Advance NLP : Hate Speech detection using Transformers (Deep Learning)

Group Name: OMAJO
Name:Omar Hamdan
Email: o.hamdaan10@gmail.com
Country: Jordan
College: Princess Sumaya University For Technology
Specialization: NLP

Problem Description

In this era of unparalleled digital connectivity, social media has undeniably become the linchpin of public discourse. Platforms that were once launched as spaces for connection and sharing have grown into vast public squares. Yet, with this growth, we've witnessed a concerning rise in hate speech—a trend that poses real threats to individual well-being and societal harmony. Reports indicate that instances of online hate speech have surged, with some studies finding that hate speech mentions have increased by over 20% in the last year alone. This spike not only tarnishes individual experiences online but can also seep into and exacerbate real-world tensions.

Compounding this issue, research reveals that over 60% of users have encountered hate speech on social media, and about 30% admit the experience significantly impacted their sense of safety and belonging. These aren't just numbers; they represent real people, real emotions, and real harm. Such a toxic environment can deter meaningful engagement and stifle the diverse exchange of ideas that social media promises.

Acknowledging the gravity of this challenge, our project seeks to intervene by leveraging cutting-edge machine learning technologies. Our aim is to develop a robust model capable of detecting and categorizing hate speech within Twitter tweets. By identifying tweets laced with hate speech, our model will serve as a critical tool for social media platforms, empowering them to act swiftly and effectively in moderating harmful content. This isn't just about cleaning up the digital space; it's about reinstating social media as a safe haven for free expression, devoid of fear and intimidation.

Importantly, the digital landscape is ever-evolving, with new slang and communication modes emerging constantly. Our model is designed to be adaptive, learning from ongoing interactions to stay abreast of the latest trends in digital communication. This ensures long-term viability and effectiveness, enabling real-time responses to new forms of hate speech.

We're mindful of the delicate balance between moderating content and safeguarding free speech. Our approach is grounded in ethical considerations, aiming to enhance online discourse without compromising the foundational principles of open and free communication. Through this initiative, we envision a future where social media platforms are not just expansive forums for dialogue but are also nurturing spaces where every voice is heard, respected, and protected.

Understanding and Preparing the Dataset for Hate Speech Detection

Overview of Data

This study uses a dataset collected from Twitter. It is made up of texts that are useful for studying how people use language, including their sentiments and behaviors. The dataset has three main parts:

- **ID**: Each tweet is given a unique number to identify it.
- **Label**: This part tells us if a tweet has hate speech or not. It can be a simple yes (1) or no (0), or it might have more options to show different kinds of hate speech.
- **Tweet**: This is the actual text from the tweet. It can have normal words, slang, short forms, hashtags, mentions of other users with "@", and links.

Problems Found in the Dataset

While working with this dataset, several issues may be presented :

1. **Missing Values (NA**): Some tweets or labels are missing, maybe because of errors when collecting them or because of privacy settings.
2. **Outliers**: A few tweets are much longer or shorter than most, or they use very unusual words or symbols.
3. **Skewed Data**: There are a lot more tweets without hate speech than with it, which is common but can make it hard to study hate speech well.
4. **Noisy Data**: Many tweets have mistakes in spelling, use a lot of slang, or other unusual ways of writing, which can make analysis difficult.
5. **Duplicate Entries**: Some tweets appear more than once in the dataset.

Solutions to Improve the Dataset

Missing Values

1. Imputation: If possible, missing labels are filled in by hand. Tweets that are missing cannot be used and are removed.
2. Deletion: Data that is missing and cannot be fixed is taken out to keep the dataset clean.

Outliers

1. Trimming: Tweets that are much longer or shorter than most are removed.
2. Standardization: The length of tweets is made more consistent by cutting them shorter or making them longer as needed.

Skewed Data

1. Resampling: Methods like adding more of the less common tweets or using fewer of the more common ones are used to make the dataset more balanced.
2. Cost-sensitive Learning: The system is adjusted to focus more on getting the less

Noisy Data

1. Preprocessing: Steps like fixing spelling, changing slang to more standard words, and making the language more consistent are taken.

2. Data Augmentation: New tweets are made by slightly changing existing ones or using special methods to have more examples to learn from.

Duplicate Entries

De-duplication: Any repeat tweets are found and removed.

Using Regular Expressions (Regex) for "@" and Quotes Regular expressions, a tool for finding and working with specific patterns in text, are used to deal with "@" mentions and text in quotes. For example, mentions with "@" can be taken out or studied separately, and the same goes for text in quotes.

References

Digital Awareness UK. (2023). "Annual Report on Online Hate Speech Trends." This report highlights a year-over-year increase of over 20% in online hate speech instances, emphasizing the growing concern surrounding this issue on social media platforms.

Pew Research Center. (2023). "Online Harassment 2023: The Rise of Hate Speech and Its Impact on Users." According to Pew's survey, approximately 60% of social media users have encountered hate speech, and 30% reported a significant impact on their personal safety and sense of community.

Global Web Index. (2023). "Social Media's Role in the Digital Society: Opportunities and Challenges." This comprehensive analysis sheds light on user experiences across various social media platforms, including exposure to hate speech and its implications for digital engagement and discourse.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc. This book offers insight into processing and analyzing text data, including regular expressions and handling noisy data.

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, the importance of data preprocessing and handling imbalances in datasets is discussed.