**Gauging Changes in Sentiment Using Emotional Classification**

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. OUTLINE OF REPORT.

**Our Data:** We used two datasets to train our final model(s). Our initial models use a dataset from Hugging Face that contains approximately 90,000 tweets that are labeled according to evenly distributed 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)

**Models:** We experimented with multiple models during our analysis out of suspicion that predicting emotions and sentiment could vary depending on the type of model implemented. Each model fine-tuned on learning rates, batch sizes, dropout rates, and decay rates, used the Adam optimizer with the sparse categorical cross entropy loss function. They also monitored validation accuracy for early stopping.

Our first model (Bert1) consisted of a standard Bert-based multi-class classifier, trained on our initial Twitter dataset using training, validation, and testing splits. Bert allows us to understand the context of a word based on all its surroundings (both left and right of the word), which is useful for capturing the contextual nuances of language that can be informal and idiosyncratic. This initial model achieved validation accuracy and F1 scores of .95, each. However, since we intended to eventually apply the model to Reddit data, we wanted to introduce Reddit comments into the training dataset to sensitize the model to the language expression and structure in Reddit comments that could differ from what is observed in Tweets. We applied Bert1 to unlabeled Reddit comments to predict the comments’ scores for each of our six emotion labels. By filtering out comments that did not have any emotion score above .6 we were able to supplement our labeled Twitter data with newly labeled Reddit data, to create a new training set. We used this new training data to train, validate, and test a new, similarly configured Bert model (Bert2). Adding in the Reddit reduced our accuracies when compared to Bert1, with validation accuracy and F1 score of .91, each

We used RoBerta sequence classifier for our second model, primarily because of its optimization of Bert’s training procedures and ability to understand complex sentence structures that are more likely to appear in Reddit comments. It was trained solely on the Twitter data and achieved similar accuracy scores as Bert1: validation accuracy and F1 scores of 0.95, each. Lastly, we sought to lean away from Bert and trained an Electra model that can learn more granular and nuanced patterns of the type of language that is often be found in social media. Unlike Bert and RoBerta, which primarily learn from masked words, Electra can offer more accurate and context-sensitive sentiment predictions that make it more suitable for the complexities and nuances found in social media texts. Our model this time also obtained similar accuracies as our Bert1 and RoBerta models, with a validation accuracy of 0.95 and F1 score of 0.95, each.

**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:** After using our application and the four model options to examining emotions, we noticed a particular distinction in the visualizations. First, and as expected, all of our models did extremely well in identifying and illustrating trends in the changes of sentiment; i.e., they each displayed similar deltas in emotion across time. However, the Bert models (Bert1, Bert2, and RoBerta) appear to label text with emotions that differ from the Electra model. While further analysis is required to hone in on the granular reason for the disparity, we strongly suspect it is related to our initial belief that different models, particularly Electra, could better capture and identify contextual relationships in social media texts.

As noted further above, all four of our models performed well from an accuracy perspective and identify similar trends across the board, but the Bert models seemed to identify emotions differently than the Electra model. Bert1, Bert2, and RoBerta consistently return similar emotional identities and scores for texts over time, while Electra seems more nuanced. After testing our models across subreddits and when considering the purpose of those subreddits, Electra’s ability to better capture nuanced meaning in text became more apparent (see Figure 1 and Figure 2). While accuracy metrics were generally equal for all of our models, Electra did a better job of matching the expected emotion of specified 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)