Toxic Comments Detection Using Machine Learning

Harshitha Marasu
Dept. of Computer Science
University of South Florida
harshitha54@usf.edu

Sri Harsha Mulluri
Dept. of Computer Science
University of South Florida
sriharshamulluri@gmail.com

Subhasya Tippareddy
Dept. of Computer Science
University of South Florida
subhasya.tippareddy@gmail.com

Chakradhar Reddy Nallu
Dept. of Computer Science
University of South Florida
nalluchakradhar@gmail.com

Ajay Babu Nannapaneni
Dept. of Computer Science
University of South Florida
ajayvidya11@gmail.com

***Abstract*— The spread of toxic comments on platforms like YouTube significantly impacts user interaction and mental health. This paper explores the development and application of a machine learning model based on BERT (Bidirectional Encoder Representations from Transformers) to detect such toxic comments effectively. Our approach involves extensive data gathering from YouTube to analyze toxicity in comments. The implemented BERT model processes comments in a bidirectional context to enhance the accuracy of toxicity detection. Our results indicate a high level of precision in classifying toxic versus non-toxic comments, thereby hoping to contribute to create a healthier online discourse environment. Future work will focus on expanding the model's applicability to different platforms and languages, assessing the impact of such technologies on societal norms, and exploring the integration of automated systems with human moderation efforts.**

I. INTRODUCTION

In today's interconnected world, social media platforms have emerged as pivotal spaces for both public and private interactions, profoundly influencing how individuals express themselves and engage with others. YouTube, as one of the world's foremost video-sharing platforms, hosts millions of users who actively participate through comments on videos spanning virtually every conceivable topic. This platform not only serves as a stage for sharing information but also as a community where discourse and dialogue flourish.

Despite these benefits, the platform is not immune to the pervasive issue of toxic comments—those that are harmful, offensive, or intended to cause distress. Such comments can significantly deteriorate the quality of discourse, affect user engagement negatively, and pose serious concerns for mental health and well-being. The prevalence of toxicity on YouTube has implications not just for individual users but also for the broader community dynamics and the platform's brand image.

Addressing this critical issue, our research is focused on developing a sophisticated approach to detect and mitigate the presence of toxic comments effectively. We harness the capabilities of machine learning and natural language processing technologies, specifically employing a BERT model. This model represents a significant advancement in understanding and processing language due to its ability to analyze the context of comments from both preceding and following texts, thus capturing nuances that are often missed by more traditional methods.

Our methodology involves a detailed analysis of comment data extracted using the YouTube Data API, which is then preprocessed and fed into the BERT model for classification. This study not only aims to refine the moderation process on YouTube by automating the detection of toxic comments but also seeks to contribute to the broader field of digital communication by enhancing our understanding of interaction dynamics on social media platforms. By improving the online environment, we aim to foster a more inclusive and respectful digital discourse community, which is crucial in the age of ubiquitous digital media.

Motivation

Toxic comments can significantly degrade the quality of interactions on the platform, discouraging engagement and adversely affecting users' overall experience. By addressing and cleaning up the commenting environment, this project aims to transform YouTube into a more welcoming and enjoyable space for user interaction, thereby enhancing user retention and engagement. Furthermore, this initiative not only helps protect users' mental well-being by reducing their exposure to potentially distressing or abusive comments but also enhances user safety. Additionally, by filtering out harmful content, the project boosts the platform’s credibility and fosters trust among users and advertisers, underscoring a commitment to maintaining a positive and secure online environment.

The research questions are:

1) How relevant is a model trained with 5 years ago (2018’s) data on today’s data?
2) How has YouTube moderation changed? Improved or same or even worse?

3) How common are toxic comments on Youtubers?

4) What are the most prevalent types of toxic behavior exhibited in YouTube comments?

II. LITERATURE REVIEW

The provided literature reviews various methodologies for detecting and analyzing toxic comments on YouTube, each employing different machine learning techniques and data analysis methods tailored to specific facets of online discourse.

The 2022 study by Dekhoda et. al. focuses specifically on Swedish YouTube channels, utilizing a machine learning model known as "hatescan" to classify toxic comments. This more localized approach contrasts with general models aimed at English-language comments, highlighting the cultural and linguistic nuances in toxicity detection.

Shubhanshu Shekhar, Akanksha, and Aman Saini's 2021 research utilizes Latent Dirichlet Allocation (LDA) for topic modeling to pinpoint abusive comments, while incorporating sentiment analysis to gauge negativity. This method differs from the use of transformer-based models, which are typically more complex and robust, offering potentially deeper insights into the structure and nature of online interactions.

Finally, Obadimu et. Al., 2019 analyzes comments from a very niche segment of YouTube—specifically, channels that are pro- and anti-NATO. This targeted approach provides a detailed look at toxicity within a highly polarized context, unlike broader models that gather data across various genres and channels to ensure a more generalized analysis of YouTube toxicity.

Collectively, these studies underscore the diverse approaches in the ongoing effort to understand and mitigate online toxicity, each adapting different technologies and perspectives to better manage the digital communication landscape.

III. DATA

A. Data Collection

Training & Testing - For training and testing our model, we utilized the 'Toxic Comment Classification Challenge' dataset from Kaggle, which is specifically designed to provide a robust framework for training machine learning models to identify and classify types of toxicity in comments. The dataset comprises 159,571 records for training and an additional 63,978 records for testing, ensuring a comprehensive dataset size that allows for significant validation and robustness checks of the model's performance. Each record within the dataset consists of user comments annotated with one or more labels from six categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate. These categories enable the model to learn from a wide array of toxic behaviors and nuances in user comments, providing a thorough grounding in the types of negative interactions that can occur online, thus helping in effectively moderating content to improve user interaction and safety on platforms like YouTube.

B. Data EDA

The Exploratory Data Analysis (EDA) on detecting toxic comments on YouTube is important in shedding light on the initial analysis of the dataset used for training and testing our machine learning models. This phase was instrumental in dissecting the characteristics of the dataset, which comprises approximately 159,571 training records and 63,978 testing records, categorized under six labels: toxic, severe_toxic, obscene, threat, insult, and identity_hate. During the EDA, we employed a variety of statistical and visual techniques to analyze the distribution and frequency of these categories. This involved generating summary statistics like mean, median, and mode, and creating visual aids such as bar charts and histograms to illustrate the prevalence of each toxicity category. Such thorough preliminary analysis is crucial for preparing the dataset for deeper processing and ensuring the effectiveness of the models in real-world applications, aiming to enhance moderation of toxic comments on YouTube.

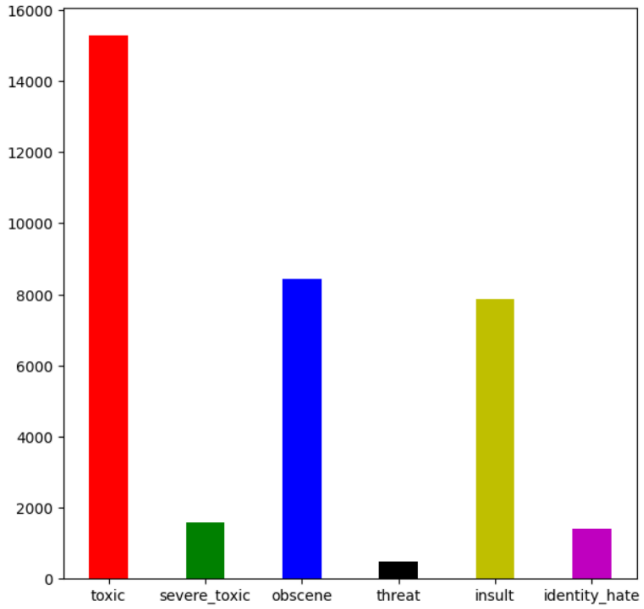

Fig. 1. Classification into different classes

C. Data Collection for Predictions - The Data Collection details the strategic approach taken to compile a comprehensive dataset for testing our machine learning model aimed at detecting toxic comments on YouTube. Utilizing the YouTube Data API v3, we meticulously gathered raw comment data from a deliberately selected variety of YouTube channels to capture the diverse nature of interactions and potential toxic behavior across different content genres. This included channels like Bloodywood, which merges Indian folk with heavy metal, offering insights into a unique cultural and musical interaction space; Jenny's Lectures, providing educational content in computer science; MKBHD, a top-tier technology review channel with a large, tech-savvy following; PewDiePie, a powerhouse in gaming and entertainment with a massive global audience that attracts a wide array of comments; and a mix of political news channels such as Lincoln Project, Fox News, and CNBC, each known for their politically

charged content that often spurs heated discussions. This selection ensures our dataset reflects the rich variety of user expressions and interactions on YouTube, enhancing the robustness and applicability of our predictive models in realistically identifying and classifying toxic comments across different contexts.

D. Data Pre-processing - The Data Pre-processing on detecting toxic comments on YouTube delineates the meticulous steps taken to refine and prepare the dataset for effective machine learning analysis. In this crucial phase, we undertook several key tasks to ensure the quality and consistency of the data. We began by removing all records containing null values to prevent any processing errors. Emojis, which could introduce ambiguity due to their varied meanings, were also stripped from the comments. Links and URLs were removed to focus the analysis strictly on textual content and avoid extraneous noise. Additionally, we eliminated non-ASCII characters to standardize text encoding across the dataset, and special characters such as @, !, #, $, etc., were also cleaned out to further reduce noise and normalize the data. These preprocessing steps were essential for creating a streamlined and uniform dataset, which is critical for training our models to accurately identify and classify toxic comments with high efficiency.

BERT Classification Model

BERT - Detecting toxic comments on YouTube includes a detailed section on BERT (Bidirectional Encoder Representations from Transformers), which is central to the project's approach for understanding and processing user comments. BERT is a state-of-the-art machine learning model developed by Google, known for its deep learning capabilities in natural language processing (NLP) tasks.

Here's an overview:

1. Model Architecture: BERT is built on the Transformer architecture, which uses attention mechanisms to understand the context of words in text. Unlike directional models, which read the text input sequentially (left-to-right or right-to-left), BERT reads the entire sequence of words at once. This allows the model to learn the context of a word based on all of its surroundings (both left and right of the word).

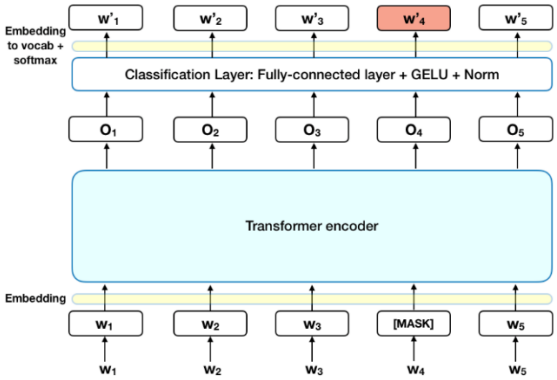

Fig. 2. Bert Architecture

2. Pre-training and Fine-Tuning: BERT is pre-trained on a large corpus of text in an unsupervised manner using two strategies: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM involves randomly masking words in a sentence and then predicting them based on the context, while NSP involves predicting whether a sentence logically follows another. After pre-training, BERT can be fine-tuned with additional output layers for a wide range of tasks, such as sentiment analysis, question answering, and, in the case of this project, toxic comment detection.

3. Application in the Project: In the project, BERT is used to enhance the detection of toxic comments by leveraging its ability to understand complex and nuanced language patterns. By fine-tuning BERT on the specific dataset of YouTube comments categorized under various forms of toxicity, the model can more accurately classify and predict toxic behavior in comments.

4. Benefits of Using BERT: Utilizing BERT allows for a more nuanced understanding of context within text, which is crucial for accurately identifying toxic comments that may not explicitly use negative words but are harmful due to the implied context. The bidirectional nature of the model ensures that the meaning of each word can be fully understood in relation to the words around it.

BERT's cutting-edge technology provides a robust framework for addressing the challenges of content moderation on social media platforms by enabling more accurate and context-aware analysis of user-generated text.

IV. METHOD

The methodology for detecting toxic comments on YouTube outlines a systematic approach utilizing the BERT model for content moderation. Initially, the method involved thorough data preparation, where we cleaned the dataset by removing non-ASCII characters, links, emojis, and special characters to ensure the model trained on high-quality, relevant textual data. We chose BERT due to its advanced capabilities in natural language processing, particularly its ability to contextually understand language from both left and right sides of a word. In specific, we employed BertForSequenceClassification since we have a multi-label classification problem. The training phase included fine-tuning the pre-trained BERT model on our specifically prepared dataset to identify various forms of toxicity in comments. This fine-tuning adjusts the model to specialize in recognizing toxic language patterns. Following training, the model underwent rigorous validation and testing to assess its accuracy and ability to generalize to new data.

In our approach, we adopted the AdamW optimizer that incorporates weight decay directly into the update step, enhancing training stability and generalization performance. The optimizer is set using with BCEWithLogitLoss loss function which we felt was more appropriate given the

binary nature of the class labels. Finally, the trained model was implemented in a real-time system to automatically moderate new YouTube comments, enhancing user interaction safety and experience by effectively filtering toxic content. This comprehensive methodological approach leverages advanced AI to address the challenges of online content moderation.

IV. RESULTS

We run our model on test data of the dataset and evaluate the model performance. The results on detecting toxic comments on different YouTube show how the model found various toxic comments across different channels.

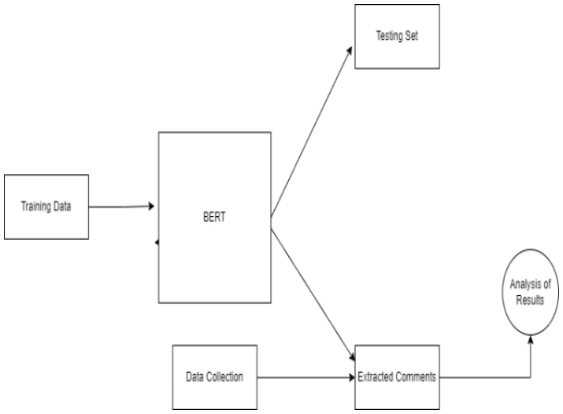

Fig. 3. Process Flow

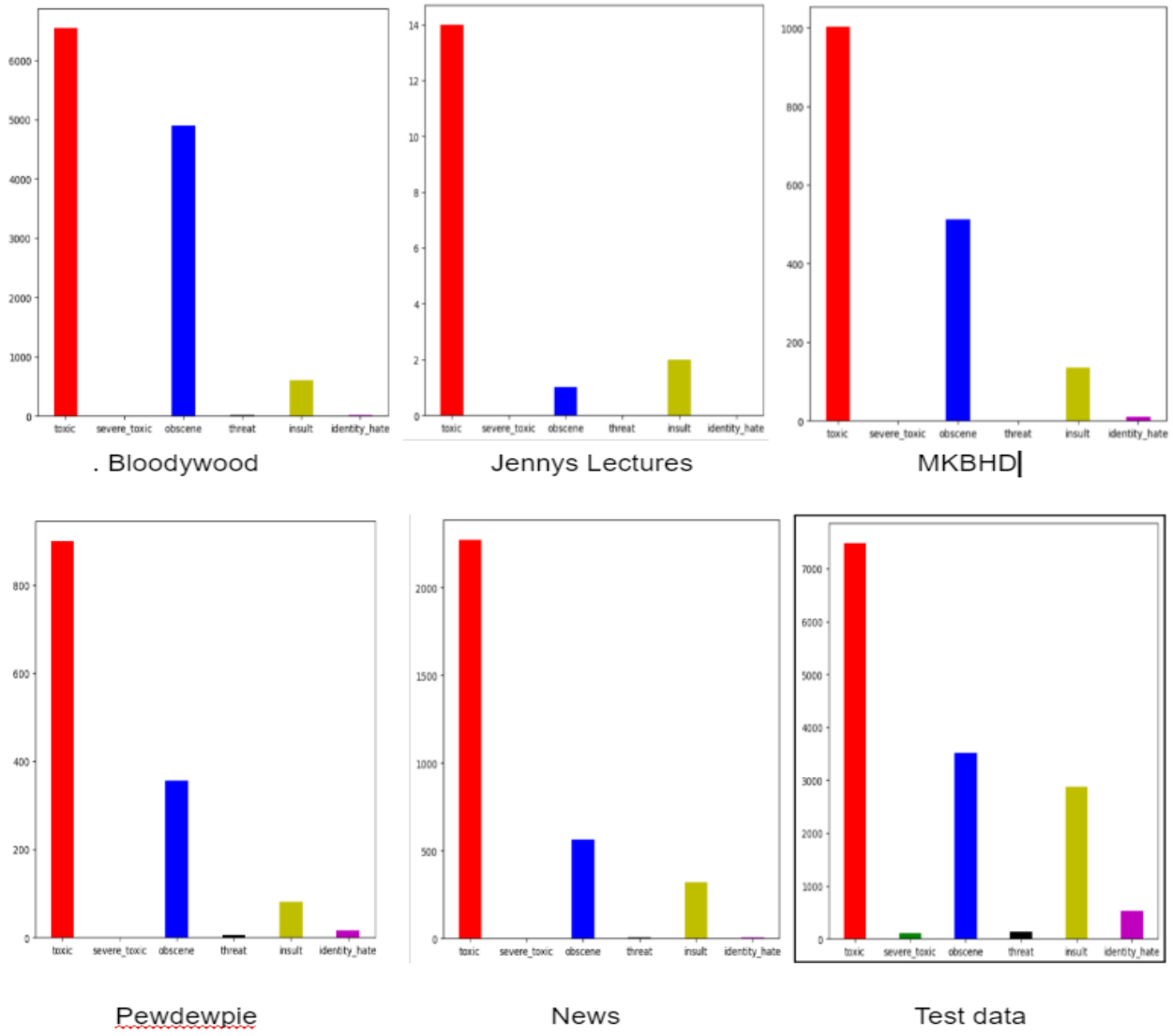

Fig 4 Results

Model Performance Metrics: Our model has a training loss of 0.0662164 and a testing loss of 0.023049, suggesting that the model fits the training data well and generalizes effectively to new, unseen data. The accuracy on the testing dataset is high at 97.695% but this can mislead model's performance. Instead, Exact Match Ratio (EMR) measures the proportion of instances in a dataset where the predicted output exactly matches the true output. The model achieved an EMR of 89.968%. These metrics indicate that the model is highly effective at correctly identifying both the presence and types of toxicity in comments.

We tested the model on comments from different channels.
The test data from Kaggle dataset used for training the model had 63,978 comments of which around 7,500 were classified as toxic, 3,500 comments as obscene, 3000 comments as insult, 100 comments as severe_toxic, 100 comments as threat and around 400 comments as identity_hate. Around 12% of the total comments were classified as one or more toxic labels.

From bloodywood music channel, we extracted around 61,000 comments of which around 6,500 were classified as toxic, 4,900 comments as obscene, 480 comments as insult, 0 comments as severe_toxic, 10 comments as threat and 11 comments as identity_hate. Around 12% of the total comments were classified as one or more toxic labels.

From Jenny lectures education channel, we extracted around 3,456 comments of which only 14 were classified as toxic, 1 comment as obscene, 2 comments as insult, 0 comments as severe_toxic/threat/identity_hate. Around only 1% of the total comments were classified as one or more toxic labels.

From MKBHD technology review channel, we extracted around 24,900 comments of which around 1,000 were classified as toxic, 492 comments as obscene, 160 comments as insult, 0 comments as severe_toxic/threat and 5 comments as identity_hate. Around 4% of the total comments were classified as one or more toxic labels.

From bloodywood music channel, we extracted around 61,000 comments, of which around 6,500 were classified as toxic, 4,900 comments as obscene, 480 comments as insult, 0 comments as severe_toxic, 10 comments as threat and 11 comments as identity_hate. Around 12% of the total comments were classified as one or more toxic labels.

From a famous content creator, Pewdewpie channel, we extracted around 13,500 comments of which around 900 were classified as toxic, 380 comments as obscene, 80 comments as insult, 0 comments as severe_toxic, 11 comments as threat and 5 comments as identity_hate. Around 4% of the total comments were classified as one or more toxic labels.

From diverse famous news channels like Fox news, CNBC, Lincoln project, we extracted 13,378 comments of which around 2,300 were classified as toxic, 600 comments as obscene, 340 comments as insult, 0 comments as severe_toxic, 4 comments as threat and 2 comments as identity_hate. Around 17% of the total comments were classified as one or more toxic labels.

Overall, the results demonstrate that the BERT-based model has achieved robust performance in detecting toxic comments, with high accuracy and low loss metrics, proving its potential as a powerful tool for enhancing content moderation on social platforms like YouTube.

V. DISCUSSION

Detecting toxic comments on YouTube provides a nuanced examination of the results yielded by the BERT model, emphasizing its success in achieving high accuracy and precision in identifying diverse types of toxicity. This segment addresses the research questions poised earlier. The model we've come with has been trained on a dataset collected in 2018 but the model is still relevant as evidenced by the model's ability to detect toxic comments from current data extracted from today's YouTube. We manually checked the predictions and almost most of them were rightly classified when cuss words were involved.

Coming to YouTube moderation, it is clearly evident from the graphs across fig 4 that YouTube platform has come to implement stricter policies to remove offensive and threatening comments from the comments section. We are pleasantly surprised to have not found a single severe_toxic comment in the entire records we've come to collect from today's YouTube. It is also evident from the graphs, that the platform took actions to limit threats and identity hate comments. These labels were extremely low especially when compared in comments from 5 years ago.

We've collected comments from diverse areas like music, education, tech reviews, news etc. and the toxic comments are comparatively low in content creators, tech reviews channels. The education channel had almost no toxic comments which is commendable but the comments from several videos of news channels constituted around 17% toxic comments which is relatively high. This reveals that barring select polarizing channels like news, rest of the youtube channels have low toxic comments.

Most of the toxic comments from collected data were of toxic, obscene, insult labels indicating that these are the more prevalent toxic behaviors in the comments sections.

Limitations on detecting toxic comments on YouTube critically assesses the challenges and constraints encountered with the BERT model used in the project. A primary issue identified is the model's occasional difficulty in accurately interpreting the context and nuances of language, such as slang or culturally specific idioms, which can lead to misclassification. For example, the model might erroneously flag positive expressions that contain traditionally negative words – consider the phrase - 'That's sick!' which the model considered to be toxic, but the same phrase can be used to express excitement. There are also limitations in handling comments in languages other than English or those that include specific cultural references.

The model's capability in detecting subtle forms of toxicity, like passive aggression, poses another significant challenge, as these require a nuanced understanding of context and intent that the current model may not fully capture. Moreover, scaling the solution to process comments in real-time on YouTube's extensive platform presents operational challenges. Lastly, the model's performance heavily relies on the quality and diversity of its training data; any inherent biases or gaps in this data can affect accuracy and perpetuate existing biases, highlighting the need for continuous refinement and testing of the model under diverse conditions. The approach we used could not work to take the emojis into account which is a limitation as emojis do have a lot of information.

REFERENCES

[1] Sasan Dehkhoda, Jasmyn Gunica on Analyzing Toxicity in YouTube Comments with the Help of Machine Learning

[2] Shubhanshu Shekhar; Akanksha; Aman Saini on Utilizing Topic Modelling to Identify Abusive Comments On YouTube Doi: 10.1109/CONIT51480.2021.9498368.

[3] Adewale Obadimu, Esther Ledelle Mead, Nitin Agarwal on Identifying Toxicity Within YouTube Video Comment Text Data DOI: 10.1007/978-3-030-21741-9_22.

[4] Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, Racial Bias in Hate Speech and Abusive Language Detection Datasets, 2019.

[5] J. Fox and W. Y. Tang, "Women's experiences with general and sexual harassment in online video games: Rumination organizational responsiveness withdrawal and coping strategies", New Media Society, vol. 19, no. 8, pp. 1290-1307, 2015, [online] Available: https://doi.org/10.1177/1461444816635778.

[6] Saeed Ibrahim Alqahtani, Wael M. S. Yafooz, Abdullah Alsaeedi, Liyakathunisa Syed, Reyadh Alluhaibi, "Children?s Safety on YouTube: A Systematic Review", Applied Sciences, vol.13, no.6, pp.4044, 2023.

[7] Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17. pp. 1217–1230. ACM Press, Portland, Oregon, USA (2017).

CONTRIBUTIONS

1. Chakradhar Reddy Nallu - Proposal, Data Preprocessing and Analysis, Model Validation, Report
2. Sri Harsha Mulluri - Data Collection, Preprocessing, Model Training, Report, Presentation
3. Harshitha Marasu - Proposal, Data Preprocessing and Analysis, Model Testing, Report
4. Subhasya Tippareddy - Model Training, Validation and Testing, Presentation
5. Ajay Babu Nannapaneni - None

Link to GitHub Repository –

https://github.com/SubhasyaTippareddy/YT_Comment_Classification