Detecting Depression in Social Media with NLP Models Trained on Journal Entry Data

Tvisha Choubey

Abstract

Writing has always been recognized as a powerful means of expressing human emotions, serving as a reflective practice that allows individuals to process and articulate their inner experiences. However, with the rise of social media, the landscape of emotional expression has shifted. This transition from private journaling to public social media posting raises important questions about how effectively these platforms serve as emotional outlets and what they reveal about users' mental health, specifically with pervasive mood disorders like depression, which affects over 18 million adults in the United States. Recently, NLP models have been noted as a promising tool for detecting underlying sentiment in text. This research explores how the Twitter posts of individuals suffering from depression compare when analyzed using a natural language processing (NLP) model trained on journal data classified by emotions. Two separate clustering approaches were used to reduce dimensionality in training data and train machine learning models, one with spectral clustering and principal component analysis (PCA), and the other with the Natural Language Toolkit (NLTK) library. The results of both machine learning approaches, with accuracy over 99%, demonstrated that tweets of depressed Twitter users are classified as more negative compared to those of non-depressed users. These findings suggest that the emotional content expressed in social media posts by individuals with depression is consistently more negative, aligning with the patterns observed in their journal entries. Ultimately, this research highlights the evolving role of social media as a platform for emotional expression and its implications for mental health monitoring.

Keywords

Natural Language Processing, depression, Twitter, spectral clustering, principal component analysis, Natural Language Toolkit library, sentiment analysis

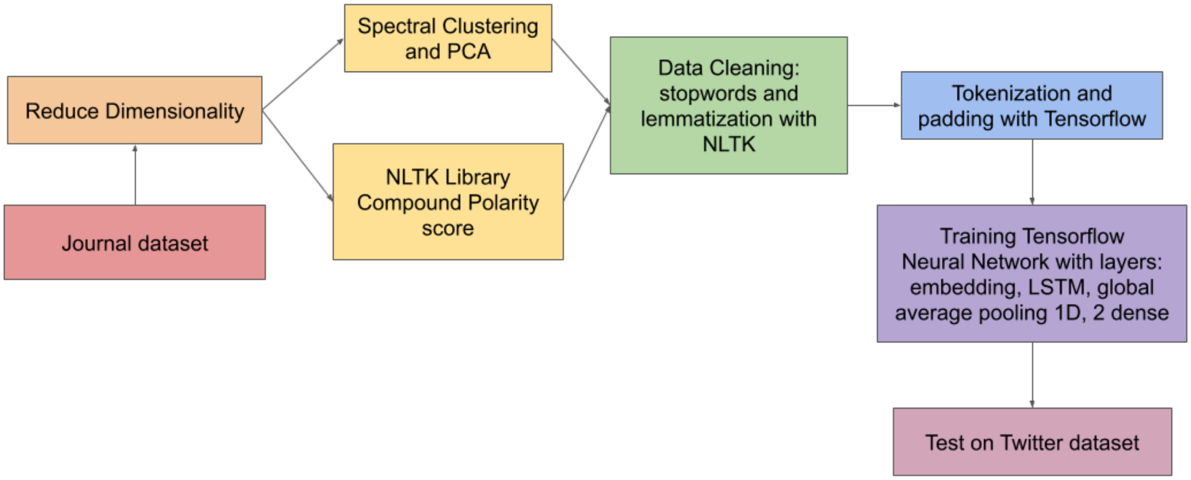
Reduce Dimensionality
Journal dataset
Spectral Clustering and PCA
NLTK Library Compound Polarity score
Data Cleaning: stopwords and lemmatization with NLTK
Tokenization and padding with Tensorflow
Training Tensorflow Neural Network with layers: embedding, LSTM, global average pooling 1D, 2 dense
Test on Twitter dataset

Figure 1: Flowchart Abstract

1. Introduction

Depression is a pervasive mood disorder that significantly impacts how individuals feel, think, and manage their daily activities, including sleeping, eating, and working, as noted by the National Institute of Mental Health (National Institute Of Mental Health, 2023). Affecting over 18 million adults in the United States each year, depression stands as a leading cause of disability among individuals aged 15-44 (*Facts about Depression | Hope for Depression*, 2013). It is also the primary factor behind the tragic loss of over 41,000 lives to suicide annually (*Facts about Depression | Hope for Depression*, 2013). Given the widespread and severe impact of depression, effective and accessible strategies for mental health management are critical. One such strategy is journaling, which has been shown to be a low-cost, low-risk therapeutic intervention that can aid in mental health management by providing individuals with a private outlet to process their emotions and thoughts (Sohal, M., Singh, P., Dhillon, B. S., & Gill, H. S., 2022).

However, in the digital age, the rise of social media has led many to seek alternative outlets for emotional expression. Platforms such as blogs and Twitter have become modern substitutes for traditional journaling, allowing individuals to vent and share their experiences in a public or semi-public forum. While this shift offers new opportunities for self-expression, it also presents challenges and raises questions about the role of social media in mental health. Unlike private journals, which are inherently introspective and personal, social media posts are often crafted with an audience in mind, potentially altering the nature of the emotional content shared.

In the context of mental health, early detection is crucial for effective intervention. Previous research has focused on identifying patterns in suicide notes to aid in the detection of individuals at risk (*Early Identification of Mental Health Issues in Young People*, n.d.). NLP models have shown promise in the ability to detect nuanced human language, including sarcasm and irony (Potamias et al., 2020). Other studies have revealed that certain emotional patterns are often present and detectable by NLP models in the writings of individuals suffering from severe mental health issues, such as those contemplating suicide (Desmet & Hoste, 2013). However, the challenge with suicide notes is that they are typically discovered too late when opportunities for intervention have passed.

This research aims to explore whether an NLP model can detect similar emotional patterns in less dire contexts, specifically in journal entries, and then apply this knowledge to differentiate between the tweets of depressed and non-depressed Twitter users. By leveraging a dataset of journal entries labeled with emotional content, the model was trained to recognize and classify emotions associated with depression. Following this, the model was tested on a secondary dataset of Twitter posts, where the goal was to identify whether it could accurately distinguish between the posts of users identified as depressed and those who were not.

The primary dataset of journal entries, classified into 18 different emotions, is used to train the model. The secondary dataset of Twitter posts then served as a testing ground, where the model's ability to generalize its understanding of emotional patterns from journals to the more public and varied expressions found on social media was evaluated.

By exploring the potential of NLP models to detect early signs of depression through the analysis of social media content, this study aims to offer new tools for mental health monitoring and early intervention, leveraging the digital footprints left by individuals in their daily online interactions. This approach not only enhances our understanding of how depression manifests in written communication but also underscores the importance of adapting mental health strategies to the realities of the digital age.

2. Materials and Methods

The objective of this research was to develop and train an NLP model using journal data and extrapolate it to classify tweets from Twitter users identified as either depressed or not depressed. To achieve this, a combination of supervised learning, unsupervised learning, and sentiment analysis tools from the NLTK library were employed.

2.1 Dataset Description

2.1.1 Primary Dataset: Journal Entries

The primary dataset consisted of 1,500 journal entries, with the largest entry being 98 words, collected by X. Alice Li and Devi Parikh (X. Alice Li & Parikh, 2020). These entries were responses to the prompt, “What were salient aspects of your day yesterday? How did you feel about them?” Each entry was annotated based on the presence or absence of 18 distinct emotions: fear, anger, anxiety, shame, awkwardness, boredom, calmness, confusion, disgust, excitement, frustration, happiness, jealousy, nostalgia, pride, sadness, satisfaction, and surprise (X. Alice Li & Parikh, 2020). However, the data was highly imbalanced, with a skew toward positive emotions, and the overall size of the dataset was relatively small, posing challenges for training the NLP model. Figure 1 shows the frequency of occurrence of each emotion in the dataset.

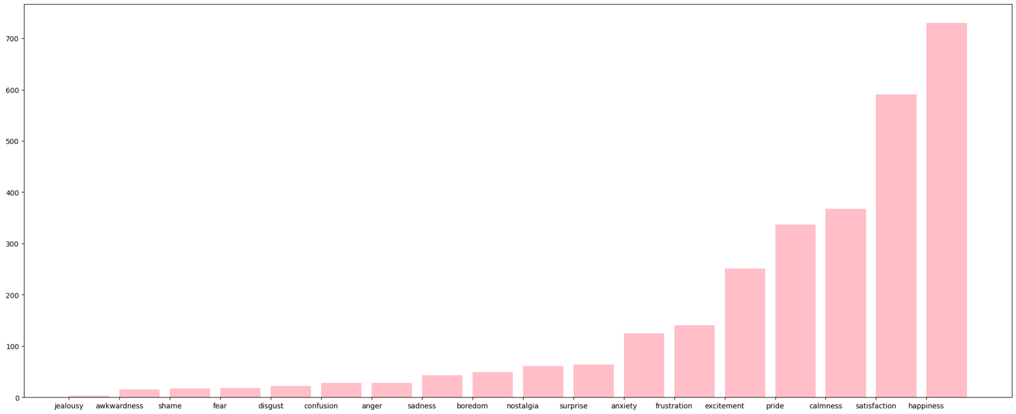

Figure 2: Frequency Of Occurrence of Each Emotion

2.1.2 Secondary Dataset: Twitter Posts

The secondary dataset consisted of 8,500 tweets, 3,500 of which were authored by users identified as depressed and 5,000 by users identified as not depressed (Hyun Ki Cho, 2021). Each tweet was labeled with usernames, allowing user-level analysis and classification.

2.2 Data Preprocessing

2.2.1 Data Cleaning

The text data underwent a series of preprocessing steps using the NLTK library. This included tokenizing the text using word_tokenize, filtering out common stopwords, and lemmatizing the words using the WordNetLemmatizer function. These cleaned and processed texts were then reassembled into complete sentences, ready for further analysis.

2.2.2 Data Splitting and Padding

The cleaned data was split into training and testing sets using an 80/20 train-test split facilitated by the Scikit-learn (Sklearn) library. Following this, the TensorFlow Tokenizer was used to convert the text into sequences of integers, which were then padded to match the length of the longest journal entry using TensorFlow's pad_sequences function. This ensured that all input data had a uniform length, a necessary condition for feeding data into the model.

2.3 Model Development

2.3.1 Approach 1: Unsupervised Learning with Spectral Clustering and PCA

Two different approaches for model development were used to develop models capable of classifying the emotional content of the data.

The first approach utilized unsupervised learning techniques, specifically spectral clustering and principal component analysis (PCA), to cluster the journal entries into two categories based on the patterns of emotions they exhibited.

- **Spectral Clustering:** This technique groups data points into clusters based on the properties of the data affinity matrix, identifying inherent patterns within the dataset (GeeksforGeeks, 2023).
- **Principal Component Analysis (PCA):** PCA was used as a dimensionality reduction method to condense the 18-dimensional emotion data into two principal components or clusters, effectively simplifying the data while retaining most of its informative content (*Principal Component Analysis (PCA) Explained | Built In*, n.d.).

The application of PCA and spectral clustering revealed that the data naturally grouped into two clusters: Cluster 0, which primarily contained positive emotions, and Cluster 1, which predominantly contained negative emotions.

2.3.2 Approach 2: Sentiment Analysis Using NLTK

The second approach involved using the NLTK library's sentiment analysis tool to classify each journal entry as either "negative" or "not negative" based on the compound polarity score. The compound polarity score is a measure that rates text on a scale from -1 (most negative) to 1 (most positive) (Mogyorosi, n.d.). Entries with scores below 0 were classified as negative, while those with scores 0 or above were classified as not negative.

Interestingly, both approaches—unsupervised learning and sentiment analysis—produced consistent results, with Cluster 0 corresponding to the "not negative" classification and Cluster 1 to the "negative" classification. This alignment between the two methodologies provided a strong basis for further analysis.

2.3.3 Data Augmentation Using Monte Carlo Resampling

Due to the small and skewed nature of the dataset, Monte Carlo resampling was employed using the Sklearn library to balance and expand it. This technique involved resampling each cluster to create 7,000 entries per cluster, which helped mitigate the effects of the data imbalance.

2.4 Model Training

The final model was built using TensorFlow and Keras, following a sequential architecture:

- **Embedding Layer:** This layer had an input dimension equal to the total number of unique words in the dataset, an output dimension of 128, and an input length matching the length of the padded sequences.
- **LSTM Layer:** A Long-Short-Term Memory (LSTM) layer with 256 units was used, incorporating a dropout rate of 0.5 and a recurrent dropout rate of 0.5. return_sequences was set to True to retain the sequence information across time steps.
- **GlobalAveragePooling1D Layer:** This layer condensed the sequence of features into a single vector, making it easier to handle the data in subsequent layers.
- **Dense Layers:** The model included a dense layer with 32 units and ReLU activation to capture non-linear relationships, followed by a dense layer with 1 unit and sigmoid activation for binary classification.

The model was compiled using binary crossentropy as the loss function, the Adam optimizer, and accuracy as the primary metric. It was then trained over 10 epochs, with verbose set to 2.

2.5 Application

The trained model was applied to classify the tweets of depressed and non-depressed Twitter users into the previously identified clusters. The final classification of tweets was conducted on a user level, allowing for the identification of broader patterns in the emotional content of the tweets.

3. Results

3.1 Hypothesis

If an NLP model is trained on short journal entry data classified by emotions present, it will detect similar patterns within other short-form text, like tweets. It should perform well on a variety of tweets, with more negative-leaning tweets predicted for the tweets of depressed users.

3.2 Approach 1: Spectral Clustering

In approach 1, spectral clustering was used to cluster the training data into 2 clusters. Figure 2 illustrates the results of this clustering process, where Cluster 0 is represented by the red dots, and Cluster 1 is represented by the gray dots. The clear separation between the clusters indicates that the algorithm was successful in distinguishing between different types of emotional content within the journal entries.

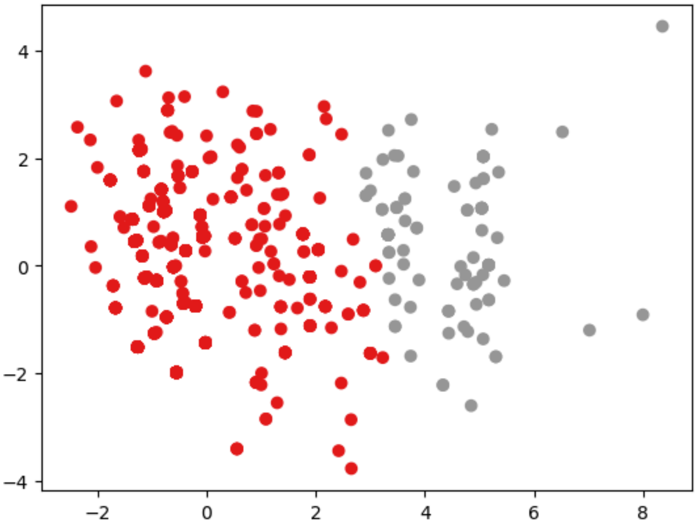

Figure 3: Spectral Clustering of Train Data

3.3 Approach 2: Sentiment Analysis Using NLTK

The second approach utilized the NLTK library to classify the training data into two clusters based on sentiment polarity. Each journal entry was analyzed, and its compound polarity score was calculated. Entries with scores below 0 were categorized into Cluster 1, which represented the negative cluster, while the remaining entries were placed into Cluster 0. Figure 3 shows the distribution of entries across these clusters, presented as a histogram. This approach allowed for a straightforward classification of the entries based on their overall sentiment.

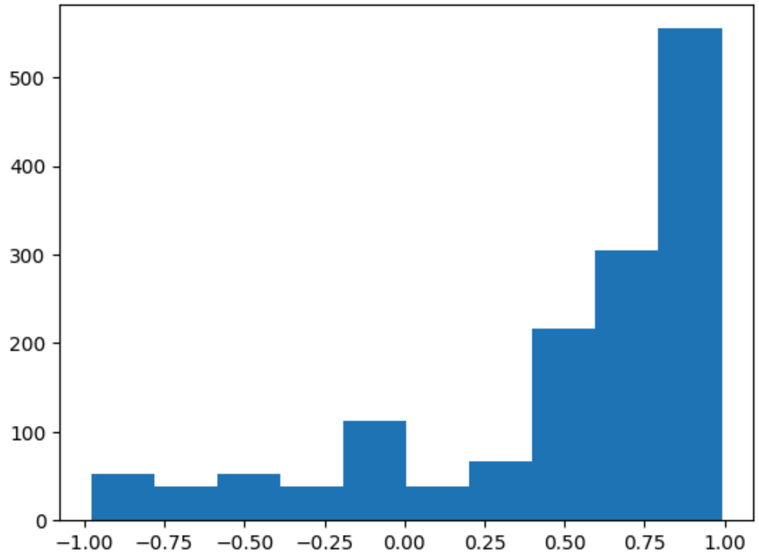

Figure 4: NLTK Analysis of Train Data

3.4 Emotion Distribution Analysis

Further analysis was conducted to compare the distribution of emotions across the clusters formed in both approaches. Figure 4 presents the emotion frequency for the clusters in Approach 1. It is evident that Cluster 1, formed through unsupervised learning, contained a higher concentration of negative emotions, while Cluster 0 predominantly included positive emotions. This pattern is consistent with the sentiment-based clustering observed in Approach 2, demonstrating that both approaches, despite their methodological differences, identified similar emotional distinctions within the data.

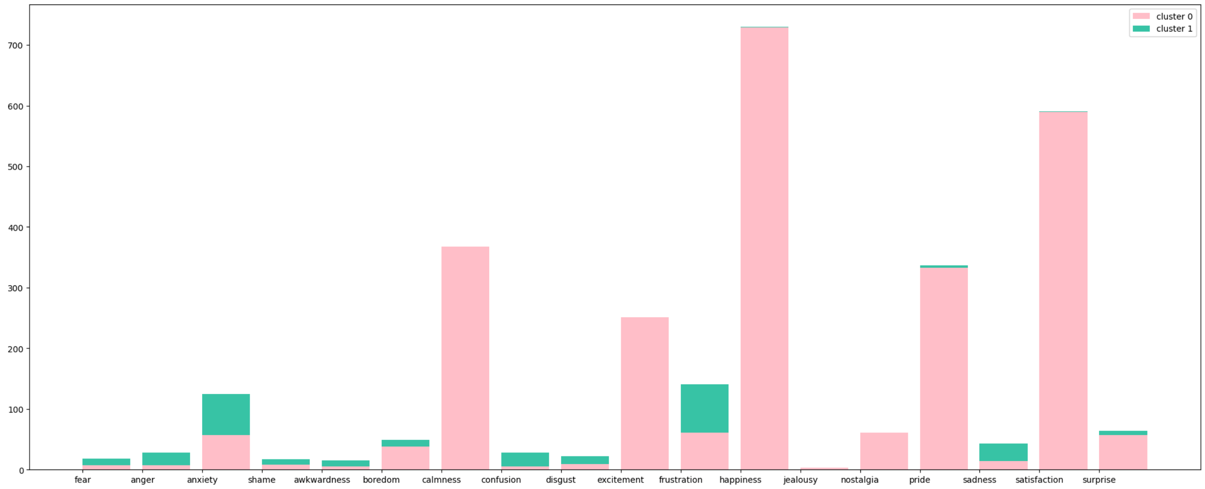

Figure 5: Approach 1 Emotion Frequency

In Approach 2, the frequency of emotions was analyzed similarly, as shown in Figure 5.

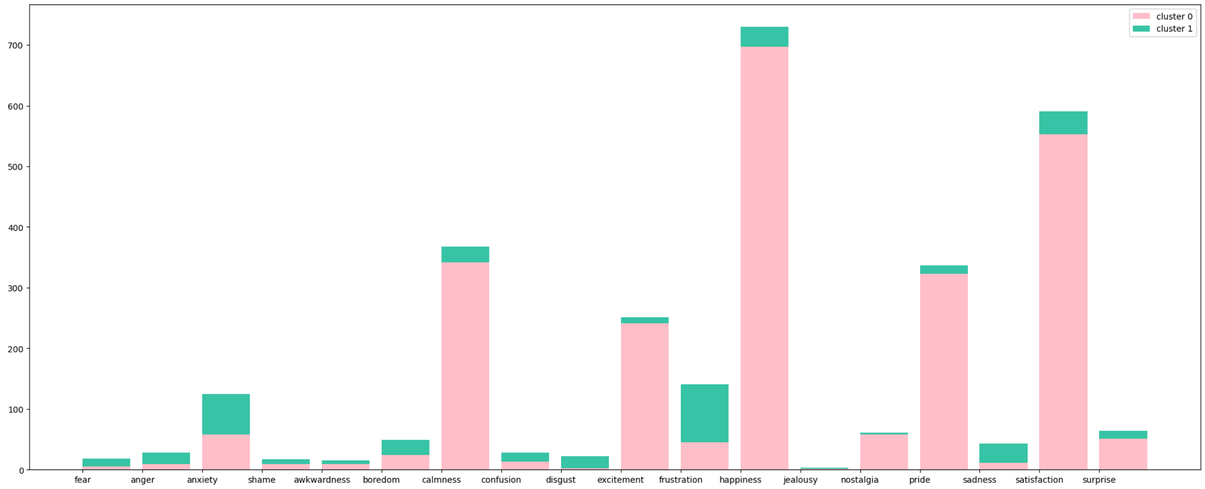

Figure 6: Approach 2 Emotion Frequency

3.5 Data Resampling and Emotion Frequency Adjustment

Due to the small and imbalanced nature of the original dataset, Monte Carlo resampling was applied to balance the dataset and address the skewness in emotion distribution. The resampling process involved expanding the dataset by equalizing the number of entries in each cluster. Figure 6 shows the emotion frequency distribution after resampling in Approach 1. The resampled data demonstrates a more balanced distribution of emotions across the clusters, which helped improve the model's performance during training.

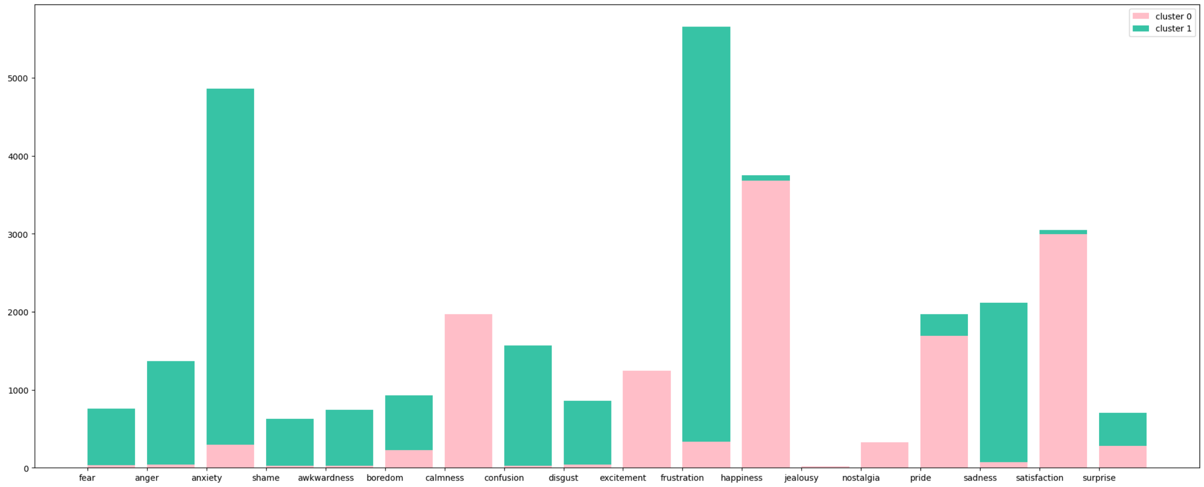

Figure 7: Approach 1 Emotion Frequency After Resampling

Similarly, the emotion frequency after resampling in Approach 2 is depicted in Figure 7. The balanced dataset provided a more equitable representation of positive and negative emotions.

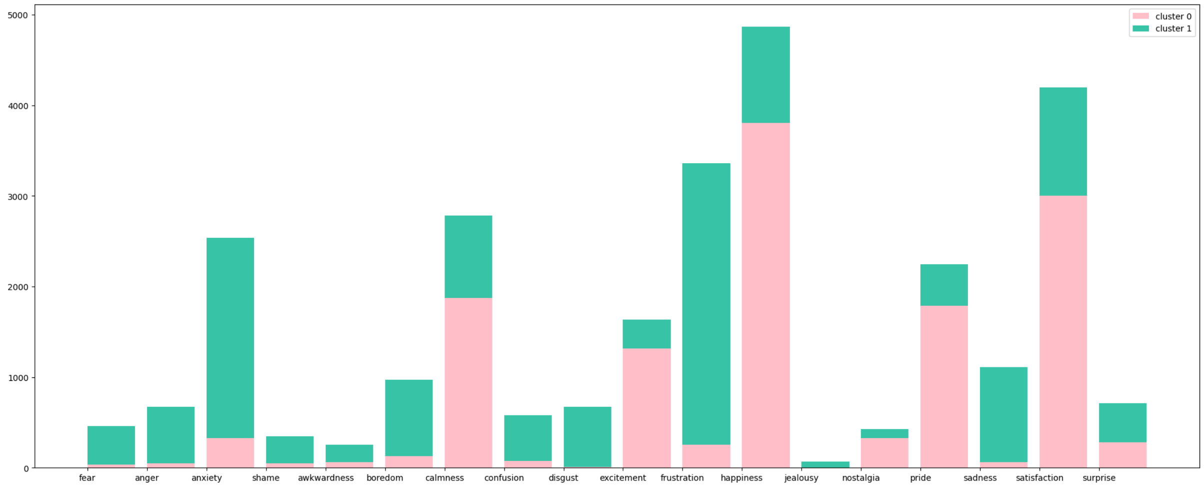

Figure 8: Approach 2 Emotion Frequency After Resampling

3.6 Model Training and Performance

The neural network was then trained on the processed and resampled journal data. In Approach 1, the model achieved a validation accuracy of 99.32% and a validation loss of 3.03% after ten epochs. The training process is illustrated in Figure 8, which shows the progression of validation accuracy and loss over the epochs. This high accuracy indicates that the model was highly effective in learning from the clustered data.

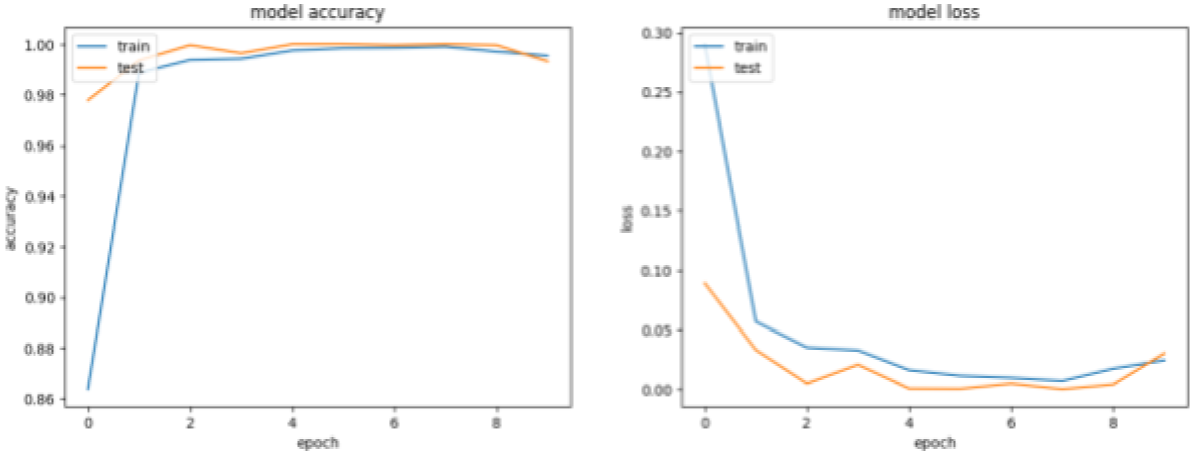

Figure 9: Approach 1 Model Training

In Approach 2, the neural network achieved an even higher validation accuracy of 99.82% and a significantly lower validation loss of 0.86%, as shown in Figure 9.

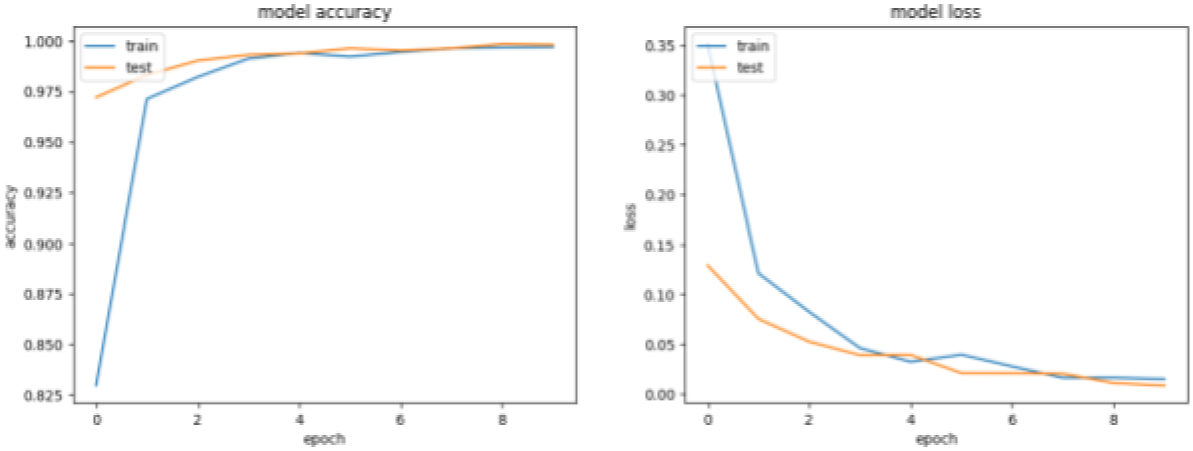

Figure 10: Approach 2 Model Training

3.7 Tweet Classification and User-Level Analysis

After training, the models were applied to classify the tweets from depressed and non-depressed Twitter users. In Approach 1, 25.2% of the tweets from depressed users were classified into Cluster 1, while 16.2% of the tweets from non-depressed users were also classified into this cluster. Figure 10 illustrates these classification results, indicating a clear distinction in the emotional content of tweets from the two user groups.

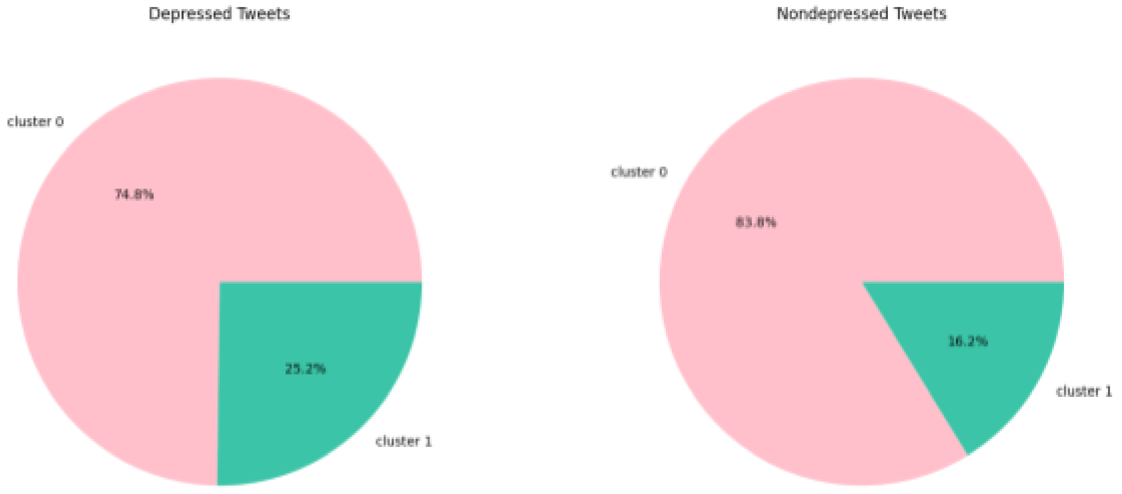

Figure 11: Approach 1 Tweet Classification

In Approach 2, a higher percentage of tweets—36.8% from depressed users and 17.6% from non-depressed users—were classified into Cluster 1, as shown in Figure 11. This approach demonstrated a more pronounced separation between the tweets of depressed and non-depressed users, suggesting that sentiment analysis provided a more sensitive measure of emotional content.

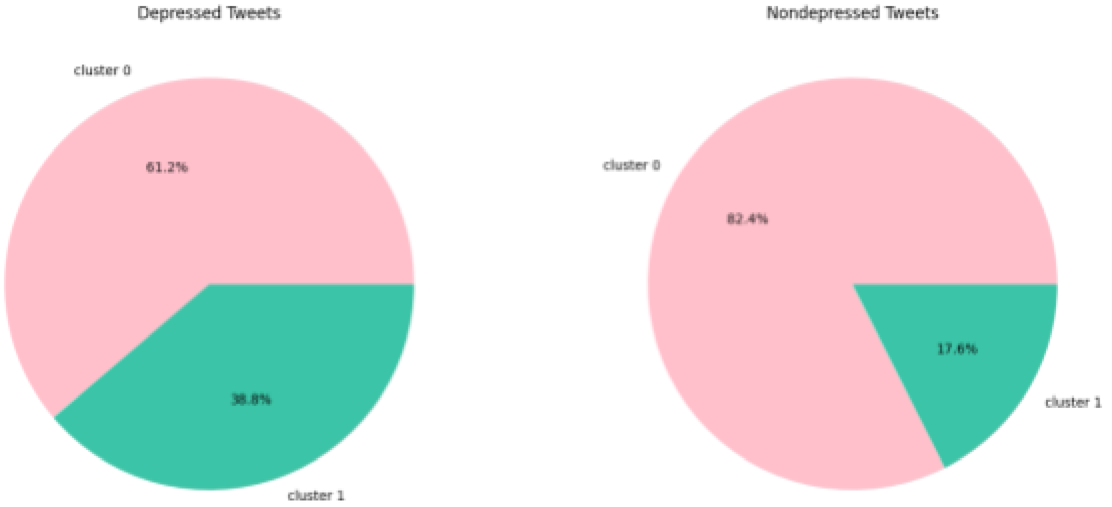

Figure 12: Approach 2 Tweet Classification

To further explore these findings, the classification results were analyzed at the user level. Figures 12 and 13 present the user-level classification for Approaches 1 and 2, respectively. The results reinforce the conclusion that there is a significant disparity in the emotional content of tweets between depressed and non-depressed users. This user-level analysis underscores the potential of NLP models to identify and differentiate between the online behaviors of individuals with varying mental health statuses.

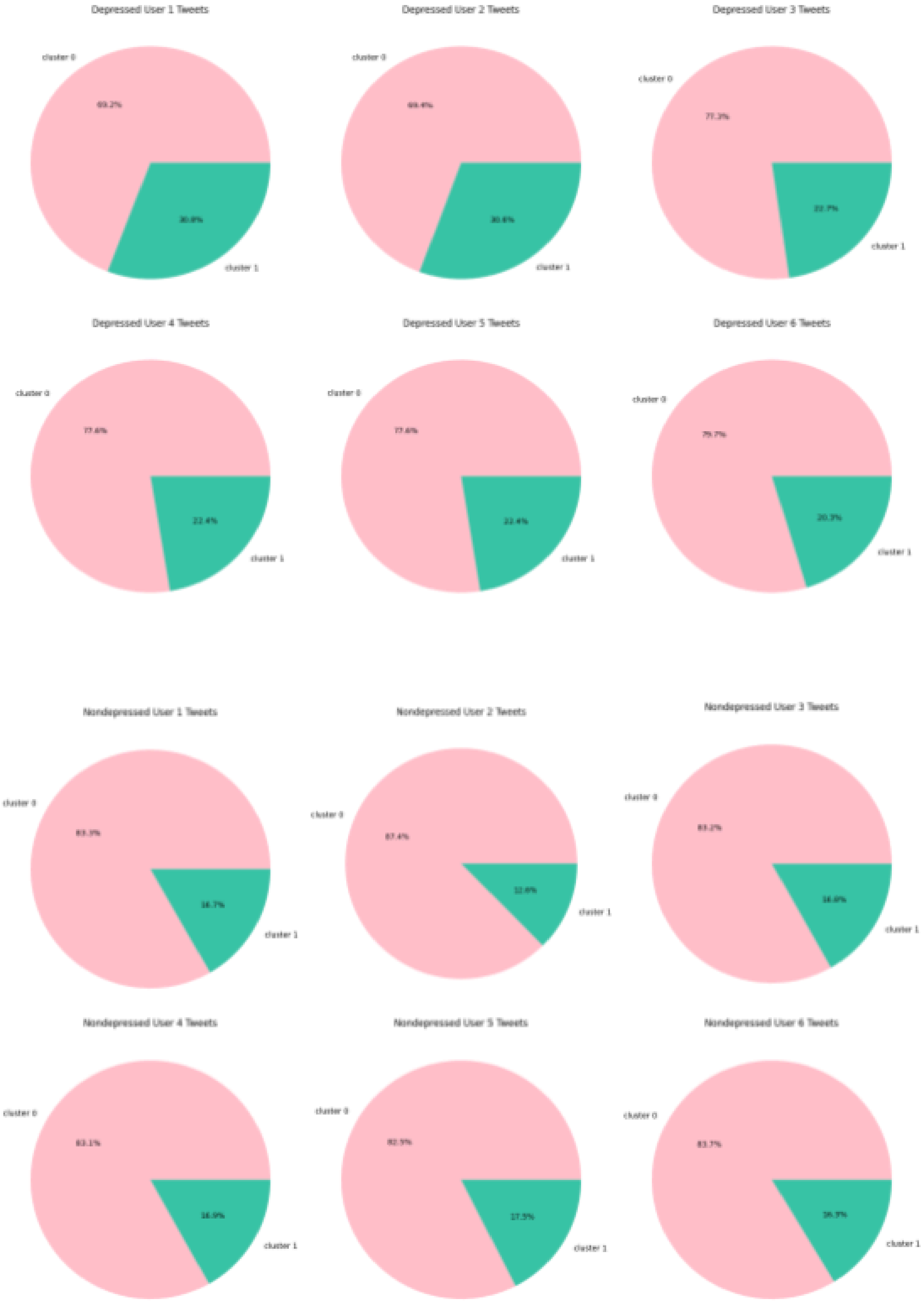

Figure 13: Approach 1 Tweet Classification by User

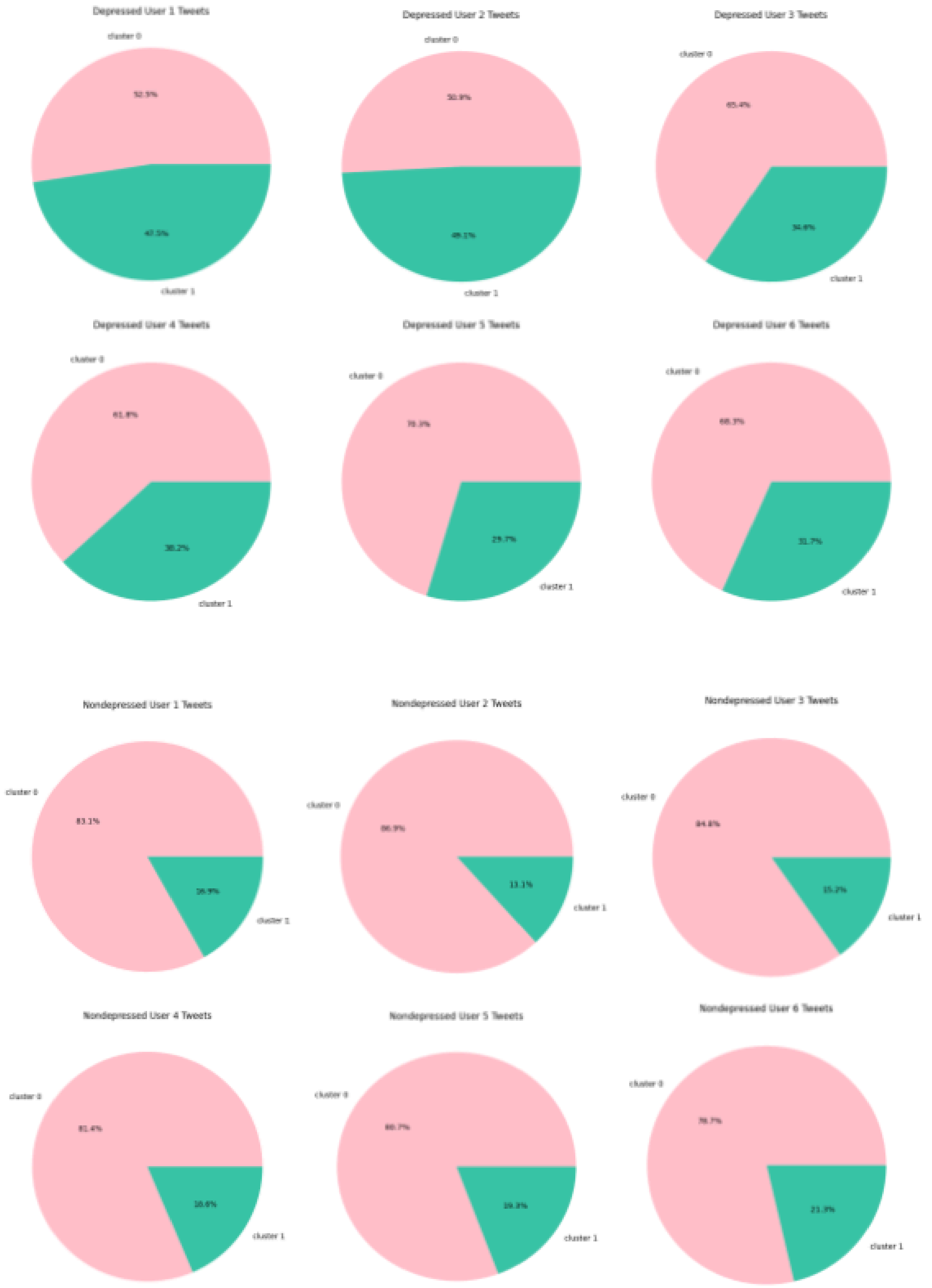

Figure 14: Approach 2 Tweet Classification by User

4. Discussion

The objective of this research was to find whether an NLP model can find patterns in journal entries and use them to detect differences between the tweets of depressed and non-depressed Twitter users. Two different approaches were used to build models to test this, one with 99.32% validation accuracy and the other with 99.82% validation accuracy. Though there were differences in the results from both models, both demonstrated that tweets of depressed users were classified more negatively than the tweets of users who were not depressed.

These findings suggest that social media can serve a function similar to traditional journaling, offering individuals a platform to express their thoughts, emotions, and experiences in a way that reflects their internal state. This parallel indicates that, just as personal journals have long been used as tools for self-reflection and emotional release, social media posts may similarly capture a person's psychological well-being, albeit in a more public and socially interactive context.

Moreover, the content shared on social media platforms often provides insights into an individual's mood, cognitive patterns, and social interactions, making it a valuable resource for mental health assessment. By analyzing the language, tone, and frequency of posts, as well as the nature of interactions with others, mental health professionals can potentially detect early signs of distress, such as depression, anxiety, or social withdrawal. This approach could complement traditional assessment methods, offering a real-time and continuous stream of data that reflects the user's daily life.

Furthermore, social media's widespread use and accessibility make it an especially powerful tool for mental health monitoring. Unlike clinical settings, where assessments are typically conducted in a controlled environment and at specific intervals, social media provides an ongoing record of a person's mental state, potentially leading to more timely and contextually relevant interventions. However, it's important to note that this method also raises ethical considerations regarding privacy and the potential for misinterpretation of posts. Therefore, any application of social media analysis for mental health purposes would need to be conducted with caution, ensuring that it respects users' privacy and is used along with other diagnostic tools to provide a holistic view of an individual's mental health.

This study was constrained by the availability and quality of the training data, which significantly impacted the model's performance and generalizability. The journal dataset used to train the model was not only limited in size but also heavily skewed. This imbalance likely introduced biases into the model, affecting its ability to accurately identify and assess mental health indicators in various social media users.

The small size of the dataset also restricted the model's capacity to learn complex patterns and nuances, which are crucial for making reliable predictions in the context of mental health monitoring. To address these limitations, future research should focus on expanding and diversifying the dataset. This could involve collecting more extensive and varied data.

Overall, despite these limitations, this study introduces a novel approach to mental health monitoring. This research represents a crucial step toward enabling social media companies to play an active role in addressing the mental health challenges that have, in part, emerged because of them. By taking responsibility for monitoring and supporting the mental well-being of their users, social media companies can begin to mitigate some of the negative effects associated with their platforms, contributing to a healthier digital environment. This approach not only highlights the potential for technological solutions in mental health care but also underscores the ethical responsibility of tech companies to safeguard the well-being of their users.

References

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*, *40*, 6351–6358. https://doi.org/10.1016/j.eswa.2013.05.050

Early Identification of Mental Health Issues in Young People. (n.d.). Mental Health America. https://mhanational.org/issues/early-identification-mental-health-issues-young-people

Facts about Depression | Hope for Depression. (2013). Hope for Depression. https://www.hopefordepression.org/depression-facts/

GeeksforGeeks. (2023, December 14). *Spectral Clustering A Comprehensive Guide for Beginners*. GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/spectral-clustering-a-comprehensive-guide-for-beginners/#

Hyun Ki Cho. (2021). *Twitter Depression Dataset*. Kaggle.com. https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset?resource=download

Mogyorosi, M. (n.d.). *Sentiment Analysis: First Steps With Python's NLTK Library – Real Python*. Realpython.com. https://realpython.com/python-nltk-sentiment-analysis/

National Institute Of Mental Health. (2023, March). *Depression*. National Institute of Mental Health. https://www.nimh.nih.gov/health/topics/depression

Potamias, R. A., Siolas, G., & Stafylopatis, A. - G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

Principal Component Analysis (PCA) Explained | Built In. (n.d.). Builtin.com. https://builtin.com/data-science/step-step-explanation-principal-component-analysis#:~:text=necessary%20for%20context.-

Sohal, M., Singh, P., Dhillon, B. S., & Gill, H. S. (2022). Efficacy of journaling in the management of mental illness: a systematic review and meta-analysis. *Family medicine and community health*, *10*(1), e001154. https://doi.org/10.1136/fmch-2021-001154

X. Alice Li, & Parikh, D. (2020). *Lemotif: An affective visual journal using deep neural networks*. https://arxiv.org/abs/1903.07766

Supplemental Materials

Code:
https://github.com/tvisha763/Detecting-Depression-in-Social-Media-with-NLP-Models-Trained-on-Journal-Entry-Data