IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 6, Ser. 2 (Nov. – Dec. 2023), PP 54-57
www.iosrjournals.org

System For Detecting Toxic Speech

Indranil Banerjee[1], Soyeta Dey[2]
[1]*Assistant Professor Cse, Nshm Knowledge Campus, Durgapur, West Bengal, India*
[2]*B.A. English Honours , Durgapur Government College, Durgapur, West Bengal, India*

Abstract:
Detecting toxic speech is a valuable tool for addressing the proliferation of harmful content, particularly on social media platforms. Many methods, including recent deep learning-based approaches, have been developed for this task. Various datasets have also been created to illustrate different aspects of the toxic speech detection problem. In this study, we conduct a comprehensive empirical comparison of deep and shallow toxic speech detection methods using three commonly utilized datasets. Our primary aim is to offer insights into the field's progress and highlight both strengths and weaknesses of current state-of-the-art techniques. Our analysis focuses on practical performance metrics such as detection accuracy, computational efficiency, the capability to utilize pre-trained models, and domain adaptability. Through this research, we aim to provide practical guidance for toxic speech detection applications, assess the current state-of-the-art, and pinpoint potential avenues for future research.

Key Word*: Toxic, deep-learning, detection accuracy, computational efficiency.*

Date of Submission: 13-12-2023 Date of Acceptance: 23-12-2023

I. Introduction

In today's contemporary society, individuals enjoy the freedom to express their thoughts through social media platforms, where they can share their creations, images, and engage in discussions by commenting on others' content. However, in the present day, there is an abundance of comments across various social networks, news websites, and forums, and a significant portion of these comments contain toxic or abusive content. Given the sheer volume of comments, manual moderation is impractical, leading to the widespread use of automated methods that employ machine learning models for the detection of toxic content. In this research, we will conduct a systematic review of the current state-of-the-art in the field of toxic speech detection using machine learning techniques.

II. What is toxic speech

Toxic speech, or hate speech, can be characterized as any form of expression that singles out individuals or groups based on attributes such as their race, religion, ethnicity, national origin, sexual orientation, or gender identity. Hate speech is frequently employed to promote animosity and prejudice, and it can serve as a means of intimidating and menacing individuals. This kind of speech has the potential to induce feelings of isolation, anxiety, and fear, and it can even contribute to the occurrence of hate crimes. Furthermore, hate speech can strain relationships between various groups of people. The identification and detection of hate speech are of utmost significance because they can help mitigate these harmful consequences.

III. Feasibility Study

While social media offers numerous advantages, it is not without its downsides. Within this digital environment, malicious individuals engage in unethical and deceptive activities aimed at causing emotional harm and tarnishing the reputations of others. In recent times, cyberbullying has emerged as a significant problem in the realm of social media. Cyberbullying, or cyber-harassment, involves the use of electronic means to engage in bullying or harassment, and it is commonly referred to as online bullying. With the expansion of the digital landscape and advancements in technology, cyberbullying has become increasingly prevalent, especially among adolescents. Around 50% of teenagers worldwide encounter cyberbullying, and this form of bullying exerts both physical and psychological effects on its victims. The trauma inflicted by cyberbullying can drive victims to engage in self-destructive behaviors, including contemplating suicide, as it can be exceptionally distressing to endure. Consequently, recognizing and preventing cyberbullying is vital to safeguard the well-being of teenagers.

IV. Literature Review

In 2019, d'Sa, A. G., Illina, I., & Fohr, D. presented a paper [1] where they focused on designing binary classification and regression-based methods to predict the toxicity of comments. The study involved comparing various unsupervised word representations and DNN-based classifiers. Additionally, they investigated the robustness of their approaches against adversarial attacks, specifically by introducing one additional word (either healthy or toxic). The evaluation of their methodology was conducted on the English Wikipedia Detox corpus. The experiments demonstrated that fine-tuning BERT surpassed feature-based BERT, Mikolov's, and fastText representations when combined with different DNN classifiers.

In 2019, Abderrouaf, C., & Oussalah, M. introduced a paper [2] aiming to make advancements in the field through the application of innovative techniques in natural language processing, machine learning, and feature engineering. The proposed approach advocates for a classification-like technique that incorporates a unique data design procedure. This procedure ensures a balanced training scheme by exploring the negativity within the original dataset, leading to the creation of new transfer learning paradigms. The study contrasts two classification schemes utilizing convolutional neural network and LSTN architecture, both using FastText embeddings as input features, with baseline models comprising Logistic regression and Naive Bayes classifiers. The validity and utility of the proposal are tested using the Wikipedia Comment dataset, which includes data on Personal Attack, Aggression, and Toxicity.

In 2019, Arango, A., Pérez, J., & Poblete, B. conducted a study [3] in which they delved into the apparent contradiction between existing literature and real-world applications. Their analysis focused on scrutinizing the experimental methodology employed in previous research and its applicability to different datasets. The results revealed methodological issues and a notable bias in the dataset, leading to a significant overestimation of the performance claims made by the current state-of-the-art methods. The identified problems were primarily associated with data overfitting and sampling issues. In response, they discussed the implications for ongoing research and conducted new experiments to present a more accurate assessment of the current state-of-the-art methods.

In 2021, Nguyen, L. T., Van Nguyen, K., & Nguyen, N. L. T. introduced a paper [4] wherein they established a dataset named UIT-ViCTSD (Vietnamese Constructive and Toxic Speech Detection dataset) comprising 10,000 human-annotated comments for constructive and toxic speech detection. They proposed a system for these tasks, leveraging the state-of-the-art transfer learning model in Vietnamese NLP, known as PhoBERT. The system achieved F1-scores of 78.59% and 59.40% for the classification of constructive and toxic comments, respectively. Additionally, they implemented various baseline models, including traditional Machine Learning and Deep Neural Network-Based models, to assess the dataset. The outcomes enabled them to address several challenges related to online discussions and develop a framework for automatically identifying the constructiveness and toxicity of Vietnamese social media comments.

In 2021, Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. presented a paper [5] proposing an automated classification approach for tweets into three categories: Hate, offensive, and Neither. Initially, they conducted experiments using a public tweet dataset to construct BI-LSTM models with both empty embedding and pre-trained Glove embedding. Subsequently, they explored a transfer learning method for detecting hate speech, employing pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), DistilBert (Distilled version of BERT), and GPT-2 (Generative Pre-Training). The authors conducted an analysis of hyperparameters for their best model (BI-LSTM), considering various neural network architectures, learning rates, and normalization methods. After fine-tuning the model with the optimal parameter combination, they achieved a test data accuracy of over 92 percent. Additionally, they developed a class module encompassing key functionalities like text classification, sentiment checking, and text data augmentation, positioning this model as an intermediary module between users and Twitter.

In 2022, Velankar, A., Patil, H., & Joshi, R. extensively delved into the complexities of automatic hate speech detection, presenting a hierarchical framework to organize these challenges. Their emphasis was on the obstacles encountered by machine learning or deep learning approaches in identifying hate speech. The top-level categorization included challenges at the data level, model level, and human level. They proceeded to conduct a thorough analysis of each level within the hierarchy, supplemented by illustrative examples. This survey served as a valuable resource for researchers, aiding in the more efficient design of solutions within the realm of hate speech detection.

In 2023, Mazari, A. C., Boudoukhani, N., & Djeffal, A. conducted a study [7] focusing on multi-aspect hate speech detection, which involved classifying text into multiple labels such as 'identity hate,' 'threat,' 'insult,' 'obscene,' 'toxic,' and 'severe toxic.' The proposed approach centered around leveraging the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, combined with Deep Learning (DL) models, to construct various ensemble learning architectures. The DL models utilized in this study were constructed by stacking Bidirectional Long-Short Term Memory (Bi-LSTM) and/or Bidirectional Gated Recurrent Unit (Bi-GRU) on GloVe and FastText word embeddings. These models, along with BERT, were individually trained on a multi-label hateful dataset and then combined for hate speech detection tasks on social media. The results demonstrated that encoding texts using recent word embedding techniques such as FastText and GloVe, in conjunction with Bi-LSTM and Bi-GRU, could create models that, when combined with BERT, significantly improved the ROC-AUC score to 98.63%.

V. How can machine learning be used to detect toxic speech

Machine learning, a subset of artificial intelligence, is a valuable tool for deriving insights from data, including the identification of patterns. It is also capable of detecting toxic speech by employing algorithms that analyze text and ascertain its tone. This functionality is particularly important for uncovering disguised hate speech or toxic content presented as humor or sarcasm, which is prevalent on social media platforms.

Automated toxic speech detection is a crucial instrument in the battle against the proliferation of hate speech, especially in the context of social media. Techniques for identifying toxic speech through machine learning encompass traditional classifiers, deep learning, transfer learning-based classifiers, or combinations of these methods. Deep learning, a machine learning approach, is used to extract patterns from data, while transfer learning allows the utilization of previously acquired knowledge from other machine learning algorithms. These techniques have significantly advanced various domains, including hate speech detection, within machine learning and natural language processing (NLP). Consequently, they play a pivotal role in the identification of toxic speech.

VI. Methodology

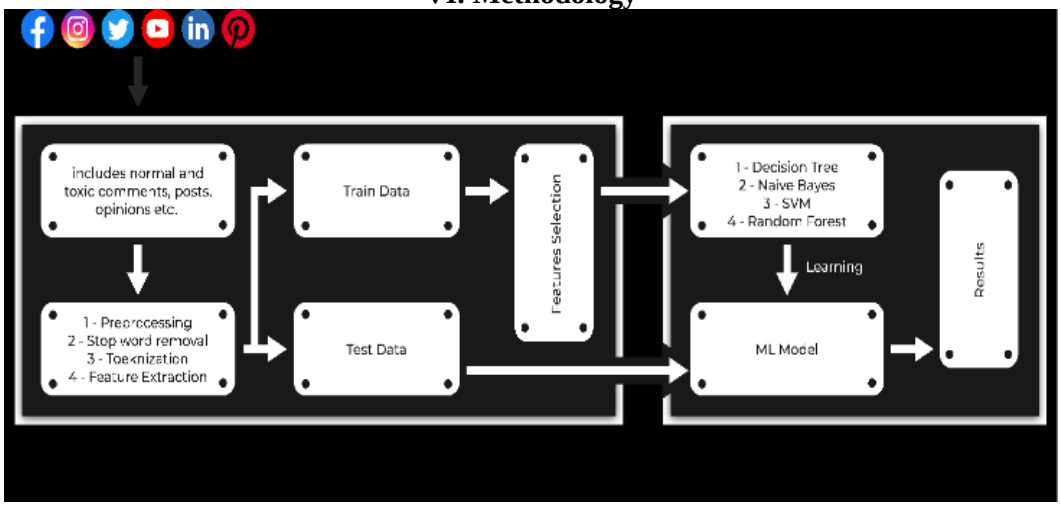

Figure 1: Flow chart of Toxic Speech Detection System

Phase Overview

Raw Data:

Raw data constitutes a collection of unprocessed information received from a specific data source, and it has not undergone any manipulation by either machines or humans. This data is typically sourced from online platforms to gain in-depth insights into users' online behaviors. Leveraging this information, marketers can effortlessly craft personalized online campaigns and effectively target users with precise messages at the opportune moment. It is important to acknowledge that raw data, in its unprocessed state, lacks utility.

Typically, it appears as a string of code, such as a user cookie, which on its own does not offer substantial information. However, when this data is integrated with relevant user profiles, it becomes a valuable resource for marketers and business analysts.

Data Labelling:

In the realm of machine learning, data labelling involves the task of taking raw data, which can be in the form of images, text files, videos, and more, and annotating it with one or more meaningful labels that offer context. These labels serve the purpose of enabling a machine learning model to gain knowledge from the data. For instance, labels may specify whether a photo features a bird or a car, transcribe the spoken words in an audio recording, or indicate the presence of a tumor in an x-ray. Data labelling is an essential process across various applications, including computer vision, natural language processing, and speech recognition.

Tokenization:

Tokenization represents a fundamental element within the broader framework of natural language processing (NLP). This straightforward procedure takes unprocessed data and transforms it into a structured data string. While tokenization is commonly recognized for its application in fields like cybersecurity and the development of NFTs, it also plays a crucial role in the NLP workflow. In the realm of natural language processing, tokenization is employed to segment paragraphs and sentences into smaller units, facilitating the assignment of meaning to these units. The figure below illustrates several examples of tokenizers.

VII. Conclusion

The persistent issue of toxic speech underscores the growing demand for automated toxic speech detection systems. We have discussed existing approaches for addressing this problem and introduced a novel system that demonstrates a commendable level of accuracy. Furthermore, we have put forward an innovative approach that not only excels in comparison to current systems but also offers enhanced interpretability. While many challenges persist, it is clear that further research is warranted on this matter, encompassing both technical and practical aspects.

References

[1]. d'Sa, A. G., Illina, I., & Fohr, D. (2019). Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN. Arxiv Preprint Arxiv:1911.08395.
[2]. Abderrouaf, C., & Oussalah, M. (2019, December). On Online Hate Speech Detection. Effects Of Negated Data Construction. In 2019 IEEE International Conference On Big Data (Big Data) (Pp. 5595-5602). IEEE.
[3]. Arango, A., Pérez, J., & Poblete, B. (2019, July). Hate Speech Detection Is Not As Easy As You May Think: A Closer Look At Model Validation. In Proceedings Of The 42nd International Acm Sigir Conference On Research And Development In Information Retrieval (Pp. 45-54).
[4]. Nguyen, L. T., Van Nguyen, K., & Nguyen, N. L. T. (2021). Constructive And Toxic Speech Detection For Open-Domain Social Media Comments In Vietnamese. In Advances And Trends In Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference On Industrial, Engineering And Other Applications Of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34 (Pp. 572-583). Springer International Publishing.
[5]. Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. (2021). Offensive Language And Hate Speech Detection With Deep Learning And Transfer Learning. Arxiv Preprint Arxiv:2108.03305.
[6]. Velankar, A., Patil, H., & Joshi, R. (2022). A Review Of Challenges In Machine Learning Based Automated Hate Speech Detection. Arxiv Preprint Arxiv:2209.05294.
[7]. Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2023). BERT-Based Ensemble Learning For Multi-Aspect Hate Speech Detection. Cluster Computing, 1-15.