Proposal: Detecting Offensive Language in Tweets with Attention-fused Text and User Embeddings

Viha Gupta,[1] **Casey Primel**[1]

[1] Dept. of Computer Science
vg2237@nyu.edu, ctp219@nyu.edu

Content moderation is a pressing concern for online communities. For over a decade, automated systems have been employed to identify abusive behaviors and offensive language in user-generated content. Early systems relied on hand-crafted rules to extract linguistic features from textual data. Such rule-based systems have largely been replaced by neural network-based systems that can identify more subtle patterns in users' linguistic behaviors. However, research has shown that textual features are not always sufficient for identifying the nature of user-generated content. The rise of social media and the corollary ability to query the social graphs formed by users' activity has opened up a possible data source for constructing the requisite context. The remaining question is how to most effectively provide such context from the available data. Our project focuses on one solution in the context of Twitter: the fusion of text and user embeddings via attention mechanisms.

Recent work exploring the application of deep learning to offensive language detection on Twitter can be divided into two different approaches. The first approach focuses solely on text data (e.g., tweets) and the use of pre-trained language models fine-tuned for this domain-specific task (Liu, Li, and Zou 2019). The second approach combines text data with information about users and community structure. Recent exemplars of this approach have used graph neural networks to learn the features of the social graph: Mishra, Yannakoudakis, and Shutova (2019) constructs a heterogeneous graph of users and tweets and applies a graph convolutional network to learn its features. Miao et al. (2022) begins by independently extracting text and user embeddings and then fusing them via attention mechanisms. The latter serves as the starting point for our project.

Datasets and models

For this project, we will be using the English-language dataset collected, labeleled and made publicly available on GitHub by Miao et al. (2022).[1] The data was obtained via the Twitter API[2] and contains data from 1,260 users and 12,780 of their tweets posted between Janauary 2018 and March 2021. The dataset also captures the social community in terms of the follower-network between these users with 8,877 such relationships.

We will be following Miao et al. (2022) in using a graph attention network for learning user embeddings and attention mechanisms for fusing user features with tweet features; however, rather than using text embeddings from BERT, we will use the RoBERTa-based model pre-trained on tweets and fine-tuned for offensive language detection (Barbieri et al. 2020).[3]

Deliverables

By the end of this project, we expect to produce a model which performs on par or exceeds the performance of Miao et al. (2022) (F1 score: 89.94) on standard benchmarks for offensive language identification. In addition to the aforementioned switch between BERT- and RoBERTa-based text embeddings, we will also explore different strategies for augmenting user embeddings. We will collect our methodologies and all results in a detailed report. All code and example notebooks will be published in a GitHub repository.

References

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.

Liu, P.; Li, W.; and Zou, L. 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 87–91. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Miao, Z.; Chen, X.; Wang, H.; Tang, R.; Yang, Z.; and Tang, W. 2022. Detecting Offensive Language on Social Networks: An End-to-end Detection Method based on Graph Attention Networks.

Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2019. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods.

[1] https://github.com/mzx4936/GF-OLD-Dataset

[2] https://developer.twitter.com/en/docs/twitter-api

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive