IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Utilizing Machine Learning And Text Mining For Toxic Comment Classification On Social Media

[1]Anuj Kumar Pal, [2]Sakshi Rai

[1] Assistant Professor, [2]Assistant Professor

[1]Computer Science and Engineering

[1]LNCT University, Bhopal, India

Abstract: As social media usage rises, unethical activities like defamation, spreading hatred, and sharing pornography have become easier due to easy accessibility. This study examines text mining methods for better feature extraction and further employs machine learning to classify comments by toxicity. Results present a comparing among SVM, RNN and BERT. Three models were developed, trained on a Jigsaw's Kaggle dataset, and evaluated using a large set of labeled comments. The BERT model achieved an accuracy of 92.32% and an F1-score of 0.952571

Index Terms – BERT, Recurrent Neural Network, Text mining, Toxic text classification, Text classification

I. INTRODUCTION

The Internet, a major 21st-century innovation, has rapidly advanced computer science since the World Wide Web's inception in 1990 [1, 2, 3]. Initially, email was the primary communication method, but spam required the development of filtering algorithms [4]. Today, networking sites have dramatically increased internet data flow.

Social media facilitates easy information exchange in virtual networks [5]. A 2020 Hootsuite survey showed a 17% increase in internet usage in Indonesia, rising from 160 million to 175.4 million users [6]. Among these users, 88% use YouTube, 84% WhatsApp, 82% Facebook, 79% Instagram, and 56% Twitter.

Natural Language Processing (NLP) uses computational methods to analyze and represent text data, aiming to process language similarly to humans.

Hate speech is communication targeting sexual orientation, race, nationality, color, ethnicity, gender, religion, or other characteristics.

1.1 Challenges for Natural Language Processing:

Short-of-Vocabulary- A common issue in the task is the occurrence of words that are absent from the training data, such as slang, leading to a limited vocabulary

Large length Dependencies- In initial comments, toxicity is expression-dependent. With comments expanding by 50 words, this issue worsens, potentially nullifying previous parts' impact on the outcome.

Multi-word phrases- Algorithms capable of recognizing multi-word phrases as single hateful expressions can detect repetitions of such phrases and their toxicity.

1.2 Text classification problems

Computers struggle to differentiate between images and text, operating solely with binary code. Classification algorithms for online comments utilize Natural Language Processing (NLP), Data Mining, and Machine Learning techniques [7,8,9].

C. Classification problem

In supervised learning, a classification algorithm assesses training data to classify new observations. It learns from the dataset and categorizes new results into classes like 0 or 1, Yes or No, Spam or Not Spam. The main goal is to predict the output, with classifiers divided into Binary (Two Classes) and Multi-class (More than two classes).

II. LITERATURE REVIEW

Badjatiya et al. examined diverse deep learning architectures for semantic word embeddings in toxic comment classification, with extensive experiments [10]. Multiple classifiers were utilized, and prior studies have also explored similar speech analysis using neural network techniques.

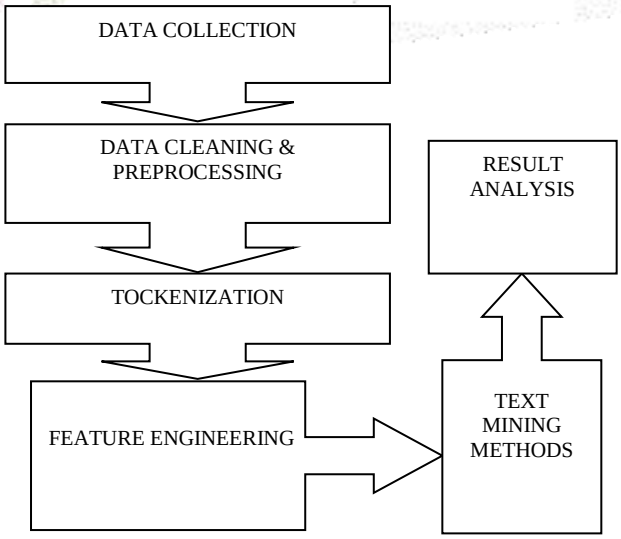
Another study introduced a deep neural network model for sentiment analysis in YouTube videos with 70-80% accuracy [11]. The author addressed issues by analyzing user comments and categorizing them based on video quality, coverage, and relevance, determining classifications as neutral, positive, or negative from viewer comments.

The author [12,13] developed a machine classification system to detect religious cyber hate in Twitter posts, utilizing diverse neural network methods to create a more generalized model.

In cyberbullying detection, supervised learning methods are prevalent, as evidenced by Farag and El-Seoud [14], as well as Karlekar and Bansal [15]. They have applied such techniques to tackle personal sexual harassment and online abuse. This study introduces the task of automatically categorizing and analyzing different forms of sexual harassment.

III. METHODOLOGY

Natural Language encompasses human communication through speech and writing in diverse languages. Unlike humans, computers communicate only in binary code, unable to grasp natural language. The data from human communication is invaluable, providing profound insights. Hence, it's crucial to develop computers that can intelligently comprehend, mimic, and respond to human speech.



The steps to perform preprocessing of data in NLP include [18, 22]:

Segmentation:- To begin, we break down the text into sentences by segmenting it into sections and removing all punctuation, including commas and periods.

Data Cleaning:- Here, we lowercase uppercase words, remove punctuation, and exclude redundant words like "was," "in," "is," and "the," which could hinder learning.

Tokenizing:- To enable the algorithm to understand sentences, we extract and explain each word independently, breaking down our statement into constituent parts and identifying each word as a token.

Lemmatization- Identifying a word's root stem involves finding its base form listed in a dictionary, from which it originates. We can also determine root words for many terms, accounting for factors like tense, mood, and gender.

Feature Engineering:- The text document was converted into a vector representation using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

III. DATASET

Jigsaw's Kaggle dataset for the Toxic Comment Classification Challenge is widely used [16], referenced in 22 primary studies. It consists of 153,164 Tweets from the Twitter API. We selected this dataset to demonstrate our approach's effectiveness in handling multi-class problems and its compatibility with Tweets, which have a unique structure due to character constraints. Features considered for results include toxicity, severity, obscenity, threats, insults, and identity-based hate.

V. MACHINE LEARNING ALGORITHMS

5.1. Support Vector Machine

SVM addresses an optimization problem via structural risk management [19, 20]. Over time, SVM has been applied in various fields and discussed in terms of decision boundaries [21]. It's extensively used in classification and regression, known for its kernel-based approach, separating data into categories using hyperplanes and a decision function [21].

It aims to maximize the margins between the hyperplane and the closest training samples. A hyperplane, depicted in Figure 02, is a pivotal method wherein learning occurs within a higher-dimensional feature space. In SVM, the hyperplane separates two classes of data points, maximizing the margin between training points, with H representing a dot product space, Xi € D denoting training points, and w indicating a weighting vector.

SVM's test phase moves slowly. A basic application of SVM might be as follows: given a training set of (m_1, n_1), (m_2, n_2),...., (m_r, n_r) with $a_i = (m_1, m_2.....m_r)$ as the input vector and n_r as the class label.

A positive class and a negative class, designated as bi€{1,-1}, would result from this. However, the linear function can be stated as follows:

$$f(m) = (w,m) + n$$
$$f(m) = +1 \quad \text{if } w.m + n > 0$$
$$f(m) = -1 \quad \text{if } w.m + n > 0$$

The decision boundary, which is denoted by the hyper-plane that divides the negative class from the positive class, can be written as (w.m) + n= 0. SVM seeks to isolate the hyperplane with the highest margin in order to reduce the error bound.

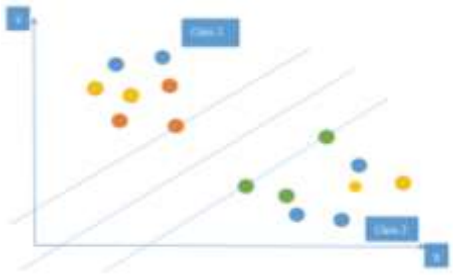


Fig-2- SVM between two classes with the highest margin hyper plane

5.2. Recurrent Neural Network (RNN)

RNNs excel in sequential information processing [22], unlike traditional neural networks where inputs and outputs are independent. Considering word order in a sentence is crucial for many tasks. RNNs derive their name from their consistent output for each input sequence, relying on past calculations [23], akin to a memory storing previous computations.

5.3. BERT (Bidirectional Encoder Representations from Transformers)

BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT is conceptually simple and empirically powerful. As a result, the pre-trained BERT model can be fine tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications. BERT alleviates the previous uni-directionality constraint by using a “masked language model” (MLM) with pre-training objective. BERT's pre-training serves as a base layer of knowledge from which it can build its responses. From there, BERT can adapt to the ever-growing body of searchable content and queries, and it can be fine-tuned to a user's specifications. This process is known as transfer learning.

VI. RESULT

This study compared four models: Logistic Regression, SVM, RNN, trained and tested on the same dataset for fair comparison. Metrics such as accuracy, recall, and F1-score were used to evaluate their performance. Accuracy measures correctly classified instances; recall estimates true positives, and the F1-score balances precision and recall. These metrics are crucial for assessing the models' accuracy in classifying instances in the test dataset.

A Confusion Matrix

Table 1- Confusion Table of SVM

True		Non-Toxic	Toxic
	Non-Toxic	0.91	0.09
	Toxic	0.29	0.71

Predicted

Table 2- Confusion Table of RNN

True		Non-Toxic	Toxic
	Non-Toxic	0.94	0.06
	Toxic	0.26	0.74

Predicted

Table 3- Confusion Table of BERT

True		Non-Toxic	Toxic
	Non-Toxic	0.96	0.04
	Toxic	0.22	0.78

Predicted

Table 4: Comparison between different approaches used for classification

Model	**F1**	**Recall**	**Accuracy**
SVM	0.938105	0.921618	86.7 %
RNN	0.931148	0.874565	87.1%
BERT	0.952571	0.847389	92.32%

As shown in Table 4, the SVM model achieved 86.7% accuracy, RNN model 87.1% and BERT model achieves 92.32% accuracy. This indicates that BERT model classifies toxic comments most accurately.

VII. CONCLUSION AND FUTURE WORK

This study explored various machine learning and natural language processing methods for classifying toxicity, revealing that poor data quality often causes errors. The BERT model achieved an F1-score of 0.952571 and 92.32% accuracy. Future work can enhance accuracy with word embeddings and more Deep Learning techniques, which could automatically select the best-fitting models for more robust and accurate classifications.

REFERENCES

[1] Mansourifar, H., Alsagheer, D., Fathi, R., Shi, W., Ni, L., & Huang, Y. (2021, September). Hate speech detection in clubhouse. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 341-351). Cham: Springer International Publishing.

[2] Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572.

[3] Zaheri, S., Leath, J., & Stroud, D. (2020). Toxic comment classification. SMU Data Science Review, 3(1), 13.

[4] Ghiassi, M., Lee, S., & Gaikwad, S. R. (2022). Sentiment analysis and spam filtering using the YAC2 clustering algorithm with transferability. Computers & Industrial Engineering, 165, 107959.

[5] Maillard, P. (2022). Object Classification and Tracking for Augmented Reality Applications (Master's thesis, ETH Zurich).

[6] S. Kemp, "Digital 2020 indonesia," 2020.

[7] Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, 3, 238-248.

[8] Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC medical research methodology, 22(1), 181.

[9] Zhao Jianqiang, Gui Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis", IEEE Access 5, 2870-2879 (2017).

[10] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760, 2017.

[11] Alexandre Ashade Lassance Cunha, Melissa Carvalho Costa, and Marco Aur´elio C Pacheco. Sentiment analysis of youtube video comments using deep neural net- works. In International Conference on Artificial Intelligence and Soft Computing, pages 561–570. Springer, 2019.

[12] Tiwari, V., Ashpilaya, A., Vedita, P., Daripa, U., & Paltani, P. P. (2020). Exploring demographics and personality traits in recommendation system to address cold start problem. In ICT Systems and Sustainability: Proceedings of ICT4SD 2019, Volume 1 (pp. 361-369). Springer Singapore.

[13] Keita Kurita, Anna Belova, and Antonios Anastasopoulos. Towards robust toxic content classification. arXiv preprint arXiv:1912.06872, 2019.

[14] Kebede, S. D., Tiwari, B., Tiwari, V., & Chandravanshi, K. (2022). Predictive machine learning-based integrated approach for DDoS detection and prevention. Multimedia Tools and Applications, 81(3), 4185-4211.

[15] Sweta Karlekar and Mohit Bansal. Safecity: Understanding diverse forms of sexual harassment personal stories. arXiv preprint arXiv:1809.04739, 2018.

[16] Conversation AI. Toxic Comment Classi cation Challenge. url: https://www.kaggle.com/c/jigsaw-toxic comment-classification-challenge.

[17] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. IEEE Access, 9, 78621-78634.

[18] Patel, A.S., Merlino, G., Puliafito, A., Vyas, R., Vyas, O.P., Ojha, M. and Tiwari, V., 2023. An NLP-guided ontology development and refinement approach to represent and query visual information. Expert Systems with Applications, 213, p.118998, 2023

[19] Shounak, R., Roy, S., Kumar, V. and Tiwari, V., October. Reddit Comment Toxicity Score Prediction through BERT via Transformer Based Architecture. In 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0353-0358). IEEE, 2022

[20] Sharma, V. K., Azad, R. K., Chowdary, V. M., & Jha, C. S. (2022). Delineation of frequently flooded areas using remote sensing: a case study in part of Indo-Gangetic basin. Geospatial Technologies for Land and Water Resources Management, 505-530.

[21] Keerthika, T. (2019). A hybrid fish–Bee optimization algorithm for heart disease prediction using multiple kernel SVM classifier. International Journal of Innovative Technology and Exploring Engineering, 8(9S2), 729-737.

[22] Kunal, S., Saha, A., Varma, A., & Tiwari, V.. Textual dissection of live Twitter reviews using naive Bayes. Procedia computer science, 132, 307-313, 2018

[23] Tiwari, V., Patel, H., Muttreja, R., Goyal, M., Ojha, M., Gupta, S., & Jain, S. (2021). Real-time soybean crop insect classification using customized deep learning models. In Data Management, Analytics and Innovation: Proceedings of ICDMAI 2021, Volume 1 (pp. 143-156). Springer Singapore.