Toxic Speech Detection System

Arnab Chakraborty[1], Soumalya Das[1], Rohit Mondal[1], Biswajit Gope [1]

[1]NSHM Knowledge Campus, Durgapur

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 utilised 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 utilise 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.

Keywords: Toxic, deep-learning, detection accuracy, computational efficiency.

1. 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.

Toxic speech, or hate speech, can be characterised 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.

2. 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. 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 analyse text and ascertain its tone. This functionality is particularly important for uncovering disguised hate speech or toxic content presented as humour 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 utilisation 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.

3. Methodology

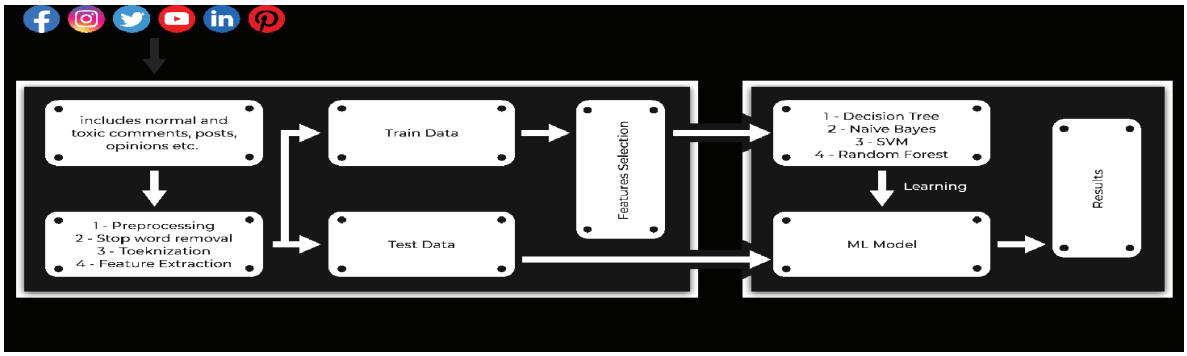

Figure 1: Flow chart of Toxic Speech Detection System

3.1 Phase Overview

3.1.1 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 behaviours. Leveraging this information, marketers can effortlessly craft personalised 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.

3.1.2 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 tumour in an x-ray. Data labelling is an essential process across various applications, including computer vision, natural language processing, and speech recognition.

3.1.3 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.

4. 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.

Reference

1. https://www.resurgentindia.com/paper-feasibility-study-and-its-importance-in-paper-management#
2. Battery Power Saving Profile with Learning Engine in Android Phones - Rimpy Bala & Anu Garg, International Journal of Computer Applications (0975 – 8887) Volume 69– No.13, May 2013
3. Web Performance with Android's Battery-Saver Mode - Utkarsh Goel, Stephen Ludin & Moritz Steiner
4. https://www.researchgate.net/publication/361470812_Battery_Saver_in_Android_Studio Battery Saver - Full Android Source Code by LubuTeam | Codester