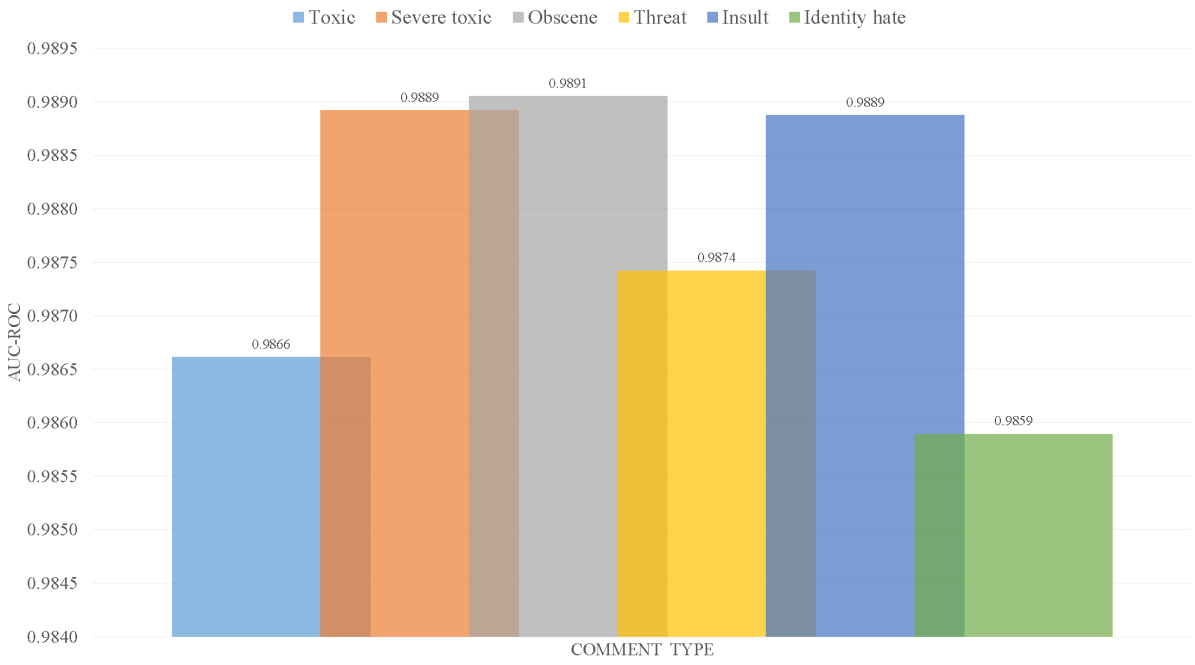


Fig 3: Result analysis

The model architecture consists of several layers for text classification. First, an input layer takes input sequences of maximum length (maxlen). Then, an embedding layer maps the words in the input sequences to dense vectors using pre- trained GloVe word embeddings, with the weights initialized from an embedding_matrix_glove. The embedding layer is set to be non-trainable to keep the pre-trained embeddings unchanged. To prevent overfitting, a SpatialDropout1D layer applies a dropout rate of 0.2 to the embedded sequences. Next, a bidirectional LSTM layer with 128 hidden units is employedto capture contextual information from both past and future tokens in the input sequences. Dropout of 0.1 is applied during training, and recurrent_dropout of 0.1 is used during recurrent connections to regularize the LSTM layer.

## V. CONCLUSION

In this study, we addressed toxicity detection in social media using deep learning techniques. We adopt the Bidirectional Encoder Representations from Transformers (BERT) to classify toxic comments from usergenerated data in social media, such as tweets. The BERT-base pre-trained model was Zine-tuned on a well known labeled toxic comment dataset, Kaggle public datasets. Moreover, the proposed model was tested on real-world data, two different tweets datasets, collected in two different periods based on a case study of the UK Brexit. The evaluation outcomes showed that BERT has the ability to classify and to predict toxic comments with a high accuracy rate. Moreover, we compared the BERT-base model to three models, called Multilingual BERT, RoBERTa, and DistilBERT. The BERT-base model outperformed all compared models and achieved the best results.

## REFERENCES

1. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hussien, A.G.; Khasawneh, A.M.; Alshinwan, M.; Houssein, E.H. Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis. Algorithms2020,13,345.[CrossRef]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204354 |

2. Uhls, Y.T.; Ellison, N.B.; Subrahmanyam, K. Benefits and costs of social media in adolescence. Pediatrics 2017,140,S67–S70.[CrossRef][PubMed]
3. Souri, A.; Hosseinpour, S.; Rahmani, A.M. Personality classification based on profiles of social networks' users and the five-factor model of personality. Hum. Centric Comput. Inf. Sci. 2018, 8, 24. [Google Scholar] [CrossRef][GreenVersion]
4. Morente-Molinera,J.A.;Kou,G.;Samuylov,K.; ren a R.;Herrera-Viedma,E.Carrying out consensualGroup Decision Making processes under social networks using sentiment analysis over comparative expressions. Knowl.BasedSyst.2019,165,335–345.[GoogleScholar][CrossRef]
5. Risch, J.; Krestel, R. Aggression identification using deep learning and data augmentation. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018;pp.150–158..[Google Scholar]
6. Subramani,S.;Wang,H.;Vu,H.Q.;Li,G.Domestic violencecrisis identificationfromfacebookpostsbasedon deeplearning.IEEEAccess2018,6,54075–54085.[GoogleScholar][CrossRef]
7. Subramani, S.; Michalska, S.; Wang, H.; Du, J.; Zhang, Y.; Shakeel, H. Deep Learning for Multi-Class Identification From Domestic Violence Online Posts. IEEE Access 2019, 7, 46210–46224. [Google Scholar] [CrossRef]
8. Abualigah, L.M.Q. Feature Selection And Enhanced Krill Herd Algorithm For Text Document Clustering; Springer:Berlin/Heidelberg,Germany,2019..[GoogleScholar]
9. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hamad, H.A.; Omari, M.; Alshinwan, M.; Khasawneh, A.M. Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering. Electronics 2021, 10, 101. [Google Scholar][CrossRef]
10. Ahmad, S.; Asghar, M.Z.; Alotaibi, F.M.; Awan, I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Hum. Centric Comput. Inf. Sci. 2019, 9, 24. [GoogleScholar] [CrossRef][GreenVersion]
11. Chiranjeevi, P.; Santosh, D.T.; Vishnuvardhan, B. Survey on Sentiment Analysis Methods for Reputation Evaluation. In Cognitive Informatics and Soft Computing; Springer: Berlin, Germany, 2019; pp. 53–66. [GoogleScholar]
12. Alaei,A.R.;Becken, S.; Stantic,B. Sentiment analysis in tourism: capitalizing on big data.J.TravelRes. 2019, 58,175–191.[[GoogleScholar][CrossRef]
13. Cury, R.M. Oscillation of tweet sentiments in the election of oa o Doria Jr.for Mayor. J. Big Data 2019, 6, 42. [Google Scholar][CrossRef]
14. Al Shehhi, A.; Thomas, J.; Welsch, R.; Grey,I.; Aung, Z. Arabia Felix 2.0: a cross-linguistic Twitter analysis of happiness patterns in the United Arab Emirates. J. Big Data 2019, 6, 33. (Google Scholar] [CrossRef][Green Version]
15. Pong-inwong, C.; Songpan, W. Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining. Int. J. Mach. Learn. Cybern. 2018, 10, 2177–2186. [CrossRef]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details