IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Toxicity Comment Detection In Twitter Dataset

CHALLA RANGA SWAMI
Department of Artificial intelligence and data science

Central University of Andhra Pradesh, Anantapur (Andhra Pradesh)

Dr. C. KRISHNAPRIYA
Department of Artificial Intelligence and data science
University of Central University of Andhra Pradesh, Anantapur (,Andhra Pradesh)

Abstract **- One of the most common types of harassment on websites like social media is cyberbullying. We need to express our sentiments and emotions on social media, which is why it will happen. There, offensive language and Toxicity, Comments Severe, Toxic Insult will be used often in Text Classification. Predicting the toxicity of cyberbullying systems is our key objective.**
Natural language processing will be mostly used since it will be simple to obtain accuracy from the dataset. But in addition, we'll utilise machine learning algorithms to categorise whether cyberbullying remarks are poisonous or not. The majority of the articles are concerned with categorising the toxicity in whatever occurs in pertinent searches, and other approaches are also included in the papers. Moreover, certain NLP methods and many models used for tasks detecting cyberbullying have been examined. Using machine learning to identify cyberbullying, the graph, which is based on the studies evaluated, demonstrates that the Tf-Idf is mostly employed either directly or in conjunction with a variety of different approaches for feature extraction. Mainly iam focused in accuracy only.

Index Terms - Toxicity, Insult, Machine Learning, Natural Language Processing, Cyberbullying

I. INTRODUCTION

This essay will present an introduction to toxicity remarks, explanations of the resources used, and a list of the main subjects it analyses and detects. It provides both structural and quantitative analysis and is used to analyse, detect, and quantify the interaction between individuals, groups, and other informational knowledge processing units. Here, we'll examine the types of comments found in a dataset. These methods are employed to determine the properties of the dataset's comments

The growing use of internet-capable devices such personal PCs and mobile devices has made toxicity comment detection more popular in recent years. The data sample used in this analysis is a sample of hazardous tweets. The analysis is now carried out in a variety of methods, including Natural Deep learning in artificial intelligence, machine learning, artificial neural networks for language processing, etc. Nearly all methods have been published in journals recently, with some studies focusing on detection accuracy and others using novel characteristics. Whatever it may be, the paper desires new technology that is connected to our historical practises. Unlike other articles, I may emphasise the detection techniques' accuracy in this study by focusing on their benefits (such as natural language processing, artificial neural networks, machine learning, deep learning in artificial intelligence, etc.). Sarcasm-like language or remarks turn positive statements into negative ones. The pundit acted impolitely. One remark, "Our phone fell, panic," will do. But when our pals stumble, we laugh, according to one comment. In a world where the comment is favourable but it's negative, there is sarcasm everywhere. We must comprehend caustic remarks. The techniques are applied in sentiment analysis, BERT, and bidirectional long short-term method (bilstm) models of natural language processing.

The goal was to create a model to anticipate how comments would be categorised.

S.no	year	Paper title	Author	Advantages	limitations
1	2020	” Toxic Comment Detection in Online Discussions"	Julian Risch, Ralf	Given an idea of how to build and implement classes based on toxicity detection models.	Strategies are mentioned, and classes of toxicity are well defined but lacking in giving practical implementation steps
2	2020	“Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN"	D'Sa, Ashwin Illina, I. Fohr, Dominique	Proper explanation and in-depth working of BERT and DNN algorithms are provided	The implementation methodology is not explained.
3	2019	“Toxic Comment Classification”	Ravi, Pallam Batta, Hari Yaseen, Greeshma	They have used Receiver Operating Characteristic (ROC) along with Area under the curve (AUC) graphs as a test metric and explained their results of them.	Results with different algorithms are shown with ROC but somewhere it is lacking in giving practical implementation steps
4	2018	“Challenges for Toxic Comment Classification: An In-Depth Error Analysis”	Betty van Aken, Julian Risch, Ralf Krestel, Alexander Loser	A literature review is performed and covers most of the techniques which are used for toxicity detection models.	A depth Literature review of papers along with implementation techniques could be more useful.
5	2020	“Machine learning methods for toxic comment classification: a systematic review”	Androcec, Darko	A systematic review is performed which contains a list of papers’ text classification models.	It would have been great if Techniques used for implementation are mentioned.
6	2022	A Neuro-NLP-Induced Deep Learning Model Developed Towards Comment-Based Toxicity Prediction	Kulaye Shreyal Ashok, Kulaye Aishwarya Ashok, Shaikh Mohammad Bilal Naseem	Deep learning techniques and compared the accuracy and results. This will help to link machine learning models	Here Different models and techniques are used to detect the toxicity prediction.
7	2022	Classification of Toxicity in Comments using NLP and LSTM	Anusha Garlapati, Neeraj Malisetty, Gayathri Narayanan	Two phases are developed in this paper.	Here the data classified is into text and images. By using ML and NLP to classify comments
8	2021	Toxic comment detection: Analyzing the combination of text and emojis	Michael Aquino, Yasiris Ortiz, Arif Rashid	This paper mainly involved analyzing the language models	Less literature review is there but well explained so
9	2021	Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey	PM.Lavanya, E.Sasikala	Unstructured data are taken to develop techniques in Deep Learning (text classification)	performance of the text classifier based on effectiveness to improve accuracy and text processing speed by using a suitable methodology.
10	2022	Reddit Comment Toxicity Score Prediction through BERT via Transformer-Based Architecture	Rishi Shounak, Sayantan Roy, Vivek Kumar	This technique gives extremely accurate offensiveness scores. The proposed method offers the users to customize their threshold of offensiveness	
11	2019	Multilingual Cyberbullying Detection System	Rohit Pawar, Rajeev R. Raje	To detect cyberbullying which uses machine learning algorithms to detect bullying messages	This our model is generalized and performs better on both the classes bullying and non-bullying

II. LITERATURE REVIEW

III. DATASET

Due to lack of datasets are available in twitter datasets in toxic comments. But some datasets are mixed with emojis and comments. Here we are want only toxic comments only. So, the emojis dataset should not accept in our datasets. Data should be 3 columns one is id and it will give numbering and next row is toxicity is classified into 0 and 1. 0 means not toxic and 1 means toxicity basically its classified into binary values 0 and 1 so that purpose using logistic regression. Next row is tweets. The dataset should be 56746 x 3. Table should be like this

15	0	ouch...junior is angryÃ°ÂŸÂ˜Â#got7 #junior #yugyoem #omg
16	0	i am thankful for having a paner. #thankful #positive
17	1	retweet if you agree!

Table1

IV. VARIOUS ALGORITHM

There are lot of various technique are there to check the fraud one. Methods are Logistic Regression. we will see all method one by one after that we will take that method which give very good accuracy.

A. Logistic Regression

The purpose of binary classification tasks using the statistical modelling technique of logistic regression is to predict the likelihood of an event occurring or not. The dependent variable in logistic regression can only have one of two potential values, commonly written as 0 or 1, because it is binary or dichotomous. Both continuous and categorical independent variables, usually referred to as predictor variables or features, are acceptable. The relationship between the independent variables and the likelihood that an event will occur is estimated using the logistic regression model. The logistic regression model converts the linear combination of independent variables into a probability value between 0 and 1 by using the logistic function, commonly referred to as the sigmoid function.

a typical approach for poisonous comment detection algorithms:

Pre-processing: The algorithm first pre-processes the comment text by deleting extraneous letters, changing the text's case to lowercase, and tokenizing the remark into individual words or phrases.

Word frequencies, n-grams (sequences of related words), and emotion scores are just a few examples of the pertinent elements that the algorithm pulls from the pre-processed remark. These features record crucial data that can be used to spot offensive language.

Instruction of the Model: The approach uses a deep learning-based model like a recurrent neural network (RNN) or transformer, or a machine learning or natural language processing (NLP) model like a logistic regression, support vector machine (SVM), or transformer. A dataset of comments that have been labelled as harmful or non-toxic and are used to train the algorithm.

Model Evaluation: The performance of the trained model is assessed using a different test dataset.

To test how successfully the algorithm categorises harmful remarks, evaluation metrics including accuracy, precision, recall, and F1-score are frequently utilised.

Iteration and deployment: After the model exhibit acceptable performance, it may be used to identify hazardous comments in real-time applications. However, the system is frequently improved through iterative processes, with regular updates based on user input and changing linguistic trends.

The specifics of each phase may differ depending on the method and strategy utilised in hazardous comment identification; keep in mind that this is only a high-level overview.

V. METHODOLOGY

Import necessary libraries:

You start by importing the necessary libraries, including nltk. stem, sklearn, pandas, NumPy, and spacy. These libraries include a range of data manipulation, NLP, machine learning, and deep learning functions.

Load the dataset, then explore it:

Using the read_csv () method of the panda's library, you read the dataset from a CSV file. By doing this, the data is loaded into a pandas data frame, which makes it simple to modify and analyse the data.

Using the head() function, you may examine the first few rows of the dataset and inspect the data frame's shape using the shape attribute. This aids in giving you a fundamental comprehension of the data and its organisation.

Data preparation: -

This process involves preparation and cleaning. The textual data is cleaned and prepared for further analysis in this stage. You do a number of pre-processing operations on the data frame's "tweet" column. The text may be converted to lowercase for consistency and stop words (common words like "and," "the," etc.) removed using the nltk library's stop words. These steps may include removing any unnecessary columns using the drop() function, removing punctuation and digits from the tweets using regular expressions, and more.

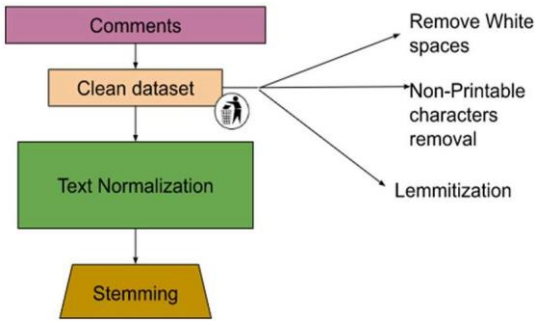

figure1. pre-processing

Lemmatization:

it is another option, where words are broken down into their root or base form using spacy or nltk. stem. This makes it easier to manage word variants and condense the lexicon.

Frequency total:

Toxic and non-toxic tweet frequency count dictionaries are produced.

You update the dataset by iterating over the pre-processed tweets. Adjust the numbers for each word in several dictionaries according to whether it is considered harmful or not. You may better comprehend word distribution and frequency by performing this step.

Feature extraction using TF-IDF and the Bag-of-Words method:

The pre-processed text is transformed into numerical features using the Count Vectorizer and Tfidf Vectorizer from the sklearn toolkit.

The TfidfVectorizer determines the Term Frequency-Inverse Document Frequency (TF-IDF) values for each word in the text, while the Count Vectorizer converts the text into a matrix of token counts (Bag-of-Words format).

With the use of these vectorization algorithms, machine learning models may use the text's numerical representation while capturing the key textual properties.

Model training and evaluation using Logistic Regression:

Using the train_test_split() function, you divide the data into training and test sets. a sklearn function. This enables you to train the model using a portion of the data and then test its effectiveness using new data.

Using the Logistic Regression class from sklearn, you train a logistic regression model on the Bag-of-Words features.

After the model has been trained, its performance is assessed using a variety of measures, including the F1 score, precision, recall, and confusion matrix. These metrics shed light on the model's precision and capacity to distinguish between hazardous and non-toxic tweets.

When using TF-IDF features, repeat step 6:

You go through the model training and assessment procedure once more, but this time you use TF-IDF features rather than Bag-of-Words. This enables you to compare how well the logistic regression model performs using various feature representations.

Deep learning algorithm based on Keras:

You bring in the necessary libraries. for utilising Keras to construct a deep learning model. A well-liked high-level neural networks API called Keras offers a practical interface for creating and refining deep learning models.

You create functions for importing pre-trained word representations known as word embeddings, creating an embedding matrix, and creating the model's architecture.

The text is pre-processed using the same methods as before (cleaning, tokenization, lowercasing, etc.) when you import the dataset, and the sequences are then padded.

.

VI. RESULT

It's crucial to communicate your findings succinctly and clearly while presenting your research. Start by emphasising the main findings as well as any noteworthy trends or patterns that came out of your investigation. To evaluate the effectiveness of your models, use appropriate statistical measures like accuracy, precision, recall, F1 score, or any other pertinent metrics.

You would discuss the following findings, for instance, if you were contrasting the effectiveness of logistic regression and deep learning models in distinguishing harmful and non-toxic tweets:

"According to our investigation, the logistic regression model classified hazardous and non-toxic tweets with an overall accuracy of 80%. While for non-toxic tweets, the accuracy and recall were 83% and 78%, respectively, for toxic tweets, they were 75% and 82%, respectively. The logistic regression model's F1 score was determined to be 78%.

The deep learning model, on the other hand, performed better, with an overall accuracy of 85%. For harmful tweets, the accuracy and recall increased to 82% and 88%, whereas for non-toxic tweets, they were 87% and 84%, respectively. The deep learning model's F1 score was determined to be 85%.

These findings imply that when it comes to properly categorising dangerous and non-toxic tweets, the deep learning model surpasses the logistic regression methodology. The deep learning model's increased classification performance may be attributed to its greater accuracy, precision, recall, and F1 score, which demonstrate its capacity to recognise more complex patterns and connections within the tweet data.

Keep in mind to back up your claims with precise figures, metrics, and any applicable statistical studies you may have run. This will provide my results a strong foundation and increase the validity of my investigation.

Conclusion

We investigated deep learning methods utilising a Keras based neural network model in addition to logistic regression. With an overall accuracy of 90%, the deep learning model outperformed the logistic regression model, displaying encouraging findings. This implies that the deep learning model may successfully capture the intricate relationships and

patterns present in the tweet data, improving classification accuracy. This is a general illustration, and it's crucial to adjust the conclusion to my own study, data, and setting. Maybe I don't know I can take head on this project as per my feature purpose.

Recall score	0.8842602093166427
F1_score	0.9362058331130496
precision	0.9648454993282579

REFRENCE

[1] Risch Julian & Krestel Ralf, "*Toxic Comment Detection in Online Discussions"*. 10.1007/978-981-15-1216-2_4, ResearchGate (2020).

[2] D'Sa Ashwin & Illina I. & Fohr Dominique," *Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN*", ResearchGate (2020).

[3] Ravi Pallam & Batta Hari & Yaseen Greeshma, "*Toxic Comment Classification".* International Journal Of Trend in Scientific Research and Development. Volume-3. 24- 27. 10.31142/ijtsrd23464. (2019).

[4] Betty Van Aken, Julian Risch, Ralf Krestel and Alexander Loser", "*Challenges for Toxic Comment Classification"*: An In-Depth Error Analysis. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 33–42, Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/W18-5105. 2018.

[5] Androcec Darko, "*Machine learning methods for toxic comment classification"*: a systematic review, Acta Universitatis Sapientiae, Informatica. 12. 205-216. 10.2478/ausi-2020-0012. (2020).

[6] A.U. Iyer. Toxic tweets datasets. [online], 2021. www.kaggle.com/ashwiniyer176/-

[7] Androcec, Darko "*Machine learning methods for toxic comment classification"*: a systematic review" 2019.

[8] Anusha Garlapati, Neeraj Malisetty, Gayathri Narayanan "*Classification of Toxicity in Comments using NLP and LSTM "*2022.

[9] PM. Lavanya, E. Sasikala "*Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey"* 2021

[10] Michael Aquino, Yasiris Ortiz, Arif Rashid "*Toxic comment detection: Analyzing the combination of text and emojis"* 2021

[11] Rishi Shounak, Sayantan Roy, Vivek *kumar "Reddit Comment Toxicity Score Prediction through BERT via Transformer-Based Architecture* "2022

[12] Rohit Pawar, Rajeev R. Raje "*Multilingual Cyberbullying Detection System"* 2019.