Hate Speech Identification Case Study

A DS 4002 Case Study by Vivian Jiang

Social media platforms shape global conversations, yet they also struggle with the rapid spread of harmful content. Misclassifying posts, either by leaving hateful messages online or censoring harmless ones, can damage public trust and put vulnerable communities at risk. Unlike profanity detection, identifying hate speech requires understanding *context*: who is targeted, how language is used, and whether meaning shifts across speakers or communities.

Hate speech and offensive language often overlap, but they are not the same. A single word may be playful in one setting and deeply harmful in another. This complexity makes automated moderation difficult and highlights the need for more reliable, transparent approaches. Recent developments in natural language processing, especially transformer-based models like BERT, offer new ways to capture nuance that simpler models miss.

In this case study, you will step into the role of a data scientist tasked with analyzing real tweets from the CrowdFlower Hate Speech Identification dataset. Your goal is to explore linguistic patterns that differentiate hate speech, offensive language, and neutral content. You will then build and compare models, ranging from interpretable baselines to context-aware transformers, to evaluate how effectively these patterns can be detected. Finally, you will interpret your results and consider how annotation subjectivity and data limitations affect model performance.

Your final deliverable will be a concise analysis that communicates which features matter, how your models perform, and what your findings imply for automated content moderation. Full details are provided in the rubric.

References:

Ingale, P. (2025). *Training Machines to Tackle Hate: My Master's Research on Hate Speech Detection.* Medium.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*