Paper Title:
Toxicity Detection Using State of the Art Natural Language Methodologies

Paper Link:
https://ieeexplore.ieee.org/document/10155587

1 Summary

1.1 Motivation

The motivation of this paper is to explore the use of state-of-the-art natural language processing methodologies, such as sentence transformers, supervised machine learning algorithms, and BERT transformer architecture, to accurately detect and measure the toxicity level of text. The paper aims to develop a deep learning-based methodology that can effectively identify objectionable expressions, such as insults, violence, or obscene expressions, in order to provide control over the sharing of such content on online platforms.

1.2 Contribution

In this paper, the author conducted experiments and studies to detect objectionable expressions in text, with a specific focus on toxicity detection. They employed cutting-edge natural language processing and machine learning methodologies, including sentence transformers, supervised machine learning algorithms, and the BERT transformer architecture. The paper explained the workings of transformer architectures, particularly BERT, which assess the interplay between words and their contextual meaning. The use of supervised machine learning algorithms like logistic regression, random forest, XGBoost, and KNN for classification tasks was also discussed. The results of the experiments, along with data preprocessing steps and performance metrics for evaluation, were presented. Ultimately, the author concluded that deep learning-based methodologies, particularly those utilizing BERT, are effective in measuring text toxicity and enabling control over objectionable content on online platforms.

1.3 Methodology

The methodology section of the study provides an overview of the transformer architectures, with a focus on BERT, and introduces the supervised machine learning algorithms used, including Logistic Regression, Random Forest, Xgboost Classifier, and K-Nearest Neighbors (KNN). It explains how BERT is fine-tuned using transfer learning and mentions Huggingface as the implementation source for these transformer models.

1.4 Conclusion

The conclusion of the document is that deep learning-based methodologies, specifically the use of state-of-the-art natural language processing techniques, can effectively detect the toxicity level of text. By training models using labeled text data and methodologies such as sentence transformers, supervised machine learning algorithms, and BERT transformer architecture, objectionable expressions in text, such as insults, violence, or obscene expressions, can be detected. The

document highlights the importance of detecting toxicity in text to improve the quality of websites and enhance the reader experience. The experiments conducted in the document show promising results, with performance metrics such as accuracy, precision, recall, and F-score indicating the effectiveness of the methodologies used. Overall, the document concludes that using deep learning-based methodologies can provide control over objectionable content shared on online platforms.

2 Limitations

2.1 First Limitation

The paper mentions the use of datasets obtained from the Kaggle platform, but it does not provide detailed information about the size, diversity, or representativeness of these datasets .The use of a limited dataset can restrict the generalizability and reliability of the findings.

2.2 Second Limitation

The paper primarily focuses on the application of state-of-the-art natural language methodologies for toxicity detection without comparing the performance of these methodologies with alternative approaches or existing models. A comparison with other methods could provide a more comprehensive understanding of the strengths and weaknesses of the proposed methodologies.

3 Synthesis

The paper explores using advanced NLP methods to detect and control toxic text on online platforms. It emphasizes the use of transformer architectures like BERT for effective toxic content detection. These methods can automatically measure toxicity in user-generated text, improving content quality and user experience. Websites can assign toxicity scores to texts and use them to monitor and regulate content sharing. The same methodologies can be applied to advertisement filtering, ensuring ad appropriateness on online platforms.