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 administrators to monitor their online communities for significant swings in sentiment that could have broadly negative implications if left unchecked. OUTLINE OF REPORT.

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]](#footnote-1) We also use a dataset available on Kaggle that contained almost 8,500 unlabeled Reddit comments from various subreddits.[[2]](#footnote-2)

We experimented with multiple models during our analysis all of which implemented fine-tuning on learning rates, batch sizes, dropout rates, and decay rates. Each model also used the Adam optimizer with the sparse categorical cross entropy loss function, and monitored validation accuracy for early stopping.

* Our first model (model1) consisted of a standard Bert-based multi-class classifier, trained on our initial Twitter dataset using training, validation, and testing splits. This initial model achieved validation accuracy and F1 score of .95, each. However, since we intended to apply the model to Reddit data, we wanted to introduce Reddit comments into the training dataset so the model is sensitized to Reddit comment expression and structure. In order to get Reddit data with labels that matched our Twitter dataset, we applied our initial model to unlabeled Reddit comments that predicted scores for each of our six emotion labels. We then removed comments that did not contain at least one emotion with a score higher than .6 to increase our confidence that the comments contained at least one detectable emotion. These now-labeled Reddit comments were added to our labeled Tweets to create a new combined dataset. After training a similar standard Bert-based multi-class classifier, this new model (model2) achieved a validation accuracy and F1 score of .91, each.
* ELECTRA
* ALBERTA

6. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

In summary, we trained two models to predict the emotion and general sentiment of Reddit posts/comments over time, and created an application the displays the change in attitudes over a specified time frame. This tool would offer the most value to administrators and moderators of online communities where text-based communication is both frequent and engaging. They would be able to regularly monitor for abrupt changes to typical behaviors and views in discourse that could threaten the integrity of the subreddit and its reputation. Looking forward, our model would benefit from additional training on more Reddit-specific data. As is, Tweets dominated the training process which could have negatively impacted the model’s predictive power on

8. References. In addition to references used for background information or for the written portion, you should provide the links to the websites or github repos you borrowed code from.

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