Two papers were reviewed. The first paper lacks associated code and will serve as a reference for comparing the performance of different classification models in my project. The second paper includes a GitHub repository with codes but does not provide the data. Using similar data, I was able to replicate some of the results, although exact replication was not feasible due to differences in the datasets. Nonetheless, this paper offered valuable insights into potential analyses that can be performed with text data.

1. Supervised machine learning models for depression sentiment analysis

https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1230649/full

This paper analyzes Twitter data consisting of approximately 1.6 million entries, categorized as positive, negative, or neutral, to identify signs of depression from a user's tweets. The authors first preprocessed the data using NLTK, implementing steps such as tokenization, stop word removal, stemming, lemmatization, and creating bigrams/trigrams. Special characters and links, including URLs, were removed to improve tweet quality. Numbers were excluded since they do not contribute to sentiment analysis, and all text was converted to lowercase for consistency.

The study utilized four machine learning classifiers: XGB Classifier, Random Forest Classifier, Logistic Regression, and Support Vector Machine. Their performances were evaluated using metrics like accuracy, precision, recall, and F1 score, along with comparisons of their processing speeds.

The results showed that the Logistic Regression model performed best, achieving an accuracy of 96.3% with a fast computation time of just 0.29 seconds.

The paper does not provide the code or dataset used in the analysis, making it impossible to replicate the results. However, in my project, I plan to preprocess the data and test various machine learning and deep learning models and compare their performance in a similar manner.

2. Context is Important in Depressive Language: A Study of the Interaction Between the Sentiments and Linguistic Markers in Reddit Discussions

https://paperswithcode.com/paper/context-is-important-in-depressive-language-a
Code:
https://github.com/nehasharma666/depression

This research paper examines how the topic of discussion influences language patterns and emotional expression in people with depression, using data from Reddit to explore these interactions. The authors first aimed to determine whether differences exist between the depression and control groups regarding the topics discussed and the sentiments expressed within the dataset. They then sought to understand which psycholinguistic factors, extracted using the LIWC tool, contribute to the sentiment differences observed between individuals with depression and the control group across various topics.

2.1 Methods

An existing Reddit-based data set comprising posts from users with self-reported mental health diagnoses (SMHD) is used. The analysis is performed only in the part where users have depression or normal.

Sentiment Analysis with RoBERTa
Sentiment analysis was performed using a pre-trained RoBERTa model to classify each text as negative, neutral, or positive based on the sentiment. The results, including sentiment scores and labels, are stored in a new DataFrame and saved to a CSV file for further analysis, such as topic modeling.

Topic Modeling with BERTopic
The BERTopic model is used to extract interpretable topics from a large collection of textual data. The process begins with loading and preprocessing the data, followed by converting the text into sentence embeddings using a pre-trained SentenceTransformer model. UMAP is then applied for dimensionality reduction, and HDBSCAN is employed to cluster the data. BERTopic generates topics while incorporating additional representation models like KeyBERT for enhanced topic extraction. The results, including topic information and sentiment labels, are saved into CSV files for further analysis, with the option to reload the trained model later.

2.2 Results:

Data Availability:

The dataset discussed in the paper is the SMHD dataset, which comprises self-reported mental health diagnoses from Reddit. It includes over 20,000 clinical users and approximately 300,000 control users, with diagnoses covering nine mental illnesses, including depression. The authors focused on a subset of this dataset, specifically selecting 1,316 clinical and 1,316 control users, resulting in a total of approximately 550.6k posts. However, the original data is not publicly available, making it impossible to replicate the exact results.

For our analysis, we used an alternative Reddit depression dataset from Kaggle, which consists of 7,650 unique entries without user identification numbers. This limitation prevented user-specific analyses. Given the dataset differences, we aimed to replicate the analysis using the available code, focusing on identifying qualitatively similar patterns rather than exact results.

https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

We aimed to replicate the sentiment analysis and topic modeling sections from the research paper. While we did not reproduce all the results, we will summarize the remaining findings in the research paper below to gain insights into additional types of analyses that can be conducted.

Replicated parts:

We began by preprocessing the data through standard cleaning and lemmatization techniques. Next, we applied the Roberta model to generate sentiment scores for each post. We then used the Bertopic model to extract topics from the data. Finally, we conducted sentiment and topic-related analyses.

Sentiment Analysis

In the original paper, the depression group showed 6% more negative sentiments and 4% more positive sentiments. In our replicated results, we found that the depression group exhibited approximately 28% more negative sentiments and 8% fewer positive sentiments. A Chi-square test confirmed that the variation in sentiment distribution was statistically significant. Additionally, we observed that posts from the depression group were significantly longer than those from the control group.

(Can be seen in the jupyter file named Sentiment Distribution and Topic Analysis)

Topic Modeling

The original paper's topic modeling identified 4,187 topics, analyzing their frequency and sentiment distribution across user groups. In our replicated results, we generated 85 topics. Given that our dataset is significantly smaller, we did not reduce the number of topics. Based on the representative words, we identified potential topic names without further refining them. If further analysis were conducted, some similar topics could potentially be merged. The 17 most frequent topics are as follows:

```
"life","anxiety","sleep","depression","work","twitter","music","mental
health","health and emotional well-being","TV shows","school","food","body
parts","job","public transport","french","weather"
```

Frequency distributions for these topics were analyzed. The results indicate that the control group frequently discusses topics related to daily life, school, work, music, Twitter, and public transportation. In contrast, the depression group's discussions are more focused on health-related issues such as anxiety and depression.

Topic and Sentiment Interactions:

We also analyzed sentiment distribution across topics. Our observations reveal that the depression group expresses significantly higher negative sentiment towards topics related to depression or health related topics compared to the control group. Conversely, for topics such as work, school, TV shows, and Twitter, the depression group shows less negative or even more positive sentiment than the control group. These findings are consistent with the patterns observed in the research paper.

Here are the remaining models and results from the paper:

These sections were not recreated due to the lack of exact data. Nonetheless, they offer valuable insights for the main project.

LIWC (Linguistic Inquiry and Word Count) Analysis

LIWC (Linguistic Inquiry and Word Count) is a software tool developed by Pennebaker et al. that analyzes written text to reveal psychological and emotional information. Words are mapped to psycho-linguistic attributes, resulting in the proportion of words in various categories such as Summary Variables, Linguistic Dimensions, Psychological Processes.

There are four summary variables: Analytic Thinking, Clout, Authenticity, and Emotional Tone. Other categories are:

- **Linguistic Dimensions**: Percentage of specific types of words (e.g., pronouns, verbs).
- **Psychological Processes**: Attributes related to cognitive, emotional, and social processes.
- **Extended Dictionary**: Includes broader aspects like cultural and lifestyle factors.

User-based LIWC Analysis

In the dataset each user has many posts. Posts for each user are aggregated first. The feature values for users are calculated. The difference in LIWC attributes were analyzed. Welch'st-test with adjusted p-value using Bonferroni correction was performed.

Topic-Specific LIWC Analysis

A linear regression model was built using LIWC attributes as predictors and the sentiment differences between depression and control groups as the response variable.

- **Dependent Variable:** The sentiment score difference is the dependent variable, measuring the difference between positive and negative sentiments in both groups. For each topic, it captures whether the depression group or control group shows more positive or negative sentiment.If the difference is positive, it indicates that the depression group expresses more positive sentiment on that topic. Conversely, if the difference is negative, it indicates that the control group expresses more positive sentiment on that topic.
- **Independent Variable:** First, mean aggregated LIWC scores are calculated for both depression and control groups for each topic. The difference between these scores is then computed. A positive difference indicates a higher proportion of the attribute in the depression group, while a negative difference indicates a higher proportion in the control group.

These independent features, calculated across 4187 topics and 63 attributes, are used to fit a linear regression model using ordinary least squares. Statistically significant coefficients for the LIWC attributes suggest that these attributes have an impact on the sentiment differences between the groups.

Topic-Specific LIWC Analysis results

A linear regression model with 63 LIWC attributes as predictors and sentiment differences as response variable was fitted. The model explained 26.6% of the variation in sentiment difference, $R^2 = 0.266$, $p < 0.001$, and found 25 attributes to be statistically significant at ($p = 0.05$).

Since positive attribute values refer to the depression group expressing more of that attribute in a topic, higher positive coefficient shows that these features are linked to a higher level of positive sentiment in the depression group's posts across topics.

The findings show that the depression group generally has more negative sentiments than the control group. Key variables like post length and emotional tone align with expected sentiment patterns. Surprisingly, words related to anger and sadness correlate with positive sentiment in some cases. Further analysis suggests that mixed emotions, where both positive and negative words are used, may contribute to these unexpected results, particularly in the depression group. More details can be found in the paper.

2.3 Discussion

The strengths and importance of the paper lie in its methodological approach and insights into the role of discussion context in understanding language used in depression. The methods employed simplify the analysis, making it more manageable and effective. Furthermore, the paper underscores the significance of discussion context, demonstrating that the emotional intensity and meaning of linguistic markers can vary depending on the topic being discussed. This highlights the nuanced ways in which language reflects emotional states, providing a deeper understanding of how depression is expressed across different contexts.

Limitations in the research paper:

There are several limitations in the study. First, the sentiment analysis model, trained on Twitter data, may not perform optimally on Reddit posts, which are generally longer. Manual review of a sample showed the model often misclassified positive and negative posts as neutral, likely leading to underestimation of those sentiments. Additionally, the SMHD dataset used spans 2006-2017, making it uncertain whether all users labeled as depressed had consistent depression during that period. Lastly, the control group, drawn from non-mental health subreddits, may still include users with undiagnosed mental health issues. Additionally, the LIWC tool struggles with contextual nuances, leading to potential misclassifications, and the regression analysis only explains 26% of sentiment variance, indicating other influencing factors. The dataset's focus on depression also limits the generalizability of the findings.

Limitations of the regenerated work:

Due to the smaller dataset size, extracting a significant number of meaningful topics was challenging, and the topics identified had low frequencies.