A Data Mining approach for Classification of Toxic Comments and Empowering Positive Discourse

Deeraj Nair *deenair@iu.edu*
Malhar Dhopate *mdhopate@iu.edu*
Rishit Puri *rispuri@iu.edu*

project-deenair-mdhopate-rispuri

Abstract

In an era of ubiquitous online communication, hate speech and toxic remarks have attracted a lot of attention. This project analyses this issue through data mining and machine learning algorithms. The primary focus of this project is to classify toxic comments with a high recall - as it is important to correctly identify toxic comments. We aim to identify the unfavorable content by utilizing natural language processing and sentimental analysis models. Additionally, based on these models we developed a recommendation system which promotes the users to opt for positive discourse. In addition to being a technical undertaking, this project is aimed towards fostering a more positive, respectful and inclusive online environment.

Keywords

Logistic Regression, Support Vector Machine (SVM), Sentimental Analysis, Classification, Recommendation, Neural Network, Tokenizer.

1 Introduction

The digital age has brought an unprecedented level of connectivity and communication through online platforms. In-spite of opening avenues for interactions, these platforms have become incubators for toxic and harmful comments. The presence of such comments disrupts meaningful conversations, threatens the emotional and psychological well-being of an individual and leads to a hostile online environment.

In response to this problem - instead of only identifying a toxic comment, our project aims to develop a reliable system to proactively recommend the users to revise their comments to foster a healthier and respectful online environment. This will be achieved by utilizing a variety of data analysis techniques, including sentiment analysis and machine learning models like Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, Neural networks, and Natural Language Processing (NLP) models.

The project makes use of sentiment analysis to determine the emotional tone and purpose underlying the remarks. By integrating various machine learning model outputs, with sentiment analysis insights the study seeks to advance the creation of efficient comment moderation system for online platforms.

Previous work

There have been several studies conducted to classify the toxic comments with a variety of machine learning models, and there have been several studies to determine the effects of toxic comments on the online environment and an individual's mental and emotional health.

The report [1] by Etibar Aliyev presents a practical implementation of a neural network-based model for toxic comment classification. The model is trained to classify comments into six different categories of toxicity: toxic, severe toxic, obscene, threat, insult, and identity hate. The model is compiled with the Binary Cross entropy loss function and the Adam optimizer. The model predicts the probabilities of each toxicity category for the input text. A threshold of 0.5 is applied to convert the probabilities into binary predictions and have been evaluated using precision, recall and accuracy metrics.

[2] aims to understand the effectiveness of deep learning models compared to machine learning models along with the most common models used by researchers in the last 5 years. The paper has also provided insight on the most common data sets utilized by researchers to detect toxic comments. To achieve this, they have compiled the datasets of research papers and analyse the algorithm used. The findings indicate that Long Term Short Memory is the most routinely mentioned deep learning model with 8 out of 26 research papers. There have been attempts to combine more than one deep learning algorithms, however these hybrid models might not result in a better accuracy than an original model. In conclusion LSTM has better accuracy with highest being at 97 percent compared to other models. GRU also yields good accuracy while logistic regression having the highest accuracy of 96 percent. However, neural networks with TF-IDF pre-processing has the best accuracy. The most frequently used datasets are from Kaggle. Specifically, Wikipedia's talk page edits datasets that have over 150000 samples with a total of 13 out of 26 research papers using this particular datasets.

[3] identifies the types of online users' by analyzing their online activity in-terms of the comments posted. This research introduces two metrics, the F score and G score, to identify types of toxic online user behavior. Analyzing 4 million Reddit comments, the study classifies users into four categories: Steady Users, Fickle-Minded Users, Pacified Users, and Radicalized Users based on their toxicity scores. The paper segments the users based on their toxicity scores based on their comments. The paper concluded by saying that the most toxic behavior is when a user switched constantly between toxic and non-toxic commenting. The paper concludes by saying that additional research is required to analyze as to why the observed patterns arise.

[4] has conducted studies using LSTM and Naive Bayes models to classify toxic comments and observed that the LSTM model had a higher true positive rate (20% more) compared to the Naive Bayes model. This paper also refers studies that have proven that a higher number of toxic comments are harmful to an individual's emotional and psychological health. The findings emphasize the potential of data science in creating a healthier online environment, achieving a promising accuracy of over 70% using LSTM. Additionally, the paper describes the integration of Amazon Web Service (AWS) for efficient algorithm execution.

[5] the paper "Machine Learning Methods for Toxic Comment Classification: A Systematic Review" explores the problem of toxic comment classification in online platforms. With the growing number of comments, manual moderation becomes unfeasible, so the study focuses on automatic methods for discovering toxicity using machine learning models. The authors conducted a systematic review of 31 primary studies to understand the state of the art in this field. They analyzed various aspects, including publication trends, dataset usage, evaluation metrics, machine learning methods, classes of toxicity, and comment languages. The research reveals that this area gained significant attention from 2018, with most studies being conference papers, and it is still an evolving research topic. The primary dataset used in many studies is Jigsaw's dataset, and deep neural networks are among the most effective machine learning methods. The study also identifies gaps in current research and suggests future research themes, such as the use of transformers for toxic comment classification and

multilingual toxicity detection.

[6] The paper by Julian Risch, Ralf Krestel talks about how crucial the comment sections are for online news platforms which is often misused by spammers and haters. Sentiment analysis can aid in both content moderation and understanding online discussion dynamics and help in toxicity detection.

2 Methods

Initially, the project will begin by gathering a substantial dataset of comments that are known to contain dangerous or toxic content. These remarks can be retrieved from social networking sites, forums, publicly accessible sources, pre-compiled datasets, and any other relevant source that contains such remarks. We will ensure that the dataset accurately reflects the harmful remarks that people could run into in everyday life.

Using sentiment analysis tools, the emotional tone of the gathered comments will be understood. Sentiment analysis is the process of analyzing comments to ascertain if they have a positive, negative, or neutral emotional tone using Natural Language Processing (NLP) tools and techniques. This study will assist in categorizing the remarks as offensive, hateful, or generally unfavorable. To determine the degree of toxicity in comments, the project will make use of machine learning techniques like Naive Bayes, Logistic Regression, SVM, Random Forest, Neural Networks and maybe others. The labeled dataset, which consists of comments classified as harmful or non-toxic according to their content, will be used to train these models. To assess if a remark is toxic, the machine learning models will consider many aspects of the text, including sentiment scores, keywords, and context.

The project will use recall as a measurement metrics - as we want to correctly identify toxic comments, using which the model will offer recommendation to the platform's moderators or users. A recommendation to delete or change the remark or take the required action will be given if it is determined that the comment is extremely toxic and/or contains negative emotion.

Lastly, we mine latest data of comments by using a web-scraper and test our model to determine how well it performs in a real-life implementation and its usability.

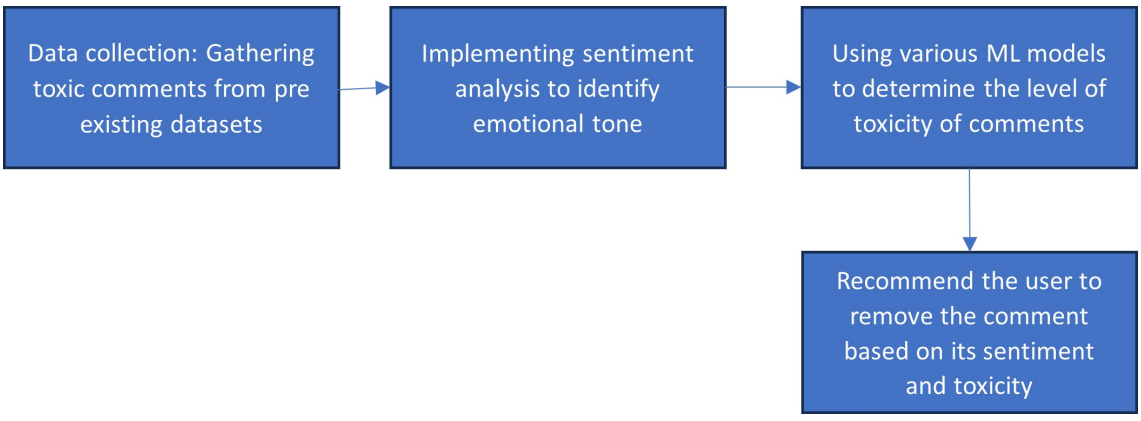

Figure 1: Methodology

References

[1] Etibar Aliyev. 2023. *Toxic Comment Classification: Identifying Harmful Content Using Recurrent Neural Networks*. medium.com

[2] Felix Museng, Adelia Jessica, Nicole Wijaya, Anderies Anderies, Irene Anindaputri Iswanto. 2022. *Systematic Literature Review: Toxic Comment Classification*. ieeexplore.ieee.org

[3] Raghvendra Mall, Mridul Nagpal, Joni Salminen, Hind Almerekhi, Soon-Gyo Jung, Bernard J. Jansen. *Four Types of Toxic People: Characterizing Online Users' Toxicity over Time*. https://dl.acm.org/doi/abs/10.1145/3419249.3420142?casa_token=O3sEJNOHJVcAAAAA:JAk37Wgzdn_2PZ1Nu0Ny802-eVUxnIJDSiowmjjbmdSaRfrBqa7NF81AChDvGz0ptegSyn_UgeFU

[4] David Stroud Sara Zaheri, Jeff Leath. 2020. *Toxic Comment Classification*. https://scholar.smu.edu/cgi/viewcontent.cgi?article=1134&context=datasciencereview

[5] Darko Androcec, *Machine learning methods for toxic comment classification: a systematic review,* researchgate.net, 2020

[6] Julian Risch, Ralf Krestel, *Toxic Comment Detection in Online Discussions*, link.springer.com, 2020