MLOffense: Multilingual Offensive Language Detection and Target Identification on Social Media Using Graph Attention Transformer Model

Wang, Grant (School: Williamsville East High School)

With the increasing use of social media in our daily lives, it is crucial to maintain safe and inclusive platforms for users of diverse backgrounds. Offensive content can inflict emotional distress, perpetuate discrimination towards targeted individuals and groups, and foster a toxic online environment. While AI-based natural language processing (NLP) has been employed for automatic offensive language detection, most studies focus on English only, leaving languages other than English understudied due to limited training data. This project fills this gap by developing a novel multilingual model for offensive language detection in 100 languages, leveraging existing English resources. The model employs graph attention mechanisms in transformers, improving its capacity to extend from English to other languages. Moreover, this work breaks new ground as the first study ever to identify the specific individuals or groups targeted by offensive posts. Statistical analysis using F1 scores shows high accuracy in offensive language classification and target recognition across multiple languages. The trained model is deployed on an application programming interface (API), enabling users to explore the model's capabilities of multilingual offensive language detection and prevention in social media settings. Furthermore, a pilot study is conducted to showcase the model’s potential for providing valuable data to inform behavioral and social science research. This innovative model represents a significant step forward in the field of offensive language detection, paving the way for a safer and more inclusive social media experience for users worldwide.

Awards Won:

Third Award of $1,000

Central Intelligence Agency: First Award: $1000 award

National Security Agency Research Directorate : Third Place Award “Cybersecurity”