Signals of Mental Health in Twitter Language

Group 18

Taylor Sehnem 406109624
and Andrew Drabkin 406174488

Abstract

The goal of our project was to be able to identify and label plain twitter tweets that were scrapped as anxious, normal, stressed, or lonely. We started with a data set that has pre labeled tweets to train and build a model that we could run our unlabeled tweets through to provide a label to. It required a lot of data cleaning within the groups. We tokenized them, removed stop words, punctuation, and common words. Then we created a TF-IDF weighting to capture the significant meaningful words in each group that led them to be in the emotional category they are in. We computed the cosine similarity between the tweets with four thresholds to identify any tweets that share similar language. Because of the large sample size, we had to randomly select 50 tweets from each category of each emotion and combine them into a dataset and run a model on that set. We ran that 50 times and took the average evaluation score. Then we used three approaches to find the best accuracy and determined that doc embeddings and random forest yielded us the highest accuracy. So then with the unlabeled data set we applied that method to predict an emotional label on each tweet after applying similar data cleaning. After we had the labels we used BERTopic to create 10 clusters and we were able to observe and analyze the gender breakdown within each cluster. Our main findings were that what we could see in the UMAP projections showed the emotional states distributed across all topics demonstrating a relationship between language use and emotional content.

Introduction

Our motivation behind this project came from an article we read in the World Health Association. They estimate that 1 in every 8 people are living with a mental illness. This problem is becoming more and more prevalent and there is large research on the subject. When mental illnesses are recognized early, better steps and support can be offered, helping individuals avoid a downward spiral and preventing the condition from worsening. One real world event that inspired us was people being isolated during the covid-19 pandemic. Everyone, including both of us, were separated from friends and some family members and we were in a way forced to turn to social media as a form of social interaction. During this time, a spotlight was turned to the fact that some of the issues that people faced were being brought out and expressed to others online. Twitter is a great place to study these hypotheses due to its “real-time nature, high availability, and large geographical coverage” [6]. The National Library of Medicine conducted a study to evaluate the feasibility of using Twitter to estimate the prevalence of mental disorder symptoms such as anxiety and depression. They did this by examining correlations between Twitter-based data and national surveillance statistics, as well as tracking how these correlations evolved over time [6]. This was something that spurred us to do our own study. After digging even deeper into the topic, we realized that this idea has been around for a long time. In 2014, John Hopkins University did their own study on the relationship between tweets and mental illness. They focused specifically on PTSD and depression. Interestingly enough, they even had a hackathon with the data they collected on this very topic. This backs up the claim “using quantitative techniques to predict the presence of specific mental disorders” can be done and is a valid way of finding symptoms early [2]. This line of thinking is not new, but with the new frontier of machine learning, we seek to run our own study that will hopefully recognize signs of mental illness in tweets which can lead to detection, better-targeted support, and improved mental health outcomes. So with our interest in mental health and previous work to back up the platform we chose to study, we wanted to study how tweets on twitter can be analyzed to identify early mental health signs like stress, anxiety, and loneliness. This is a new approach to the field because there is limited research in the early detection of mental health struggles on twitter. We were able to obtain over 20,000 tweets and run a model to apply a label of prediction for stress, anxiety, loneliness, and normal (which included every emotion but the three we were interested in studying) on plain text within minutes. The data set we acquired had some demographics on the user that allowed us to dig deeper into what we had labeled and uncover any trends within emotions. We will review some previous studies that have done related research on our topic, dive deeper into the data set and methods, present our analysis, and finally discuss the implications of our findings and where we could go in the future and what we were limited on for this paper.

Related Works

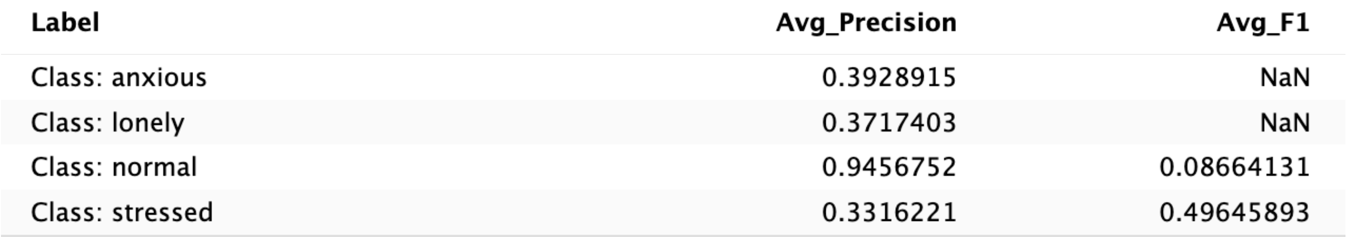
There is a growing body of research surrounding the topic of mental health and how language patterns on social media can act as a proxy for mental health status, like anxiety, loneliness and stress. Coppersmith et al. performed early work demonstrating that Twitter that can self detect themselves as an individual having mental health conditions exhibited linguistic patterns that can be predictable - such as negative emotional words and elevated use of single person pronouns [1]. Continuing this, Chancellor and De Choudhury gave a critical review of predictable methods that can be applied to supervised and unsupervised models to infer an individual's mental health status using contextual language clues, user demographics, and temporal patterns [2]. Another study on postpartum depression was one of the first that used machine learning and linguistic analysis to predict a mothers risk based solely on social media activity [3]. This study was notable for its methods but also its consideration for ethics and demographic variability. While social media can provide a lot for predictive pattern research, the privacy and protection of the user is also something that needs to be balanced. Exploring Reddit and Twitter discourse, De Choudhury and colleagues demonstrated that an online expression of struggle could mirror an offline serious mental health condition , suggesting that social platforms can serve as a source for mental health surveillance [4]. And while these researches provided huge context into the field, much of the work surrounded detecting depression or PTSD, limiting its attention towards emotions like loneliness and stress. Furthermore, our data set contained dimensions such as gender and bot inference which are underexplored. Our project aims to address these gaps by including emotional states such as stress and loneliness alongside with anxiety and normal and by using BERTopic modeling to associate emotional content with clusters of user identity. We contributed to this growing field by showing that a plain text tweet, when labeled using a random forest classifier on trained pre labeled emotional tweets, can reveal enriching patterns in emotional expression. The detection of clusters based on common words can be broken down into gender breakdown which includes institutional voices aligns with De Choudhury's finding but extends the research by incorporating clustering techniques and user demographic breakdown. Our work reinforces the hypothesis that twitter language can portray emotional states highlighting the potential our tweets hold in early detection.

Data and Methods

Our data came from the popular website kaggle, known for good data sets. The first source of data is titled "Behavioural Tweets." This page contained four data sets each prelabeled with the four following emotions: normal, stressed, lonely and anxious. Each emotion had close to 10,000 tweets each. The second data set was titled "Twitter User Gender Classification." The owners of this data set randomly scraped twitter and then classified them into a gender. Tweets that didn't cleanly fit into male or female or where suspected of being bots were labeled as Brand.

The tweets were already cleaned, so we went straight into training a model to be able to classify tweets into different emotions. After joining the four data sets of tweets plus emotions into a single data frame, we immediately ran into some issues. When trying to convert the tweet text into a numerical representation using the TF-IDF technique, we constantly encountered errors saying R had to abort the process due to excessive memory and processing time requirements. We attributed this to the joint data frame containing about 40,000 rows. This was because the confusion matrix would be n x n, and in other words way too big, and it did not even work with the doc embedding either. With both of us being statistics majors, we knew immediately how to fix this. To get the joint data frame smaller, we randomly sampled 50 tweets from each emotion *before* combining them. This ensured that there would still be an equal amount of observations of each emotion, but to a size that we knew the models would be able to perform on. But a random sample like that could potentially still have bias, so to try and remove this, we repeated sampling the 50 tweets fifty times, trained the model on the smaller set, and then logged the performance results of F1 scores, recall, precision, and accuracy. Because we took a small random sample and repeated it, we had no missing or N/A values in our data set used to train the model so there was no imputation method to choose. After repeating 50 times, we took the averages of each and used this when deciding which model to move forward with.

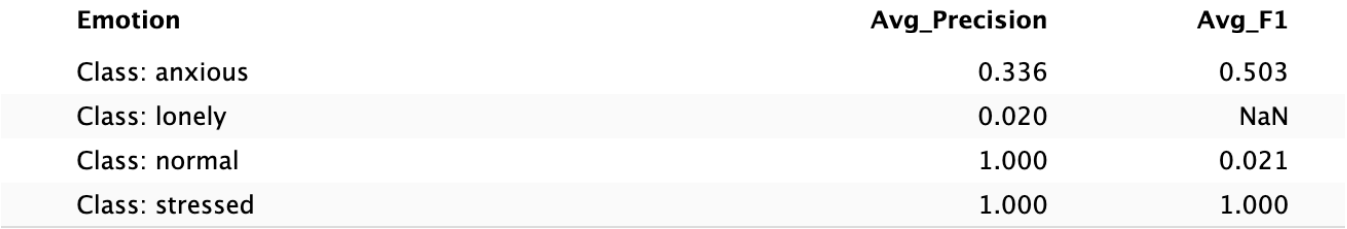
Three three models we considered were doc embedding, TF-IDF, and doc embedding with random forest. Like we mentioned before the first step was random sampling 50 tweets for each emotion. Then we tokenized the tweets, constructed a term co-occurrence matrix , and used it to train GloVe word embeddings. Each tweet was then converted into a document embedding by averaging the GloVe vectors of the words it contains. We then used k-means clustering to group the document embeddings into four clusters. After matching up the clusters to their prelabeled emotions we computed the performance metrics, repeated this process 50 times, and took averages of each.



Label	Avg_Precision	Avg_F1
Class: anxious	0.3928915	NaN
Class: lonely	0.3717403	NaN
Class: normal	0.9456752	0.08664131
Class: stressed	0.3316221	0.49645893

Table 1: Average precision and average F1 scores of the GloVe word embedding model

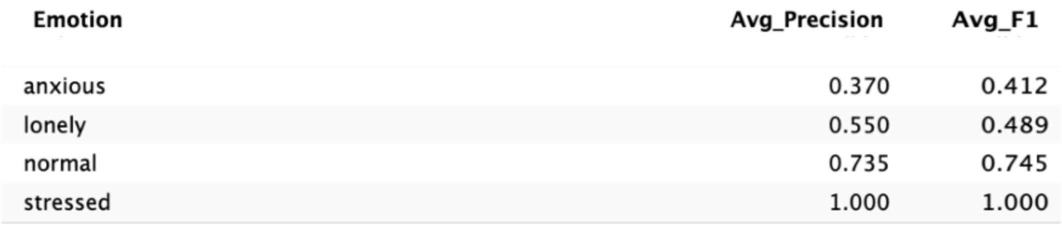
Next we did the TF-IDF model. The tweet text from the combined data frames is tokenized and transformed into a document-term matrix using a vocabulary constructed from the data. This document term matrix is then transformed using TF-IDF weighting, which assigns importance scores to words based on their frequency in individual tweets relative to the whole corpus. Then, same as above, we used k-means clustering and assigned the clusters to their emotion. We then repeated this process 50 times and took average metrics.



Emotion	Avg_Precision	Avg_F1
Class: anxious	0.336	0.503
Class: lonely	0.020	NaN
Class: normal	1.000	0.021
Class: stressed	1.000	1.000

Table 2: Average precision and average F1 scores of the TF-IDF model

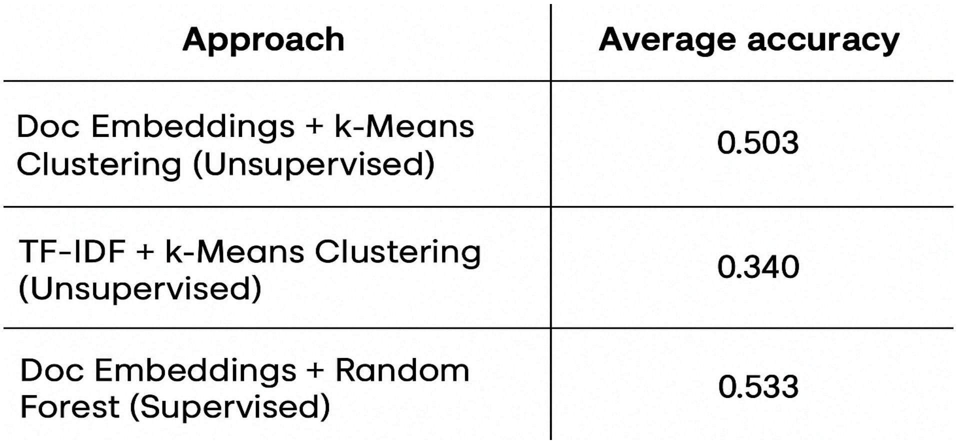
The final model we made was with supervised learning on the doc embeddings. We constructed a term co-occurrence matrix which was used to train GloVe word embeddings similar to before. This was done with the text2vec package in R. Each tweet was embedded by averaging the GloVe vectors of the words it contained, producing fixed-size document embeddings for each tweet. These embeddings were then combined into a feature matrix, where each tweet was labeled with its corresponding emotion category. Next came the supervised part. We split the resulting dataset into an 80% training set and a 20% test set. A Random Forest classifier was trained on the training set using 5-fold cross-validation. The package we used for random forest was the ranger package. The parameters were tuned to avoid overfitting and underfitting where the arguments mtry and min.node.size. Mtry affects the number of variables randomly selected at each split and min.node.size changes the minimum number of observations required in a terminal node. The model was tuned to maximize the F1 score, which balances precision and recall across all emotion categories. The trained model was then used to predict the emotion labels of the test set. We encountered a limitation when attempting to repeat the process 50 times, as the model was too computationally intensive for our computer to handle, causing R to terminate the execution. We could only repeat it 5 times before it would terminate the call, so these results are only the average of those 5 samples. While this is not ideal, we still ran this a couple and saw little changes in the scores below.



Emotion	Avg_Precision	Avg_F1
anxious	0.370	0.412
lonely	0.550	0.489
normal	0.735	0.745
stressed	1.000	1.000

Table 3: Average precision and average F1 scores of the word embedding + Random forest model

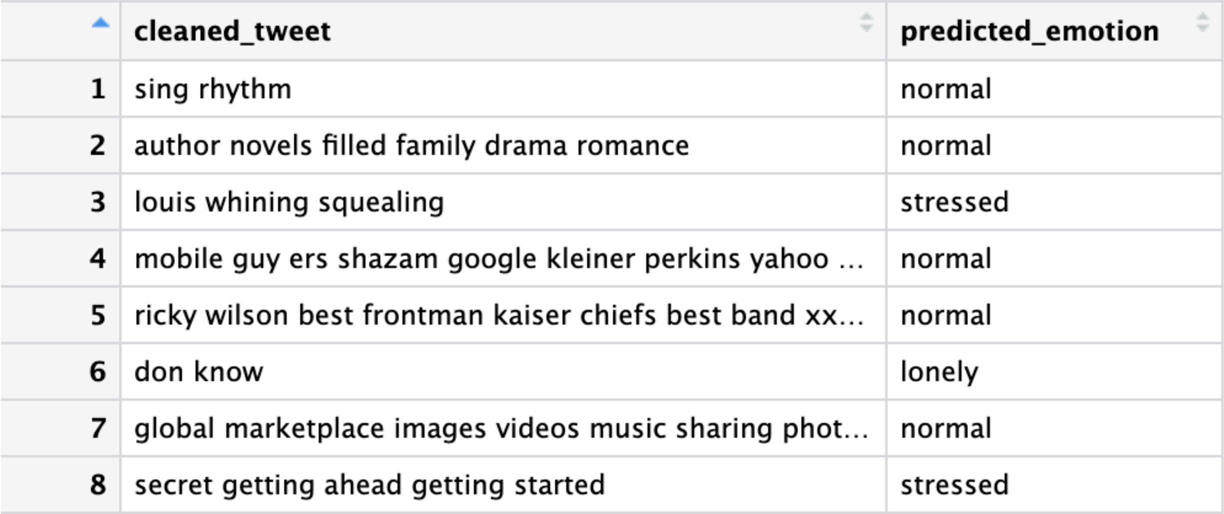
The accuracy of the three methods is shown in the table below.



Approach	Average accuracy
Doc Embeddings + k-Means Clustering (Unsupervised)	0.503
TF-IDF + k-Means Clustering (Unsupervised)	0.340
Doc Embeddings + Random Forest (Supervised)	0.533

Table 4: Average accuracy scores by model approach

Because the random forest unsupervised model had the highest accuracy score, we decided to move forward with it. Now that we did not need to repeat the model, we randomly sampled 200 tweets from each emotion to make it even more accurate. Once again training the model with a test set and tuning hyperparameters like mtry and min.node.size we created a better model that we liked. We then applied this trained model to a new dataset of tweets stored in a data frame called new_tweets. To ensure consistency in preprocessing, each tweet was cleaned by lowercased, stripped of punctuation and digits, and tokenized. Because the new dataset was large, we processed it in chunks of 1,000 tweets at a time to improve computational efficiency. For each chunk, we predicted the emotion category using the saved Random Forest model. The predicted emotion labels and cleaned text were added back to the original new_tweets data frame for further analysis. We now had a column called cleaned_tweet and predicted emotion, as well gender. Below is a sample of what the data frame looked like with the cleaned_tweet and predicted emotion column.



	cleaned_tweet	predicted_emotion
1	sing rhythm	normal
2	author novels filled family drama romance	normal
3	louis whining squealing	stressed
4	mobile guy ers shazam google kleiner perkins yahoo …	normal
5	ricky wilson best frontman kaiser chiefs best band xx…	normal
6	don know	lonely
7	global marketplace images videos music sharing phot…	normal
8	secret getting ahead getting started	stressed

Figure 1: Sample of new_tweets data frame with the columns cleaned_tweet and predicted_emotion

Now that we have labeled tweets by emotion, we could start finding clusters and similarities of the topic in the new tweets. We used sentence-transformer embeddings via the Python sentence-transformers library. Using the all-MiniLM-L6-v2 model, we generated BERT embeddings for each tweet's cleaned text. These embeddings placed the semantic meaning of each tweet into a dense vector format, which allows better clustering and visualization compared to traditional approaches like TF-IDF. After generating embeddings for all tweets, we applied k-means clustering with 10 centers to group tweets into similar clusters. We chose 10 because after running the elbow method we plotted the total within-cluster sum of squares for different values of k and the elbow is around k = 10. Each tweet was assigned a cluster ID based on its proximity to cluster centroids. These cluster assignments were saved to the new_tweets data frame which we could then make observations about.

To understand what each cluster represented topically, we used tokenization and frequency analysis. After filtering out standard stopwords, we tokenized the cleaned tweets and identified the top 10 most frequent words per cluster. These word frequencies were visualized using bar plots, allowing us to qualitatively inspect the dominant themes or vocabulary within each cluster. This provided interpretable insight into what each group of tweets was broadly about.

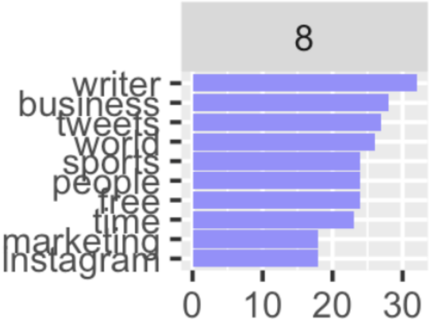


Figure 2: The list of the top 10 most common words in cluster 8

Next, we examined how the emotions were spread across the clusters derived from the BERTopic. For each cluster, we calculated the proportion of tweets associated with each predicted emotion label (normal, anxious, lonely, or stressed). The results were plotted using a bar chart, highlighting which emotional tones dominated specific semantic clusters. This helped reveal whether certain topics were more strongly associated with specific emotions.

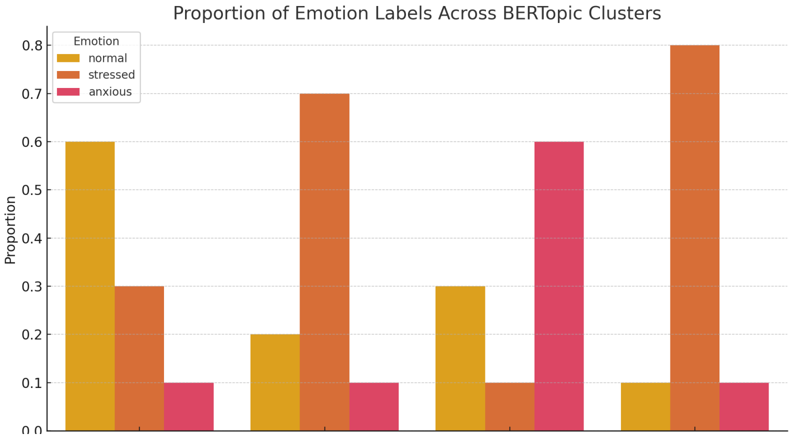


Figure 3: Example cluster emotional distribution for cluster 1-4

To visualize the high-dimensional structure of the tweet embeddings and the clusters they formed, we used Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. We projected the 384-dimensional BERT embeddings into a two-dimensional space. In one plot, we colored the points by their k-means cluster assignments (figure 4). In another plot each cluster is grouped by their predicted emotion labels (figure 5). These visualizations provided an understanding of how semantically similar tweets grouped together and how emotions were distributed across those groups.

Finally, we focused on tweets predicted to be stressed, investigating how gender identities were represented across BERT clusters. After filtering for tweets labeled as stressed and belonging to valid gender categories (male, female, and brand), we calculated the proportion of each gender within each BERT cluster. We made a bar plot to show how gender distributions varied by topic cluster, offering insight into which stressed themes were more prominent among different gender groups. This will be discussed more in the results section as well.

Results

Our project revealed several key insights into how emotional states are expressed in Twitter language and how they cluster topically using semantic embedding techniques. After applying this model to our unlabeled dataset, we predicted emotional labels—normal, anxious, stressed, and lonely—for over 20,000 tweets. We then used BERTopic (via BERT embeddings and k-means clustering) to assign each tweet to one of 10 semantic clusters. The elbow method confirmed that 10 was a reasonable choice, as the within-cluster sum of squares began to level off at this value. We next examined how predicted emotions were distributed across these semantic clusters. A grouped bar chart (Figure 3) showed that while normal tweets were the majority in each cluster, stressed and anxious tweets tended to dominate certain clusters—especially those centered around topics related to economic concerns and time pressure. This finding supports prior research suggesting that online expressions of psychological distress often emerge in topic-specific ways [1]. A UMAP projection of tweet embeddings (Figure 4) visually confirmed that clusters with high emotional content (e.g., stressed or lonely) tended to form semi-distinct regions within the semantic space, further supporting the model's ability to capture meaningful differences in emotional tone.

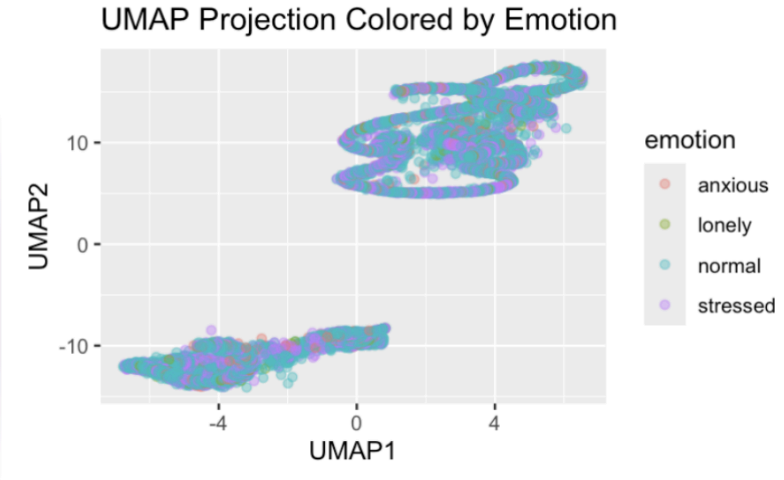


Figure 4: UMAP projection by Emotion

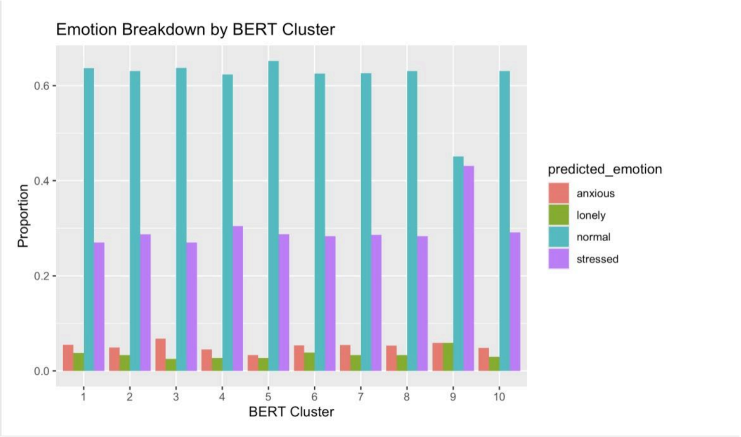


Figure 5: Emotion Breakdown by each BERT Cluster

As seen in figure 5, tweets labeled as normal are the most relevant in each cluster, which makes perfect sense. Stressed tweets are runners up in each category. To explore demographic differences, we filtered the dataset to examine only stressed tweets and assessed their gender breakdown by cluster (Figure 6). We found that female-identified users were more likely to appear in clusters centered on interpersonal and self-reflective language, while brand accounts (likely bots or organizational voices) dominated clusters focused on policy, economy, or news.

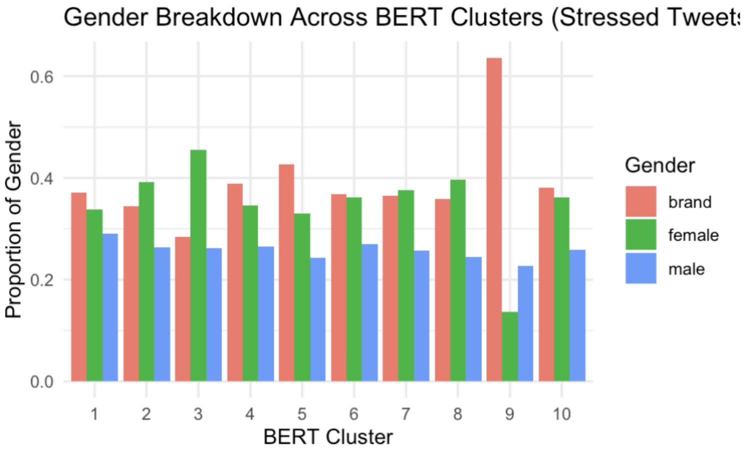


Figure 6: Gender Breakdown across clusters filtered by stressed tweets

These results affirm the feasibility of using Twitter language to detect signs of emotional stress and mental health concerns. They also demonstrate how emotion-labeled tweets cluster meaningfully into topics that reflect distinct emotional and demographic trends.

Discussion

One interesting observation that our results show is that the Doc Embeddings + Random Forest approach achieved the highest accuracy over Doc Embeddings + K means and TF-IDF + K means as well as both doc embedding methods have significantly higher averages than the TF-IDF. When researching and looking through notes on each method, we remembered how document embeddings represent each tweet as a dense, continuous vector that takes into account semantic relationships between words. This is different from TF-IDF, which treats words as independent and sparse matrices. We think that the advantage of random forest over k-means came from this method using supervised learning. Unlike unsupervised clustering, which tries to group without guidance, Random Forest directly based new predicted labels from the examples of the test set. This gives it a clear advantage in distinguishing between similar emotional categories based on the patterns in the embeddings. Our results confirm our original intuition and prior studies that mental health expression online is contextually and demographically shaped, with women more likely to share personal emotions publicly [3] and bot accounts often skewing emotional content in predictable ways (Fiorentini et al., 202). This supports prior research indicating that individuals often express psychological distress through subtle language markers such as word choice, tone, and topic focus [1] [2].

Limitations and Future Work

While our study demonstrates promising results in detecting emotional state through twitter language, there are several limitations that must be addressed. First, the emotional labels were pre-assigned and may reflect subjective categories by the user. Because we did not validate the labels with psychiatrists or clinical diagnostics, the emotional labels are best understood as a linguistic label rather than a definitive psychological diagnosis. Secondly, the demographic information on our user was limited to gender and location the tweet was posted. We were not given an age, time stamp, or cultural context on our tweets. Additionally, our approach involved training on small samples, only 50 tweets per emotional category, which helped with our computational efficiency but limits the generalizability of our study. Moreover, we assigned our emotional labels based on textual data only, not able to incorporate emojis, images, or likes/replies/retweets, which are an extremely relevant part of social media analysis. In the future we want to incorporate longitudinal data that would allow for us to detect emotional changes over time and reveal early warning signs before a mental health event takes place. Secondly, if we could integrate features such as time of day, images, emojis, etc to enhance our models sensitivity. Lastly, if we incorporated a larger more diverse data set with additional languages, we could improve our models' relevance to this field. Additionally, we should consider the ethical concerns that surround this topic such as consent for information to be revealed and addressing the risk of misclassification for future iterations.

References

[1] Coppersmith, Glen, Mark Dredze, and Craig Harman. 2014. “Measuring Post Traumatic Stress Disorder in Twitter.” *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014. https://doi.org/10.1609/icwsm.v8i1.14574

[2] Chancellor, Stevie, and Munmun De Choudhury. 2020. “Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review.” *NPJ Digital Medicine* 3 (1): 1–11. https://doi.org/10.1038/s41746-020-0233-7.

[3] De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013. “Predicting Postpartum Changes in Emotion and Behavior via Social Media.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267–3276. https://doi.org/10.1145/2470654.2466447.

[4] De Choudhury, Munmun, and Emre Kıcıman. 2017. “The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk.” *Proceedings of the International AAAI Conference on Web and Social Media* 11 (1): 32–41. https://ojs.aaai.org/index.php/ICWSM/article/view/14866.

[5] Fiorentini, C., Cotter, K., McNeill, A., & Chou, W.-Y. S. (2020). Bots and misinformation on Twitter: Examining the intersection of automation and public health messaging. Journal of Medical Internet Research, 22(5), e19361. https://doi.org/10.2196/19361.

[6] Zhang, Cai, Rui,, Zhiyu Li, Chengdong Zeng, Sijia Qiao, and Xudong Li. 2022. “Using Twitter Data to Estimate the Prevalence of Symptoms of Mental Disorders in the United States During the COVID-19 Pandemic: Ecological Cohort Study.” *JMIR Formative Research* 6 (12): e37582. https://doi.org/10.2196/37582.