Cluster Innovation Centre
University of Delhi

Detection of Offensive Content in Conversations

Group members Harshul Nanda, Keshav Sharma, and Yash Tomar

Supervised by Dr. Sachin Kumar and Dr. Nirmal Yadav

Text classification models research project submitted for the paper

SEMESTER LONG INNOVATION PROJECT

Abstract

In recent years, the analysis of conversational data has become increasingly important due to its potential role in identifying and preventing criminal activities. This poster presents a comprehensive study on the development and evaluation of various text classification models applied to a conversational dataset. Our goal is to accurately classify conversations as either safe and ethical or potentially leading to crime with the help of deep learning. We employ an array of machine learning and deep learning techniques, including Multinomial Naive Bayes, dense models, Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and state-of-the-art pre-trained models like BERT and RoBERTa. Through rigorous experimentation, we identify the best-performing model and fine-tune it for optimal performance on our dataset. The results demonstrate the effectiveness of our approach in accurately detecting conversations that could potentially lead to criminal activities, contributing to ongoing research in natural language processing and text classification for public safety and crime prevention.

Methodology

Our methodology involved the following steps:

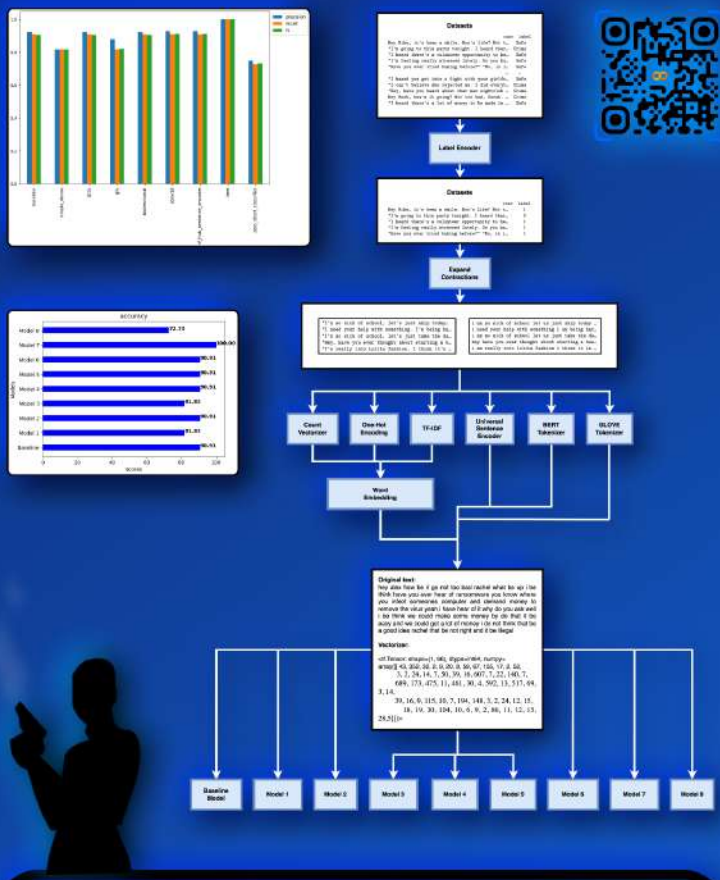
1. Data preprocessing, including expanding contractions, removing punctuations, stemming, and lemmatization.

2. Text vectorization (tokenization) to convert texts to numerical data.

3. Building, training, and evaluating various models, including dense models, LSTM, GRU, bidirectionalRNN, CNN, and pre-trained models like BERT and RoBERTa.

4. Fine-tuning the best-performing model for optimal results. Following are the models we have researched on:-

- Baseline Model - Multinomial Naive Bayes
- Model 1 - Simple Dense
- Model 2 - Long Short-Term Memory (LSTM)
- Model 3 - Gated Recurrent Unit (GRU)
- Model 4 - Bidirectional Recurrent Neural Network(Bi-RNN)
- Model 5 - Convolutional Neural Network (CNN)
- Model 6 - Pre-trained Embedding Layer
- Model 7 - BERT Modified
- Model 8 - Zero-Shot Classification with RoBERTaModel



Conclusion

Our project successfully explored and implemented multiple text classification models on a conversational dataset. The best-performing model demonstrated the ability to accurately identify conversations that could potentially lead to criminal activities is BERT Model with 100% accuracy on our testing dataset and 88% accuracy on our unknown to model dataset. This work contributes to the ongoing research in NLP and text classification, highlighting the importance of effectively analyzing conversational data for crime prevention and public safety.

References

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M.,Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., andStoyanov, V. (2019). RoBERTa: A Robustly Opti-mized BERT Pretraining Approach. arXiv preprintarXiv:1907.11692.
[2] Hochreiter, S., and Schmidhuber, J. (1997). LongShort-Term Memory. Neural Computation, 9(8),1735-1780.