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 that contained almost 8,500 unlabeled Reddit comments from various subreddits.

3/4/5. Description of the NLP model and what kind of algorithm did you use. Provide some background information on the development of the algorithm and include necessary equations and figures. Experimental setup. Describe how you are going to use the data to train and test the model. Explain how you will implement the model in the chosen framework and how you will judge the performance. What kid of hyper-parameters did you search on? (e.g., learning rate)? How will you detect/prevent overfitting and extrapolation?

* 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. 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
* We first trained it on our twitter dataset that had the six aforementioned emotions as labels. After achieving validation accuracy and F1 score of .95 for each, we applied the model to our Reddit dataset[[2]](#footnote-2) to label each comment according to its dominant emotion, with the emotions accompanying . We then examined the newly labeled Reddit dataset, removed comments that had a dominant emotion with a confidence score of less than 0.6, and added the comments to the original Twitter dataset. With our new dataset containing labeled Tweets and Reddit comments, we trained a new model, which achieved a validation accuracy of .91 and F1 score of .91. In addition to the emotions, we used the Hugging Face pipeline to derive positive and negative sentiment scores on each text.
  + During training and fine tuning, we experimented with different batch sizes, learning rates, dropout rates, and decay rates. We also used the Adam optimizer with the sparse categorical cross entropy loss function, while monitoring validation accuracy for early stopping.
* ELECTRA MODEL
  + In this case we did not

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.

7. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

* 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.

* #### Bert-based multiclass classifier is trained on approximately 90,000 tweets to a .95 validation accuracy and .95 F1 score. This model is then applied to a separate dataset of unlabeled comments. The results are filtered to only retain the reddit comments that contained an emotion with above a .6 "confidence" score. These comments and their new labels are then combined with out original twitter dataset, and used to train a new model. This second model achieves a validation accuracy of .91 and F1 of .91. The apply file loads the model and applies it to some input'ed text.
* #### Implement a scraper to get comments from Reddit posts (live) to be fed to the model. Display the sentiments (emotions, pos/neg) of the text specified by the user of the application. The application will also display the general change in sentiments for a given subreddit over X amount of time.
* #### This application is intended to illustrate changes in sentiments of a given subreddit over time. It can assist moderators identify times when intervention might be warranted, or predict and prepare for events that could have longlasting negative impacts on the direction of their subreddit.

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)