International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

TOXIC COMMENT CLASSIFICATION ON SOCIAL MEDIA USING NLP

[1*] RAMA DEVI.G ,[2*]P.VARSHITH, [3*] R.LAXMI KAMAL, [4*]S.SONIKA, [5*]P. SAI SHIVANI

[1] *Assistant Professor,* [2,3,4,5] *B.Tech Final Year*

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NALLA MALLA REDDY ENGINEERING COLLEGE

DIVYANAGAR, HYDERABAD, INDIA.

[1] ramadevi.cse@nmrec.edu.in

*Abstract-*As online content keeps to develop, so does the spread of hate speech. It is the state of abusive or threatening speech or writing that expresses prejudice towards a specific group, specifically on basis of race, religion or sexual orientation. That is affecting absolutely everyone and might be tough for them to tolerate anymore. So, we build an software to be able to discover those abusive feedback that's based totally on NLP (Natural Language Processing). NLP allows computer systems to apprehend natural language as people do. Whether the language is spoken or written, natural language processing makes use of synthetic intelligence to take real world input, procedure it, and make experience of it in a way a PC can apprehend. In this assignment we try to acquire comments from various posts on social media and then try to classify the ones comments as toxic or non-toxic using NLP. After classifying we output a listing of profiles which have published toxic remarks on the publish in conjunction with the remark. This listing can similarly be used by another device or software to delete offensive feedback as well as block the customers posting such remarks.

*Keywords-*Toxic and non Toxic Comments, NLP (Natural language Processing),

I. INTRODUCTION

With the rapid development of online media platforms, an increasing number of people participate in the sharing and dissemination of social data. Recent years have witnessed a surge in the number of toxic messages on various social platforms. The emergenceof buzzwords also makes online media an ideal place To post toxic comments, which generally refer to hate speech, insults, threats, vulgar advertisements, and misconceptions about political and religious tendencies. These spam messages have severely

affectedusers' browsing experience on the platform and hindered the healthy development of social platforms. Therefore, it is crucial to conduct research on the identification of toxic comments to filter and clean the Internet environment.

It has become evident that human behaviour is changing; our emotions are getting attached to the likes, comments and tags we receive on social media. We get both good and bad comments but seeing hateful words, slurs and harmful ideas on digital platforms on a daily basis make it look normal when it shouldn't be. The impact of toxic comments is much more catastrophic than we think. It not only hurts one's self-esteem or deters people from having meaningful discussions, but also provokes people to such sinister acts. Therefore, having a solid toxicity flagging system in place is important if we want to maintain a civilized environment on social media platforms to effectively facilitate conversations.

Our main intent is to classify content in order to detect the abusive comments. So we try to collect comments from various posts on social media and then try to find the toxicity of those comments using NLP.

II. LITERATURE SURVEY

There are several works that are related to clasiification of toxic comments where we are using NLP (Natural language processing) to classify Using toxic comment classification.[1] we can restrict users from making statements that belong to hatred, physical/verbal hard, harassment, racism, sexism, etc. By eliminating such comments and users, social media becomes a little less toxic and doesn't affect others. Many machine learning algorithms require the input to be represented as a fixed-length feature vector.In this paper, they propose Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Although the focus of this work is to represent texts, their method can be applied to learn representations for sequential data. In non-text domains where parsing is not available, we expect Paragraph Vector to be a strong alternative to bag-of-words.[2] The number of Facebook users is increasing every year, but the platform has its own merits and demerits. Cyberbullying and cybercrime through abusive language in Facebook comments are increasing, making people's living conditions miserable. Machine learning and deep learning can be used to detect cyberbullying and cybercrime. There is a lack of approaches using Bengali corpus and transliterated Bengali corpus for this purpose. A translated Bengali corpus was created with 3000 Facebook comment detetive abusive content. Supervised machine learning requires a labelled training dataset. The paper presents various algorithms for detecting abusive content in Facebook comments.[3] In this paper, they have conducted experiments on three datasets, IMDB movie reviews, Amazon product

reviews and SMS spam detection dataset. After performing sentiment analysis on these datasets using binary bag of words model and TF-IDF model they found out that the accuracy is low. But, after conducting experiment with their proposed model, i.e, using next word negation with TF-IDF, they found out a good increase in the accuracy level. So, from their experiments, they have concluded that when TF-IDF model is coupled with Next Word Negation then the performance of the sentiment classifier increases by a good percentage.[4] in this paper. We use Lexicon-based approach uses a predefined dictionary to determine the sentiment inclination of textual data.Machine learning-based techniques require a training dataset to classify input data. Hybrid techniques combine both lexicon-based and machine learning-based approaches. Support Vector Machine (SVM) is a widely used machine learning algorithm for classification. Precision, Recall, and F-Measure are used for performance evaluation of SVM. The paper is organized into sections for related work, materials and methods, classification, results, and conclusion.[5] Deep Learning models are prevalent due to their superior modeling capabilities. This work aims to assess the predictive gains of Deep Learning models and the impact of text embedding methods and data augmentation techniques. Shallow and Deep Learning models achieved competitive accuracy scores in the study. Fine-grained refinement of models or data augmentation techniques may only provide a narrow improvement margin.[6] The paper discusses different machine learning and deep learning models used for text classification and Natural Language Processing. Logistic Regression, Support Vector Machine Models with TF-IDF Vectorizer, Long Short-Term Memory with Glove and Word2Vec Embedding are used in the study. Results and analysis of the different models are presented, and the paper concludes with a summary of the findings.[7] It unfairly assigns higher toxicity scores to comments containing words referring to the identities of commonly targeted groups due to disrespectful references in training data. Marginalized groups' comments referencing their identities are often mistakenly censored. Unintended bias needs to be addressed and mitigated. Several toxicity classifiers have been constructed to reduce unintended bias while maintaining strong classification performance.[8] The paper aims to apply NLP and deep learning to determine whether a comment should be classified as abusive. Linear regression and deep learning models are used to determine the most salient features of abusive comments. The paper aims to answer the question of whether comments can be automatically moderated based on their attack scores.

The rest of the paper includes sections on relevant research, approach, experiments, and recommended next steps for future work[9] The project will implement various deep learning models, including MLP, LSTM, and CNN. The models' performance will be assessed on both binary and multi-label classification tasks. The

project will study the application of these models at both word-level and character-level granularities.[10] The paper discuss about There are two main techniques for text classification: supervised and unsupervised learning techniques. In supervised learning techniques, a previously classified dataset is used to train the machine learning algorithm. Multinomial Naïve Bayes, Max Entropy Random Forest, and Linear Support Vector Machines are popular algorithms for text classification. Three datasets were used in the study: the IMDB movie review dataset, Amazon Product review dataset, and SMS Spam Collection dataset.Three approaches were used to classify text: binary bag of words approach, TF-IDF, and TF-IDF with word negation. The base feature size was gradually increased to produce better results.[11] Most companies use at least partially automated content moderation systems.Deep neural networks are being explored as potential engines for detecting and controlling online verbal abuse. Current research focuses on text-based toxicity detection models, toxicity detection approaches relying on user metadata or non-text sources, and methods for pre-processing online comment data for NLP tasks related to abuse detection.[12] The paper presents a deep learning approach to classify abusive comments in social media. The authors trained a deep learning model using a dataset of comments labeled as abusive or non-abusive. They used word embeddings to represent the comments as vectors and then fed these vectors to a convolutional neural network (CNN) for classification.

III. METHODOLOGY

Data collection : Collecting a dataset of toxic comments is the first step. This dataset can be obtained from publicly available sources or manually labeled by human annotators.

Data preprocessing : Preprocessing the dataset involves cleaning and formatting the text data by removing stop words, punctuation, and special characters. It also involves tokenization, stemming, and lemmatization to convert the text data into a format that can be easily analyzed by NLP models.

Feature extraction: Feature extraction involves representing the text data as numerical features that can be used by NLP models. Common feature extraction techniques include bag-of-words, n-grams, and term frequency-inverse document frequency (TF-IDF).

Model training : The next step involves training a toxic comment classification model using the preprocessed and feature-extracted data. The model can be trained using different machine learning algorithms such as decision trees, Support Vector Machines (SVMs), and deep learning algorithms such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

Model evaluation: Evaluating the performance of the trained model involves using different

evaluation metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques can also be used to ensure the model's generalization performance.

Model deployment: Once the model is trained and evaluated, it can be deployed to classify new comments into toxic and non-toxic categories. This can be done using a web-based application or API.

IV. ARCHITECTURE

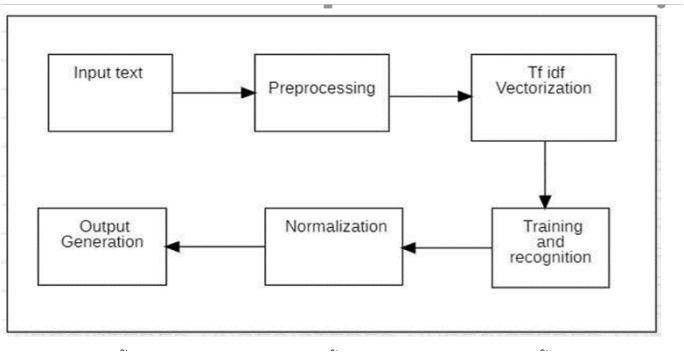

Fig – 1 System Architecture

V. MODULES

The above architecture diagram contains four stages in order to classify and detect the digits:

A. Pre-processing
B. Tf idf Vectorization
C. Training and Recognition
D. Normalization

1**. Pre-Processing**: The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical Natural Language Processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, and semantic reasoning.

Tokenization-Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. These tokens are considered as a first step for stemming and lemmatization and help in understanding the context or developing the model for the NLP. Stemming- Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. The input to stemmer is tokenized words.

Lemmatization- This is a technique which is used to reduce words to a normalized form. In lemmatization, the transformation uses a dictionary to map different variants of a word back to its root format. POS Tags- It is a process of converting a sentence to forms – list of words, list of tuples. The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on. Default tagging is a basic step for the part-of-speech tagging.

Named Entity Recognition- Named Entity Recognition can automatically scan entire articles and reveal which are the major people, organizations, and places discussed in them. Knowing the relevant tags for each article help in automatically categorizing the articles.

Chunking- Chunking is defined as the process of natural language processing used to identify parts of speech and short phrases present in a given sentence. For example, chunking can be done to identify and thus group noun phrases or nouns alone, adjectives or adjective phrases.

2**. Tf idf Vectorization**: Term Frequency — Inverse Document Frequency (TFIDF) is a technique for text vectorization based on the Bag of words (BoW) model. It performs better than the BoW model as it considers the importance of the word in a document into consideration.

3. **Training and Recognition**: In the classification and recognition step the extracted feature vectors are taken as an individual input to each of the following classifiers. In order to showcase the working system model extracted features are combined and defined using following three classifiers:

- K-Nearest Neighbor
- Random Forest Classifier
- Support Vector Machine.

4. **Normalization**:

The purpose of normalization is to transform data in a way that they are either dimensionless and/or have similar distributions. This process of normalization is known by other names such as standardization, feature scaling etc. Normalization is an essential step in data pre-processing in any machine learning application and model fitting. The need of classification of comments on social media is driven by fast-growing users with a diverse linguistic background. For Example, Most social media platforms, including Facebook, Twitter, Instagram, and LinkedIn, has huge number of users. Where those platform accept comments in the form of text documents. Users can type or paste their comments into a text box on the platform and submit them as a document. Some comments are in a mix of languages, which limits the predictive capability for models trained on monolingual datasets. Therefore, it is imperative to create a model that can capture critical semantic information from multiple languages for toxic text detection. In the classification and recognition step the extracted feature vectors are taken as an individual input to each of the following classifiers.

VI. APPLICATIONS

Social media moderation : Toxic comment classification can be used to automatically filter and remove hateful or abusive comments from social media platforms. This helps to create a more positive and healthy online environment.

Online news commenting systems : Toxic comment classification can be used to automatically filter out hateful or abusive comments on online news commenting systems. This can improve the quality of the discussion and reduce the negative impact on readers.

Customer service chatbots : Toxic comment classification can be used to train chatbots to

handle customer service interactions more effectively. Chatbots can use toxic comment classification to detect and de-escalate situations where customers are being abusive or hostile.

Brand reputation management: Toxic comment classification can be used by companies to monitor online conversations and detect any negative comments about their brand. This allows companies to respond quickly and address any concerns or issues before they escalate.

Political discourse: Toxic comment classification can be used to analyze political discourse and detect hate speech or toxic comments. This can help to identify and address issues related to hate speech in political discourse, which can have significant social and political implications

VII. CONCLUSION

In a world where technology has almost taken over and almost anybody has the right to say anything to anyone, it's very important to set some sort of boundary. Using toxic comment classification, we can restrict users from making statements that belong to hatred, physical/verbal hard, harassment, racism, sexism, etc. By eliminating such comments and users, social media becomes a little less toxic and doesn't affect others. Most people don't understand the depth of which this sort of cyber bullying can affect a person. Hence, working in this domain to be able to filter out such comments is extremely important.

VIII. FUTURE SCOPE

Text classification is a fundamental problem in NLP that involves assigning labels or tags to textual units such as sentences, paragraphs, and documents. It has a broad range of applications including sentiment analysis, spam detection, news categorization, user intent classification, content moderation, and more. With the vast amount of text data available from various sources, automatic text classification is becoming increasingly important. However, text classification poses several challenges due to the unstructured nature of text. Extracting insights from text can be time-consuming and challenging. Nevertheless, text is an invaluable source of information that can yield rich insights. In the future, advancements in NLP technologies and machine learning algorithms are expected to improve the accuracy and efficiency of automatic text classification, enabling more businesses and organizations to leverage the insights derived from text data for decision-making and other purposes.

IX. REFERENCES

[1] Q. Le and T. Mikolov “Distributed representations of sentences and documents, in the International Conference on machine learning”

[2] Darko and Drocec “Machine learning methods for toxic comment classification”

[3] “An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation”

[4] "Sentiment Analysis of Tweets using SVM, International Journal of Computer Applications"

[5] Maciej Rybinski "On the Design and Tuning of Machine Learning Models for Language Toxicity Classification in Online Platforms"

[6] Sharayu Lokhande "Multilabel Toxic Comment Detection and Classification"

[7] Elizabeth Reichert "Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers."

[8] Xuehui Jiang "A Sentiment Classification Model of E-Commerce User Comments Based on Improved Particle Swarm Optimization Algorithm and Support Vector Machines."

[9] Bijoyan Das Sarit Chakraborty "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation."

[10] Manav Kohli "Paying attention to toxic comments online."

[11] Neha Narwal "Detecting and Classifying Toxic Comments."

[12] Tanjim Mahmud "Reason Based Machine Learning Approach to Detect Bangla Abusive Social Media Comments."