2nd International Conference on
"RECENT ADVANCES IN MATERIALS, MANUFACTURING AND MACHINE LEARNING (RAMMML-2023)"
02ND - 04TH FEBRUARY- 2023
Department of Mechanical Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, India

# Toxic Comment Identification using NLP

Prof. Kalyani Pendke[1], Mandeep Sharma[2], Raunak Palewar[3] ,Sudeep Roy[4], Rohan Malhotra[5], Sameer Ambekar[6]
[1, 2, 3, 4, 5, 6]*Department of CSE, Rajiv Gandhi College of Engineering and Research, RTM Nagpur University, Nagpur, Maharashtra, India*

***Paper ID: RAMMML-369***

## Abstract

Toxic comments are the comments found in the online forums that are rude, offensive, or unfair and usually cause many users to exit the conversation. The prospect of cyberbullying and abuse restricts people's access to alternative opinions, preventing a free interchange of ideas. As per this survey of different papers the author conclude that among those papers, the techniques which were used are binary relevance, classifier chain and label power of machine learning algorithm. Among these, the binary relevance is found to give a bit higher accuracy as compared to other techniques which were used. Also, further experiment with more complex deep learning algorithms can be used to get high performance which will be beneficial for the future researchers to develop their systems more efficiently and precisely.

## Introduction

One of the best inventions of the twenty-first century is that one person can connect with another person anywhere in the world using only a smartphone and the internet, thanks to the rapid growth of computer science and technology [1] .
Comments that are rude, disrespectful, or have a tendency to make users leave the discussion are referred to as toxic comments. If these toxic comment can be automatically identified, they could have safer discussions on various social networks, news portals, or online forums. Manual moderation of comments is costly, ineffective, and sometimes infeasible. Different machine learning techniques, primarily various deep neural network architectures, are used to automatically or semiautomatically detect toxic comments [9] .
Adults can manage social media abuse to some extent, but children and teens are susceptible to serious mental health issues. Some people abuse the freedom of speech and expression provided by social media platforms to flood these platforms with offensive content. Adults can control this abusive behaviour, but children and teenagers are even more vulnerable.
There has been a 70% increase in bullying and hate speech among children and teenagers since the Covid-19 shutdown. There has been a 70% increase in bullying and hate speech among children and teenagers since the Covid-19 shutdown. With regard to social and ethical issues, there has been an increase in interest in the communities of artificial intelligence and natural language processing as a result of the global commitment to combat toxic content. This harmful content is characterised by hate, offensiveness, abuse, cyberbullying, violence, and other forms of harassment online.
To limit toxic content, most social media platforms use content moderation. Thus, require unbiased and scalable systems to identify harmful content in real-time due to the enormous scope of online content. If these systems can pinpoint the section of text that qualifies the content as toxic, people will start to trust them. To encourage positive conversations among people, toxicfree social media platforms are required [8].
To identify toxic comments, variety of identification methods and machine learning algorithms can be use on the dataset.

## Methodology

Toxic comment identification has been extensively studied in recent years, especially in the context of social media, where researchers have used various machine learning algorithms to classify toxic comments found on social media forums into different toxic classes [1]. Determine the toxicity of a user's comment based on the comment itself. The objective is to develop a classifier model that can foretell whether input text is improper (toxic).

A. Machine Learning:
Artificial intelligence (AI) systems can automatically learn from their experiences and get better over time thanks to a technique called machine learning. The development of computer programmes that can access data and use it to learn for themselves is the focus of machine learning [23] .
*B. Deep Learning:*
Deep Learning is a specialized form of machine learning (fig 1). A machine learning workflow starts with relevant feature being manually extracted from images, the features are then used to create a model that categorizes the object in the image [24] .
*C. Sentiments Analysis:-*
Sentiment analysis, also referred to as opinion mining, is a simple method for ascertaining the author's emotions in a text. What the author was trying to say when they wrote it (fig 2). It use a range of text analysis and natural language processing (NLP) tools to determine what information might be subjective [25] .
*D. NLP:*
Natural Language Processing, or NLP for short, is a subfield of computer science, humanities, and artificial intelligence (fig 3). Machines can comprehend, analyse, manipulate, and interpret human languages thanks to technology. It aids in the organisation of knowledge for tasks like topic segmentation, relationship extraction, named entity recognition (NER), speech recognition, automatic summarization, and translation [22] .

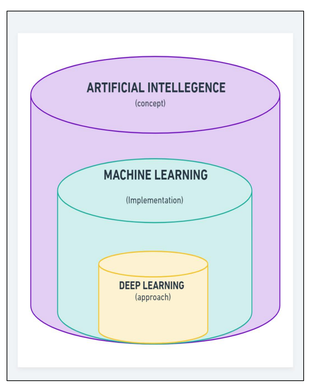

Fig 1

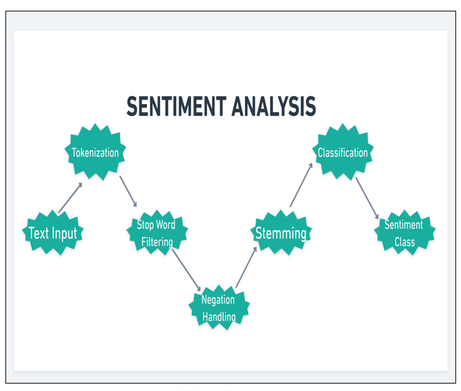

Fig 2

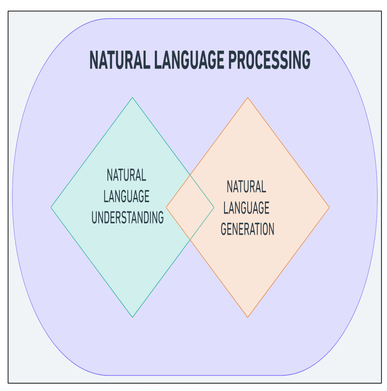

Fig 3

## Results and Discussions

| Reference | objective | Data Set | Technique used | Future Scope |
|---|---|---|---|---|
| Abhishek Aggarwal (2021) [7] | To determine the toxicity as precisely as possible. | The dataset was provided by Kaggle. | 1)Multinomial NaïveBayes 2)Random Forest Classifier 3)Bernoulli NaïveBayes 4)Nearest Centroid 5)Ridge Classifier | Accomplish multilabel classification directly using algorithm adaptation techniques also, experiment with more advanced deep learning algorithms. |
| Prof. Kiran Babu (2021) [8] | Identifying toxic spans or rationales and Toxic XLMR for bidirectional contextual embedding's. | *Kaggle-Jigsaw *TSD *Curated dataset | Multi-task neural network model, which jointly learns on the sequence classification and span prediction tasks. | To create models employing semantic embeddings that take more delicate context and text actors into consideration. |
| Prof. Darko Androcec (2020) [9] | Systematic review of toxic comment classification using machine learning methods. | Jigsaw's data collection, by Kaggle. | The systematic literature review (SLR) methodology was used. | Using transformers for the classification of harmful comments in future works. |
| Julian Risch (2021) [10] | Software solution that automatically downloads, processes, and shows a collection of more than forty datasets in a common data format. | Jigsaw's data collection by Kaggle. | GitHub repository1 and PyPI package2. | To promote the repeatability and reproducibility of toxic comment research. |

## Conclusions

The toxic comment identification helps to provide the way to tackle the problems related to increased hetaerism and toxicity on the current social media platforms. Because these things can imperil the way people participating in communication at social networks, blogs, forums, etc. The goal of this survey is to identify those harmful comments from various platforms.
Thus, the ultimate goal for this particular survey is to explore the scope of online abusive comments and categorize them into different labels to assess the toxicity as well as the accuracy, by using various machine learning algorithms combined with Natural Language Processing and compare their performance. This survey demonstrated a multi-task model that performs toxic comment identification while predicting the toxic spans as rationales. In many papers, researchers basically opted for three methods like "Binary Relevance, Classifier Chain, and Label Power "of machine learning algorithm in which Binary Relevance method gives the high accuracy. Thus, reducing the threat of bullying and abuse on the internet obstructs the free exchange of ideas by limiting people's opposing viewpoints. This identification is very helpful for reducing the toxicity from different comments from various social media handles does leading to free and toxic free exchange of comments between the people using the social Medias.
In the future, a more reliable algorithm might be approached by classifiers using the Grid Search Algorithm. Further, we can also do experiment with more complex deep learning algorithms to get high performance.

## References

[1] Agarwal A., Xie B., Vovsha I., Rambow O., and Passonneau, : " Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM", Association for Computational Linguistics United States, 2011
[2] Prof. Kiran Babu Nelatoori Prof. Hima Bindu Kommanti, "Multi-task learning for toxic comment classification and rationale extraction", Journal of Intelligent Information Systems, 2021
[3] Prof. Darko Androcec, " Machine learning methods for toxic comment classification: a systematic review", Acta Univ. Sapientiae Informatica, 2020
[4] Julian Risch, Philipp Schmidt, Ralf Krestel, "Data Integration for Toxic Comment Classification:" , Association for Computational Linguistics , 2021