Project ID: 026

Project Title: Multi-Class Hate Speech and Offensive Language Detection in Social Media Area of

Research: Natural Language Processing (NLP) Problem Statement: Hate speech and offensive language on social media platforms can lead to onlineharassment,discrimination, and a toxic online environment. Automatically identifying and categorizing such content iscrucial foreffective content moderation. However, distinguishing between hate speech, offensive language, and normal speechcan be challenging, as the boundaries between these categories are often blurred. The goal of this project istodevelop a deep learning-based system that can accurately classify social media posts into three categories: hatespeech, offensive language, and normal speech. This will enable social media platforms to better understandthenature of potentially problematic content and take appropriate actions. Dataset: Figure. Examples of annotated text data for sentiment analysis from social media. (Each entry includes a post ID, the text of the post, annotators' IDs, targets, labels (normal, offensive, hatespeech),rationales, and post tokens).

Dataset URL: https://github.com/hate-alert/HateXplain/tree/master/Data Task: Develop a deep learning model for multi-class classification of social media posts into three categories: hatespeech, offensive language, and normal speech. Relevant

Papers

[1] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: Abenchmarkdataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17),14867-14875. <https://doi.org/10.1609/aaai.v35i17.17745>

[2] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and theproblemof offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512- 515. <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>

[3] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., &Çltekin, C. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media(OffensEval 2020). Proceedings of the Fourteenth Workshop on Semantic Evaluation, 1425- 1447. <https://aclanthology.org/2020.semeval-1.188/>

[4] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hatespeechdetection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), ComplexNetworks and Their Applications VIII (pp. 928-940). Springer. https://doi.org/10.1007/978-3-030-36687-2\_77