DATS_6312_NLP Group Proposal

Team 3: Jonathan Schild, Thomas Stanton, Ashish Telukunta

Online Community Monitoring
Tracking Changes in Subreddit Sentiments

Topic Selection and Justification: We are going to work on a project to analyze sentiment dynamics in online communities, specifically within Reddit subreddits. This topic is crucial for understanding how digital discourse shapes public opinion and identifying potential misinformation.

Database/Dataset: Our data will come from two sources: a Hugging Face dataset with approximately 90,000 labeled tweets and a Kaggle dataset with about 8,500 Reddit comments. We will also use PRAW (Python Reddit API Wrapper) to scrape additional Reddit posts for a more comprehensive analysis.

NLP Methods: We plan to employ three transformer-based models: BERT, RoBERTa, and Electra, to conduct a comparative analysis (qualitative and quantitative) of their effectiveness in sentiment analysis. This will allow us to understand which model is most efficient for our specific use case.

Packages for Implementation: Our primary tools will include Hugging Face for model access and Streamlit for application development. We'll also use PRAW for data scraping from Reddit.

NLP Tasks and Analysis: The primary task is sentiment analysis, focusing on detecting emotions like joy, sadness, anger, fear, love, and surprise. We will average sentiment scores to plot average sentiment trends over time, providing a dynamic view of community sentiment shifts.

Performance Metrics: Our models will be evaluated using F1 scores, accuracy, and composite F1 scores for various emotions. These metrics will help in understanding the precision and effectiveness of our models.

Project Schedule: The project will involve stages of dataset preparation, model training and fine-tuning, application development, and performance evaluation. This will be followed by a comparative analysis of the three models and an in-depth evaluation of sentiment trends.