**Visualizing Changes in Sentiment of Online Communities**

Reddit is a popular social media forum that attracts significant numbers of users, content, and engagement from around the globe. Its ever-growing volume of online community discourse presents a platform that can inadvertently shape opinions, damage brands, and incite real-world action. Identifying and understanding swings in user sentiments is becoming more and more critical to combat inaccurate and inadequate information and general community divisiveness. Our application allows subreddit moderators to monitor their online communities for significant swings in sentiment that could have broadly negative implications if left unchecked.

**Datasets:** We used two datasets to train our final model(s). Our models primarily use a dataset from Hugging Face that contains approximately 90,000 labeled tweets distributed evenly across six emotions: sadness, joy, love, anger, fear, and surprise.[[1]](#endnote-1) We also use a dataset available on Kaggle that contained almost 8,500 unlabeled Reddit comments from various subreddits.[[2]](#endnote-2) This dataset was used to experiment with pseudo-labeling and incorporating confident test data into our initial Twitter training data.

**Models:** We experimented with different neural network architectures during our analysis out of suspicion that predictions of emotions and sentiment could vary depending on the type of model implemented. Each model tuned hyperparameters like learning rate, batch size, dropout rate, and decay rate. All of them also used the Adam optimizer with the sparse categorical cross entropy loss function, and monitored validation accuracy for early stopping.

We developed two BERT-based multi-class classifiers for emotion detection in text, as it learns contextual word representation beneficial for informal language. Model 1 (BERT1) was trained on Twitter data split into training, validation, and test sets. BERT1 achieved validation accuracy and F1 of 0.95. To improve Reddit generalization, we applied BERT1 to unlabeled Reddit comments to pseudo-label high confidence examples (threshold 0.6/label) and combined these newly labeled samples with the Twitter training set. Using this semi-supervised dataset, we trained Model 2 (BERT2) with the same BERT architecture. Incorporating Reddit data reduced performance (validation accuracy/F1: 0.91) but improved applicability to this target domain.

We utilized a RoBERTa sequence classifier for our second model, motivated by its optimization of BERT's pre-training procedures and ability to comprehend complex syntactic structures prevalent in Reddit comments. Solely trained on the Twitter corpus, it achieved validation accuracy and F1 scores of 0.95. To diverge from BERT-based models, we trained an ELECTRA model capable of learning more granular and nuanced language patterns ubiquitous in social media. Unlike BERT and RoBERTa which rely on masked language modeling, ELECTRA is context-sensitive, enabling more precise sentiment predictions suitable for the intricacies in this domain. Our ELECTRA model obtained a validation accuracy and F1 score of 0.95, comparable to our BERT and RoBERTa models.

**Implementation:** We used Streamlit[[3]](#endnote-3) to apply our models and display visualizations that help a user identify and understand changes in sentiments of online discourse, in this case subreddits. A scraper using PRAW fetches Reddit posts in real-time based on inputs provided by the user, such as subreddit, number of posts and comments, and date-range criteria. The Reddit API limits scraping to the past day, week, month, or year, instead of providing the ability to specify a date range. to ensure we could measure changes over time, we pulled the dates for each post and comment and later grouped them according to a daily, weekly, or monthly intervals extending out from the present. This allowed us to average sentiments together over the specified interval and visualize the change over time.

After pulling in the live data, our application displays a histogram that allows the user to gauge the number of comments and posts per interval to ensure certain time periods do not contain missing data and that they do not contain and significantly more or less data than other periods. We then provide the user the ability to select and apply the data using our four models (BERT1, BERT2, RoBERTa, ELECTRA). The application returns two charts: one showing the six emotions plotted on a line chart by time-interval averages, and a second chart showing the positive and negative sentiment over the same time-interval averages for easy evaluation.

**Observations:** All of our models identified and predicted similar trends in the changes of sentiment; i.e., they each displayed similar deltas in emotion across time. However, while the BERT-based architectures provided robust accuracy, qualitative analysis showed ELECTRA better captured contextual emotion semantics. This was evident when predicting sentiment across different subreddits using all four of our models and comparing their visualizations. Examples with the “funny” and “wallstreetbets” subreddits show how each model identifies similar patterns, and how the ELECTRA betters captures emotions that would be expected from more positive subreddits.

Figure 1: r/funny

A graph of different colored lines and dots

Description automatically generated with medium confidence

Figure 2: r/wallstreetbets

A group of graphs with red circles and numbers

Description automatically generated

1. https://huggingface.co/datasets/philschmid/emotion/tree/main/data [↑](#endnote-ref-1)
2. https://www.kaggle.com/datasets/prakharrathi25/reddit-data-huge/data [↑](#endnote-ref-2)
3. https://streamlit.io/ [↑](#endnote-ref-3)