International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Toxic Comments Classification in Social Networking Sites

***M. RajaBabu*[1], *J. Sirisha*[2], *G. Meghana*[3], *G. Anusha*[4]**

[1,2,3,4]Department of Information Technology, Aditya Engineering College, Surampalem, Andhra Pradesh, India.

ABSTRACT:

Classification of toxic comments with many newly proposed approaches, social networking has become an active study field. While these approaches handle some of the task's challenges, others remain unresolved, necessitating further investigation. To do this, we compare different deep learning and shallow approaches on a brand-new, sizable comment dataset and recommend an ensemble that outperforms all individual models. Using a second sample, we also confirm our findings. We can perform a thorough error analysis using the ensemble results, which reveals open problems for cutting-edge techniques as well as potential research avenues. These challenges include an absence of a paradigmatic framework and inconsistent dataset labels.

Keywords: Classification, Support Vector Machine, Toxic Comments.

1. INTRODUCTION:

Because of the increased use of the internet, the use of social media and social networking has grown tremendously over the years. On a daily basis, a flood of information emerges from online discussion because people can discuss, express themselves, and air their opinions through these platforms. While this scenario has the potential to be highly productive and improve human life quality, it also has the potential to be destructive and extremely dangerous. When a discussion or conversation begins, it is natural for disagreements to emerge due to differences in opinion. However, these discussions frequently devolve into dirty fights on social media, with one party using offensive language referred to as toxic comments. These toxic remarks can be threatening, obscene, insulting, or founded on identity hatred. As a result, these obviously pose the risk of online abuse and harassment. As a result, some people stop giving their opinions or stop seeking various perspectives, resulting in unhealthy and unfair debate. As a result, various platforms and communities struggle to enable fair conversation and are frequently compelled to either limit user comments or dissolve by shutting down user comments.

Deep learning, also known as deep neural networks, has shown tremendous promise in a variety of practical uses in recent years. Exceptional performance in a number of areas, including speech recognition, visual object recognition, and text processing, has been documented. Indeed, one could say that the network's ability to learn has been a critical factor in the recent success of pattern recognition applications. It has also been discovered that increasing the scale of deep learning in terms of the number of training examples, model parameters, or both, can significantly increase ultimate classification accuracy.

As a result, the use of GPUs has been a major advancement in recent years, making the training of modestly sized deep networks feasible. It has been known since the early days of ship detection and classification that the variability and richness of image data make it nearly impossible to create an accurate model.

Toxic comment classification on online platforms is traditionally done by moderators or with the assistance of text classification tools. With new advances in Deep Learning (DL) techniques, researchers are investigating whether DL can be used for the job of comment classification. Text classification is a well-known issue in natural language processing, and it is used in many applications such as web searching, information filtering, topic categorization, and sentiment analysis. Text transformation is the first stage in any sort of text classification.

Online comments are typically written in non-standard English and contain numerous spelling mistakes caused by typos (due to the small screens of mobile devices), but also by a purposeful effort to write abusive comments in creative ways to avoid automated filters.

2. RELATED WORK:

A systematic study of machine learning techniques for toxic comment classification [1]. Users now leave numerous remarks on social media sites, news websites, and forums. Some comments are derogatory or violent. Because directly monitoring so many comments is impractical, the majority of systems rely on machine learning algorithms to identify potentially harmful content. We used machine learning techniques in this study to perform a thorough analysis of the state-of-the-art in harmful comment classification. We gathered information from 31 carefully chosen major relevant research. We started by looking into the papers' release dates, locations, and levels of maturity. The data collection, evaluation metric, machine learning techniques, types of toxicity, and comment language of each significant research were all examined. We conclude our work by providing a thorough list of current research gaps as well as suggestions for potential future research topics related to the problem of harmful online comment classification.

Toxic Content Detection Using Convolutional Neural Networks in Sentimental Analysis [2] The use of various social media platforms has recently increased considerably, which may have both positive and negative effects on people's lives. A user's or person's remark or opinion shared on various social networking sites is one of the factors. This study examines various methods for analysing modelling and classification algorithms for detecting dangerous comments using convolution neural networks. (CNN). It also explains an algorithm based on outlier detection on a given data stream that can evaluate social media content and determine its positive and negative effect on society.

Machine Learning Methods for Toxic Comment Identification and Classification [3] Internet traffic has increased exponentially over the last four months as a result of the ongoing pandemic. This has resulted in a large number of enthusiastic new and old clients using the internet for a wide range of services, including academic, entertainment, industrial, and monitoring, as well as the emergence of a new tendency in business life known as work-from-home. As a consequence of the unexpected rise in the number of people using the internet, the number of cunning people has increased. Nowadays, the top goal for any internet platform provider is to keep inclusive and positive interactions. Twitter, an online media platform where users can express their views, is the best example that can be used. This platform has already received a great deal of criticism due to the proliferation of hate speech, insults, threats, and defamatory acts, making it difficult for many internet service providers to regulate them. As a result, studies into the classification of toxic comments are presently underway. On the dataset, we combine non-identical machine learning and other unimportant methods to propose a model that outperforms them all when compared side by side. For the reasons stated above, we used the Kaggle dataset, which is a well-known and valuable resource for academics attempting to understand the issue of toxic comment classification. The findings would aid in the development of an online interface that would allow us to determine the level of toxicity contained in a particular phrase or sentence and categorize it accordingly.

Classification of Toxic Comments [4]. Conversational toxicity is an issue that can cause people to stop being themselves and stop seeking the opinions of others because they are afraid of being harassed or attacked. In this research, natural language processing (NLP) approaches are used to detect toxicity in writing and warn people before sending potentially harmful informational messages. Natural language processing (NLP), a subset of machine learning, allows machines to comprehend human speech. A computer is capable of comprehending, analysing, manipulating, and even creating human data language. Natural language processing (NLP) is an artificial intelligence subset that allows computers to comprehend and analyse human language rather than merely reading it. Natural language processing enables computers to understand spoken or written English and carry out tasks such as speech identification, sentiment analysis, text classification, and automatic text summarization. (NLP).

Deep Learning Classification of Social Media Toxicity: Real-World UK Application Brexit [6] social media is now widely used by people to express their perspectives on a range of issues, and it is now an integral component of contemporary culture. Social media is becoming more and more necessary for the majority of people, and there have been numerous accounts of social media addiction. Twitter and other social media platforms have demonstrated how important it is to bring people from all over the globe and from various backgrounds together over time. However, they have shown negative side effects that could have a negative influence. One such unfavourable side effect is the extreme poisonousness of many social media discussions. In this study, we develop a useful model for identifying and classifying toxicity in user-generated material on social media using Transformers' Bidirectional Encoder Representations. (BERT). The BERT pre-trained model and three of its variants were enhanced using the Kaggle public dataset, a well-known labelled toxic remark dataset. (Toxic remark Classification Challenge). We also test the proposed models using two datasets gathered from Twitter during two different time periods to detect toxicity in user-generated content (tweets) using hashtags related to the UK Brexit. The findings demonstrated how well the suggested approach classified and analysed harmful tweets.

3. PROPOSED MODEL:

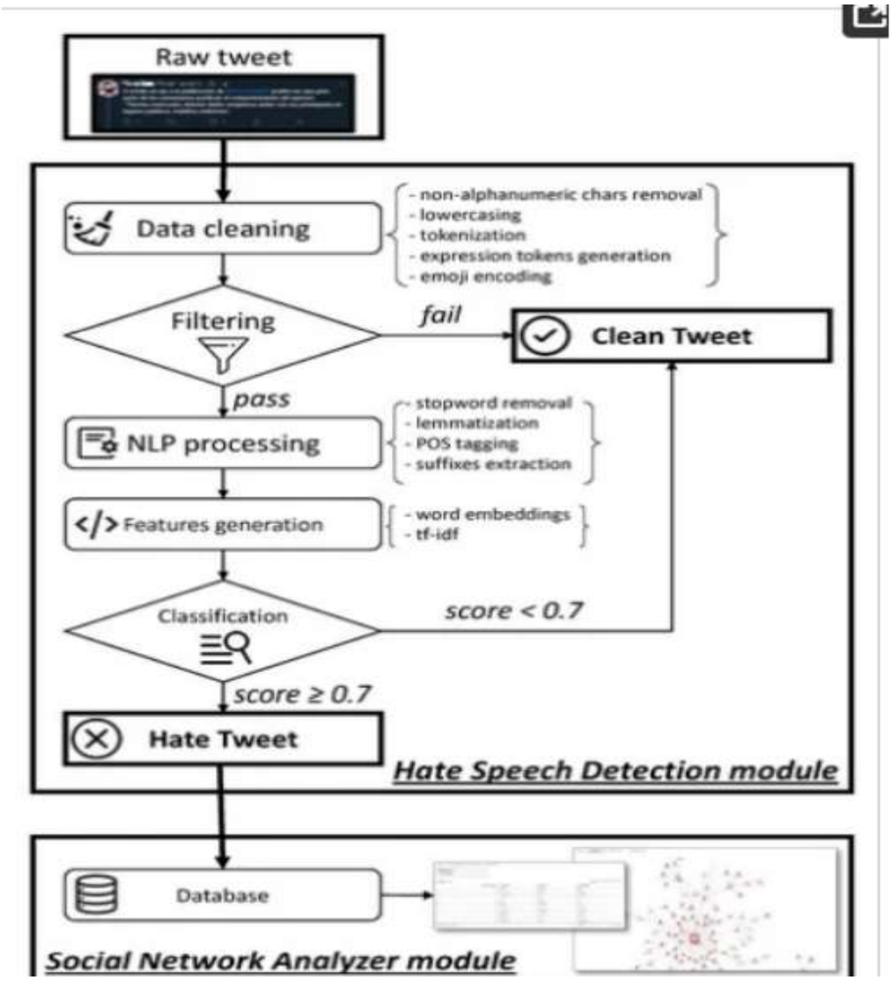
This section examines the characteristics of the data collected for this research. This comprises of data gathered with Jigsaw. A dataset of edits to Wikipedia's discussion page comments is also used. Jigsaw analyses Wikipedia comments (toxic or non-toxic) and makes the dataset available to those who want to work on the study further. The contribution of Jigsaw is to develop and show a method for analysing personal attacks that combines crowd sourcing and machine learning. This part also discusses the frequency-inverse document frequency (TF-IDF) method, as well as text mining and text processing. Confusion measures are used to evaluate the model.

We have three levels in ANN: input layer, hidden layer, and output layer.

- Input Layer: The input layer of a neural network consists of artificial input neurons that bring in the original data to be processed by following layers of artificial neurons. The procedure of the artificial neural network starts with the input layer.

- Hidden Layer: A hidden layer in neural networks exists between the algorithm's input and output, in which the function adds weights to the inputs and transmits them through an activation function as the output. In effect, the hidden layers execute nonlinear transformations on the network's inputs.

- Output Layer: The final outcome will be produced by the output layer. A neural network must always have one output layer. The output layer gathers data from the layers that come before it, uses its neurons to perform computations, and then computes the result.



An ANN is composed of hundreds or thousands of artificial neurons known as processing units, which are linked together by nodes. Input and output components are present in these processing systems. Based on an internal weighting system, the input units receive various forms and structures of information, and the neural network tries to learn about the information given to create one output report. Backpropagation, an abbreviation for backward propagation of error, is a collection of learning principles used by ANNs to perfect their output results.

An ANN begins with a training phase in which it learns to identify patterns in data, whether visually, audibly, or textually. During this supervised phase, the network compares its actual output to what it was supposed to generate. Backpropagation is used to compensate for the difference in results. This means that the network works backward, from the output units to the input units, adjusting the weight of its connections between the units to minimize the difference between the actual and planned outcomes.

Artificial neural networks are constructed in the same way that the human brain is, with neuron nodes linked in a web-like pattern. The human brain contains hundreds of billions of synapses. Each neuron is made up of a cell structure that is in charge of information processing by transporting data to and from the brain. (inputs and outputs).

4. IMPLEMENTATION:

Deep learning algorithms can be used to categorize remarks into five groups. The ANN algorithm is used to classify the comments.

The following are the techniques used for this classification:

a. Model architecture definition

b. Learning method configuration

c. Model Training

d. Model Testing

e. Prediction

Step 1: Model Architecture Definition

To classify comments, we employ an artificial neural network. This is an important step in the evolution of our deep learning model. The following operations are required to define how our model will look.

Importing Libraries:

We import all of the required libraries from the Keras module to perform ANN operations.

Creating the model:

Keras offers two ways to define a neural network:

- Sequential
- API for Functions

The Sequential class is used to describe linear network layer initializations, which are then combined to create a model. In the example below, we will create a model using the Sequential function Object() [native code] and then add layers to it using the add() technique.

Step 2: Learning Method Configuration

After specifying the training data and model, it is time to setup the learning process. This is accomplished by invoking the compile() function of the Sequential model class. For compilation, an optimizer, a loss function, and a set of metrics are needed.

Step 3: Train the model

We now have training data and a fully setup neural network to train with. All that is left is to input the data into the model and begin the training process, which is finished by iterating on the training data. The fit() technique is used to initiate training.

Step 4: Save the Model

Your model will be saved for future use. This saved model can also be used to categorize something in an android or web program.

Step 5: Forecasting

The final and most important step is to perform classifications with the Saved model. The load model method is used to import the model. We use the imread() method from the OpenCV library to receive a Comment.

5. RESULT AND OBSERVATIONS:

The Artificial Neural Network technique and the comments dataset are used to build the Toxic comments classification model. This model divides 30000 attributes into 6 groups, with 80 percent serving as the training set and the remaining 20% serving as the test set. The training set is used to train the model by defining the model design, configuring the learning process, training the model, testing the model, and making predictions. The model trained under the a forementioned conditions obtained a scale accuracy of 90%. The user interface is created with the Flask API to provide the user with an interactive UI. In this interface, the user can type a message and categorize the nature of the remark.

6. FUTURE SCOPE:

In the future, other machine learning models can be used to improve performance by determining accuracy, interference loss, and loss tracking. Deep learning techniques we examine include CNN, SVM, and TF-IDF. As a result, we can investigate a number of alternative strategies to aid in the improvement of the final output.

7. CONCLUSION:

This research intends to use logistic regression to develop a model that can automatically classify a remark as toxic or non-toxic. to develop a multi-headed model for identifying Threats, obscenity, insults, and identity hate are all examples of toxicity. A multi-headed model will identify different types of toxicity by collecting and pre-processing toxicity classified comments for training and testing, using logistic regression to train the dataset and confusion metrics to assess the model. The paper recommends four models for classifying abusive online comments: one logistic regression model and three neural network models, one of which is an artificial neural network. The combined model was found to be the most effective, with the greatest accuracy: 0.9820 and 0.9645 when tested on 0.1 and 0.33 of the training data set, respectively. All models are written in Python 3 and can be used to extract and categorize comments from social media sites based on user parameters.

8. REFERENCES:

1. The article "A comparison of deep learning models and word embeddings for detecting toxicity in online textual comments," appeared in Electronics, vol. 10, no. 7, pp. 779, in 2021. Diego Recupero, Harald Sack, and Danilo Dess.

2. Challenges for Toxic Comment Classification: An In-depth Error Analysis, Betty Van Aken et al., 2018.

3. Machine learning tools for online toxicity detection David Noever, 2018.

4. 4th International Conference on Intelligent Computing and Control Systems, Madurai, India, May 2020, Machine Learning Algorithms for Classification of Online Toxic Comments, pp. 1119–1123 (ICICCS). The group's members are Rahul, H. Kajla, J. Hooda, and G. Saini.

5. Spiros V. Georgakopoulos et al., "Convolutional neural networks for the classification of toxic comments," Proceedings of the 10th Hellenic Conference on Artificial Intelligence, 2018.