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.

Business Understanding

Objective
The primary objective of our project is to assist social media platforms in automatically identifying and moderating hate speech, thereby enhancing the quality of online discourse. By accurately detecting hate speech, our solution aims to support these platforms in maintaining community standards and complying with regulatory requirements.

Impact
The successful implementation of this project will have a significant positive impact on social media ecosystems by:

- Reducing the prevalence of hate speech.
- Protecting users from harmful content.
- Assisting in the enforcement of platform policies against discrimination and harassment.

Project Lifecycle and Deadline

Phases

1. Project Initiation: Formulate project scope and objectives.
2. Data Collection: Gather and prepare Twitter datasets for analysis.
3. Data Cleaning and Preprocessing: Address data quality issues.
4. Model Development: Design and train the machine learning model.
5. Evaluation: Assess model performance and refine as necessary.
6. Deployment: Implement the model in a real-world scenario.
7. Monitoring and Maintenance: Continuously improve the model based on new data and feedback.

Deadline

- Project Initiation: March 19, 2024 - Finalize and agree upon project scope, objectives, and initial resources.
- Data Collection: March 22, 2024 - Complete gathering of necessary datasets from Twitter, ensuring data is ready for cleaning and preprocessing.
- Data Cleaning and Preprocessing: March 27, 2024 - Finish addressing data quality issues, preparing the dataset for model development.
- Model Development: April 6, 2024 - Conclude the development and initial training of the machine learning model, ensuring it's ready for preliminary testing.
- Evaluation: April 15, 2024 - Complete the evaluation of the model, incorporating feedback for refinement.
- Deployment: April 20, 2024 - Deploy the model for real-world testing, beginning its integration into the desired application or service.
- Monitoring and Maintenance: April 30, 2024 - Start the initial phase of monitoring and maintenance, with plans for ongoing adjustments based on performance and feedback.

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.