Online Bilingual Abuse Detection:

Using Natural Language Processing and Machine Learning

Santa Sian, ID 001108298

BSc (Hons) Computer Science

University of Greenwich

School of Computing and Mathematical Sciences

Contact Information:
University of Greenwich
30 Park Row, Greenwich, United Kingdom.
Email: ss8760t@gre.ac.uk

Abstract

Cyberbullying is on the rise with the expansion of social media, now affecting 45.5% of users, a significant increase from previous years. Current measures to combat it are limited in effectiveness. This project aims to enhance cyberbullying detection on social media platforms like Twitter by analysing messages in English and Hindi (Hinglish), focusing on real-time identification. By utilising advanced Deep Learning models such as BERT, and Natural Language Processing (NLP) techniques, the project develops a novel live chat system to reduce online abuse more effectively.

Introduction

In the contemporary landscape, 2024 is witnessing a proliferation of technological innovations from digital devices to artificial intelligence, marking a relentless advancement trajectory. Particularly noteworthy is the exponential growth observed within the domain of online social networks (OSNs). These platforms allow users to foster global interconnectedness, exemplified by the prevalence of networks such as WhatsApp, Instagram, Twitter, Facebook, Snapchat, and TikTok, among others. According to recent survey data, as of January 2024, approximately 62.3% of the global populace, equating to 5.04 billion individuals, actively engage with social media platforms. Forecasts indicate a trajectory of sustained expansion in the years ahead, driven by the increasingly widespread and accessible nature of technological infrastructure.

Unfortunately, while the prevailing narrative surrounding technological advancements in OSNs tends to emphasise their positive attributes, it is imperative to acknowledge the associated emergence of the negative aspects. One such detrimental trend is the escalation of online abuse, commonly known as cyberbullying. The intrinsic principle of freedom of speech inherent in these networks implies an unrestricted ability for individuals to express both positive and negative sentiments. While most OSNs have made strides in fortifying their security measures by implementing improved algorithms targeting fake accounts and empowering users to regulate comment permissions, deficiencies persist, particularly concerning the security of direct messaging functionalities. Despite the provision for users to report cyberbullying, the current infrastructure cannot pre-emptively detect and mitigate real-time abuse within private chats. Consequently, users remain susceptible to receiving hate-filled messages without recourse or protection. Recognising this gap, the primary solution scoped out for this project is directed towards developing an online abuse detection system in the form of a chat application, mirroring the structure of typical OSN chats. An additional secondary solution entails refining context detection mechanisms within conversations, enabling the anti-abuse system to discern nuances such as the usage of derogatory language in a non-abusive context.

Hindi as Second Language

Presently, India holds the distinction of being the most populous nation globally. Moreover, it ranks second in countries boasting the highest number of social media users, with a staggering figure of 755 million, projected to rise to 1.2 billion by 2027. This demographic reality has influenced the decision to adopt Roman Hindi as the secondary language, complementing English, reflecting the nation's linguistic diversity and the growing significance of digital communication platforms.

BERT

The novelty with using BERT for this project lies in its bidirectional nature, as it looks at both the left and right context of each word. This allows BERT to capture a fuller, more nuanced understanding of sentences. This feature is particularly vital in abuse detection, where the whole meaning of a word can be entirely altered by its context. This project utilises the ”bert-base-uncased” model and its accompanying tokenizer, leveraging its ability to capture complex linguistic patterns and contextual nuances.

Product Design

In order to create the product, the backend implementation must first be conducted, to develop the model which will be intergrated into the front-end and classify the text in the chat application as toxic or non-toxic. This process is done using Google Colab.

Back-End Implementation

The dataset is first loaded as a Pandas DataFrame. Then, data preprocessing is conducted, such as renaming columns and data labels, data analysis and visualisation: Understanding label distribution, null value, and duplicate checks. The text is then cleaned, set to lowercase, removing multiple spaces, tabs, newlines, links, special characters, numbers and stopwords. Using a process called Lemmatization, the words in text are reduced to their base form. Text and labels are tokenized using the BERT tokenizer, with padding and truncation for uniform tensor creation. The data is then split into training and validation sets for model training and evaluation. Subsequently, a custom torch Dataset class is created, which allows the model to easily use tokens with their labels. The DataLoaders, for both training and validation, sets handle batches during model training. The custom PyTorch dataset acts as a container for the preprocessed data. It ensures that the tokenized texts and their corresponding labels are stored in a structured format. Data imbalance is then tackled using class weights by applying them to a cross-entropy loss function which handles data imbalance effectively. The model is then trained using training arguments, which are obtained using random search, a method to find the best parameters across a given criteria. The best parameters chosen are as follows: Epoch number: 4, Batch size: 16, loss function: Cross-entropy loss

Front-End Chat Application

Based on research and analysis conducted, the chosen network topology for the front-end of the product is a client-server topology. The server is the central authority that manages data, client authentication and rules of communication, while the clients are the chat interface used to send and receive messages.

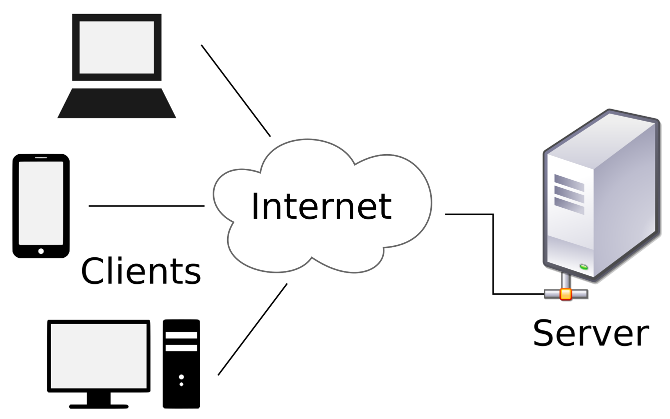


Figure 1: Client and Server Topology

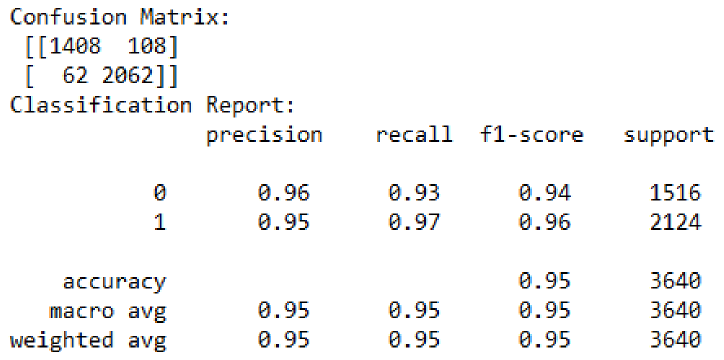
Dynamic Buffer

Each client has a rolling window (5 minutes) of messages stored in a double-ended queue. Whenever a new message is sent by a client, it is added to their individual deque, and the abuse detection is performed on a concatenated string of all their recent messages within that window. Messages older than 5 minutes are removed from the buffer, ensuring that the abuse detection is always based on recent communication. This acts as a short-term memory for the system.

Product Results and Evaluation

After training various models, and testing each one using the same type of conversation flow, a version without text preprocessing is chosen as the final model for the product. This is due to BERT's contextual undestanding, bidirectional nature and factor of being pretrained on raw text. Overcleaning the text may have removed ”noise” which is important for BERT models to learn from. Preprocessing steps may also remove subtle language cues that are important for understanding the tone or intent behind a text.

Figure 2 below shows the results obtained from training this model.



```
Confusion Matrix:
 [[1408  108]
 [  62 2062]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.93      0.94      1516
           1       0.95      0.97      0.96      2124

    accuracy                           0.95      3640
   macro avg       0.95      0.95      0.95      3640
weighted avg       0.95      0.95      0.95      3640
```

Figure 2: Confusion Matrix and Classification Report

The model performs as expected, classifying bad words as abusive, and giving a client two chances with warning messages, and upon the third abuse attempt, the server removes them completely. This creates a controlled environment for abuse. The message buffer was also tested, and performs well. It understands the words within context of conversation and does not misclassify them.

Evaluation

As shown in Figure 2, evaluation against metrics underline the models performance, having achieved a high 95% accuracy score. This shows that the trained model can generalise well to unseen data as it was trained using optimum parameters obtained through random search. The model was also compared against several works, in which it outperforms their methodologies used, however the main work used to evaluate the product against was [1], using the authors dataset and making necessary modifications to balance out the distribution of language content. [1] created their own chat application, however they used traditional machine learning methods, and although they were able to achieve similar accuracy scores, this project was able to enhance their work by adding on contextual analysis and improving on their dataset distribution.

Conclusion

To conclude, through extensive research conducted, the project was able to identify the gap currently being an inability for systems to correctly identify abusive content, especially in a bilingual setting, with English and Roman Hindi being the two languages used. The project went further by trying to detect context within conversations, with the aid of using a data structure such as a double-ended queue, the development of the dynamic message buffer within the server script was able to give a small contribution to how contextual analysis can be conducted from the front-end system rather than back-end.

Limitations and Future Work

Sophisticated model training needs considerable computational resources. The message buffer may also misclassify text due to message concatenation. The model's learning is limited to its English-centric dataset, which is lacking in an equal balance of Roman Hindi text. Future improvements can aim to partition datasets by language, use translated English datasets for better language support, and potentially include traditional Hindi script processing. Enhancements in context detection and abuse type identification, like racism or sexism, can be planned, although translating these in a multilingual context still remains challenging.

References

[1] Karan Shah, Chaitanya Phadtare, and Keval Rajpara. Cyber bullying detection for hindi-english language using machine learning. *SSRN Electronic Journal*, 2022.