CIS 600: Applied Natural Language Processing

Distinguishing Mental Health Categories on Reddit Using NLP

Spoorthi Jayaprakash Malgund (SUID: 904543696)
Aditee Suryakant Malviya (SUID: 465609468)
Adithya Neelakantan (SUID: 663182945)
Krupa Minesh Shah (SUID: 619191197)
Raj Nandini (SUID: 949041747)
Harshitha Reddy Kancherla (SUID: 519904638)
Mrunmayee Atul Jakate (SUID: 652397247)

Table of Contents

1. Introduction

The emergence of online platforms such as Reddit has transformed the way individuals convey and talk about their mental health struggles. The anonymity and transparency of the platform motivate users to discuss their experiences with conditions like depression, anxiety, and suicidal thoughts, rendering Reddit a highly useful resource for grasping the intricacies of mental health. Nonetheless, this wealth of information poses a major challenge: how can we proficiently analyze and organize these discussions to promote early intervention, establish supportive settings, and advance impactful mental health research? Natural Language Processing (NLP) presents a potential solution to this issue by facilitating the methodical examination of text data to reveal patterns, trends, and emotional nuances within conversations. Utilizing subreddits such as r/depression, r/Anxiety, and r/SuicideWatch, our project seeks to create a sophisticated model that not only classifies mental health posts but also detects high-risk signs like suicidal thoughts. In contrast to conventional methods that depend only on subreddit labels, this project explores linguistic patterns and contextual hints more thoroughly, enabling a clearer comprehension of particular mental health disorders.

Utilizing sophisticated NLP methods like Long Short-Term Memory (LSTM) networks, the initiative aims to capture the sequential character of text, which is vital for examining the changing dynamics of mental health conversations. Posts are categorized into groups such as depression, anxiety, or overall well-being, while also evaluating degrees of severity. This combined emphasis on classification and risk evaluation provides researchers, clinicians, and policymakers with practical guidance to more effectively assist individuals in need.

Subreddits that concentrate on mental health, like r/depression and r/Anxiety, provide important perspectives on the individual experiences and challenges faced by users coping with these matters. In comparison, general subreddits such as r/CasualConversation serve as a foundation for grasping routine discussions that are not connected to mental health. Through the integration of data from both varieties of subreddits, we established a balanced dataset that reflects a broad range of discussions. This method guarantees that the created models are not only proficient in detecting mental health classifications but also adaptable for use in actual situations.

The importance of this effort goes further than just technological advancement. By effectively identifying mental health categories, this initiative aids in establishing safer online environments, directing mental health strategies, and influencing policy suggestions. It highlights how technology can be utilized not only to examine language but to promote empathy, comprehension, and prompt assistance for those facing mental health difficulties. Ultimately, this initiative aims to connect the technical with the human, empowering communities and creating a lasting influence in the realm of mental health.

2. Problem Statement

Issues related to mental health like anxiety, depression, and thoughts of suicide have become more noticeable in society. Websites such as Reddit have become essential venues where people articulate their challenges, exchange experiences, and look for assistance. Subreddits focused on these subjects deliver a rich array of unorganized text data that provides essential insights into the dynamics of conversations surrounding mental health. Nonetheless, examining and classifying this information presents multiple difficulties. Individuals frequently articulate their mental health experiences using casual, vague, or context-specific language, making analysis more challenging. Furthermore, the language patterns among mental health classifications, like depression and suicidal thoughts, often intersect, complicating the differentiation between them. Additionally, broader subreddits such as r/CasualConversation contain posts that don't pertain to mental health, introducing data noise and compromising data quality. Finally, specific categories, like anxiety, are disproportionately represented in relation to others, resulting in biased data distributions that may impact model effectiveness.

This initiative seeks to categorize Reddit posts into specific mental health classifications—anxiety, depression, and general (healthy). Moreover, it aims to evaluate the seriousness of mental health disorders, particularly concentrating on recognizing high-risk signs like suicidal thoughts. The primary objective is to assist mental health researchers and practitioners by offering a structure for comprehending and examining online conversations about mental health.

2.1 Approach

In order to tackle the challenges mentioned above, this project utilizes a thorough Natural Language Processing (NLP) pipeline, which is organized as follows:

1. **Data Collection:** The initiative employs Reddit's PRAW API to gather posts from subreddits centered on mental health (e.g., r/Anxiety, r/depression, r/SuicideWatch) and general subreddits (e.g., r/CasualConversation) for comparative data. Subreddits function as identifiers that reflect particular mental health categories, degrees of severity, or general (well-being) conversations.
2. **Data Preprocessing:** Preprocessing steps include:
    - **Text Cleaning:** Removal of noise, such as URLs, special characters, and stopwords, to improve data quality.
    - **Lemmatization:** Reducing words to their base forms to retain core meanings while minimizing redundancy.
    - **Dimensionality Reduction:** Simplifying the dataset by reducing the number of features while preserving meaningful information, such as merging overlapping labels like r/depression and r/SuicideWatch under a unified "depression" category.
3. **Exploratory Data Analysis (EDA):** EDA uncovers key patterns in the data, including:
    - **Subreddit Distribution:** Identifying class imbalances and adjusting datasets to preserve diversity.

- **Text Length Analysis:** Analyzing word and character counts to standardize post lengths.
- **TF-IDF Analysis:** Highlighting the most important terms and their associations with specific subreddits.
- **Sentiment Analysis:** Measuring average sentiment scores to understand emotional trends across categories.
- **Topic Modeling:** Using techniques like Author-Topic Modeling to identify dominant themes in posts, such as therapy discussions in r/bipolar or crisis language in r/SuicideWatch.

4. **Model Development:**
   - **Baseline Models:**
     - **Most Frequent Label Classifier:** A simple model assigning the most common label (e.g., depression) to all predictions.
     - **Logistic Regression with TF-IDF:** Using term importance as numerical features to capture direct correlations between terms and labels.
   - **Advanced Models:**
     - **Structured Perceptron:** Captures dependencies between words in a sequence by labeling tokens and leveraging context within posts.
     - **LSTM:** Ideal for capturing long-term dependencies in text, LSTM models process sequences to predict nuanced mental health conditions and severity levels.
5. **Evaluation Metrics:** The models are evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics ensure robust performance, particularly in handling imbalanced and noisy data. Sequential models like LSTM are further assessed for their ability to capture temporal dependencies in longer posts.

2.2 Significance

This project is significant for its potential contributions to both computational research and mental health support:

- By analyzing textual data, the project provides insights into how mental health issues are discussed online. This can help researchers identify trends, linguistic markers, and emerging concerns in mental health discourse.
- The classification framework can assist mental health professionals in identifying high-risk posts (e.g., suicidal ideation) and responding promptly. Such tools could be integrated into platforms to alert moderators or professionals about critical cases.
- The findings can guide the creation of safer and more supportive online environments by fostering awareness of mental health themes and dynamics.
- The approach is adaptable to other datasets and platforms, enabling its application to broader mental health challenges and discussions.

By bridging advanced NLP techniques with real-world applications, this project lays the groundwork for impactful advancements in mental health analytics. It not only highlights the

value of interdisciplinary research but also emphasizes the ethical and practical importance of using technology to support mental well-being.

3. Data and Preprocessing

3.1 Data Collection

The first step in the data processing process was to collect relevant Reddit posts using the PRAW API. Various subreddits, including r/depression, r/Anxiety, r/SuicideWatch, and r/CasualConversation, were selected in order to record a range of discussions regarding mental health. Posts from these subreddits served as labelled data, representing various mental health conditions and control categories.

Subreddits for Mental Health Conditions:

- r/depression – Posts about depression struggles can be found in the /depression subreddit for mental health conditions.
- r/Anxiety – posts about anxiety-related concerns.
- r/SuicideWatch – posts indicating high-risk or crisis-related content can be found on or/SuicideWatch.
- r/bipolar, r/mentalhealth – Posts about more general mental health issues can be found at r/bipolar and r/mentalhealth.
- r/CasualConversation – posts that discuss topics other than mental health in order to provide fair comparisons.

Dataset Overview:

- Following preprocessing, roughly 6,000 Reddit posts were collected guaranteeing a dataset that was evenly distributed across categories.
- Classifying mental health conditions into predetermined categories by using subreddits as initial labels.

3.2 Data Preprocessing

Data processing is the foundation for any NLP project. In our project, distinguishing mental health categories from Reddit discussions required an extensive and structured data processing pipeline.

Effective data preprocessing was essential for ensuring data quality and model performance. Once the data was collected, we undertook an extensive preprocessing stage to clean, structure, and organize the data for analysis. This step helped reduce noise, standardize text, and ensure the dataset was ready for machine learning models.

The following sections describe our systematic approach, ensuring data quality, label accuracy, and balanced representation across various mental health categories such as anxiety, depression, and general conversations.

Handling Missing Data:

- Posts with missing or irrelevant content, particularly in the body field, were removed.
- Subreddits like r/bipolar and r/mentalhealth required special filtering to ensure meaningful content was retained.

Data Cleaning

- **Formatting Removal:** We eliminated Reddit-specific formatting like: Markdown symbols (e.g., "**", "*", ">" characters) and URLs, hyperlinks, and special characters.
  Tokenization: We split text into individual words (tokens) for better word-level analysis.
  Stopword Removal: Commonly used words like "the," "and," "is" were removed to reduce noise.
- **Punctuation Removal:** Punctuation marks were eliminated unless they added contextual meaning.

Lemmatization

- Words were reduced to their base forms using lemmatization.
- We focused on extracting nouns and adjectives since they carry significant meaning in expressing emotions and mental health states.

Data Partitioning

- The dataset was split into:
  – Training Set (80%) – Used to train models.
  – Validation Set (10%) – Used to tune hyperparameters.
  – Testing Set (10%) – Used to evaluate final model performance.

- We ensured balanced representation across mental health labels in each partition.

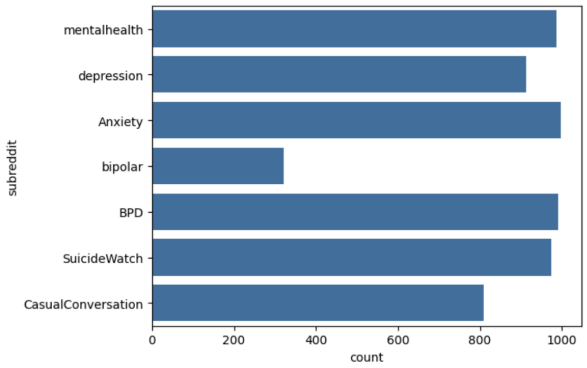
4. Exploratory Data Analysis (EDA)

4.1 Visualization & Statistical Insights

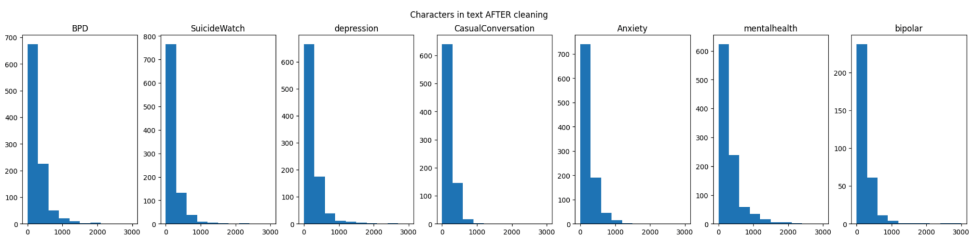
Visualization is an essential part of Exploratory Data Analysis (EDA) since it offers a clear and understandable perspective on data distribution, imbalances, and trends. In this project, visualizations were employed to evaluate the features of the Reddit mental health dataset and ready it for subsequent modeling.

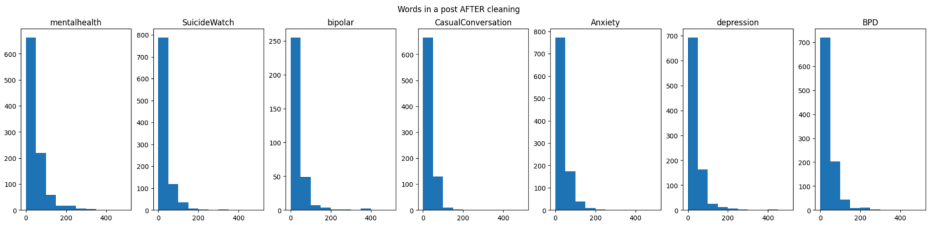
1. **Subreddit Distribution**: A bar graph was designed to show the distribution of posts in subreddits like r/depression, r/Anxiety, r/SuicideWatch, and r/CasualConversation. The chart showed notable class disparities, with subreddits such as r/depression and r/SuicideWatch featuring a higher number of posts than the rest. These disparities indicate the significant level of conversations about depression and suicidal thoughts on Reddit.

   To tackle these imbalances, the preprocessing pipeline involved removing excessively short or unrelated posts while ensuring a variety of subreddit representation. This guaranteed that the dataset was sufficiently thorough and balanced for significant analysis.



2. **Text Length Analysis**: Histograms illustrating the word and character totals prior to and following cleaning were created to evaluate the effects of preprocessing. At first, posts included unnecessary details like hyperlinks, markdown formatting, and various non-linguistic components. Histograms after cleaning indicated a significant decrease in text length, highlighting the elimination of superfluous information. For instance, r/Anxiety posts preserved their essential content while removing distractions, yielding a more refined dataset for examination.

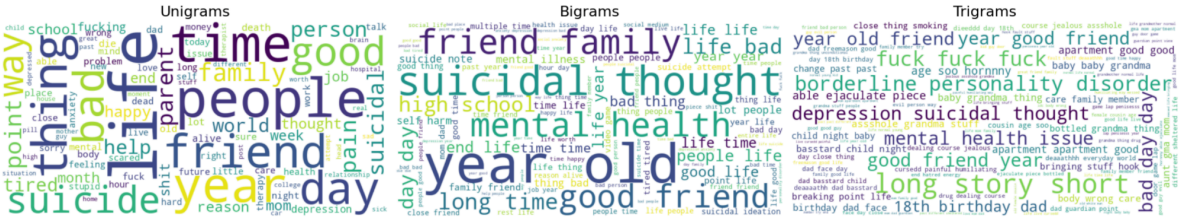




3. **Word Clouds:** Word clouds offered a visual overview of the most common terms in the dataset, presenting a qualitative viewpoint on the content of posts.

- **Unigrams**: Terms such as "life," "depression," and "help" were prevalent in subreddits focused on mental health.
- **Bigrams**: Frequently occurring phrases such as "mental health" and "panic attack" emphasized repeated themes.
- **Trigrams**: Contextual phrases, like "depression suicidal thoughts" illustrated the detailed conversations found in subreddits such as r/SuicideWatch.

These visual insights not only influenced preprocessing choices but also assisted in choosing features for modeling.



4. **TF-IDF Evaluation:** The Term Frequency-Inverse Document Frequency (TF-IDF) was computed to determine the words that best characterize each subreddit.

- In r/SuicideWatch, prominent TF-IDF terms such as "life," "people," and "suicide" highlighted the crisis-oriented character of conversations.
- In r/depression, words like "life," "time," and "depression" emphasized emotional hardships.
- Terms with low TF-IDF scores, like "sand," "scatter," and "watercolor," were either not relevant or consistently appeared across subreddits, providing little benefit for classification efforts.

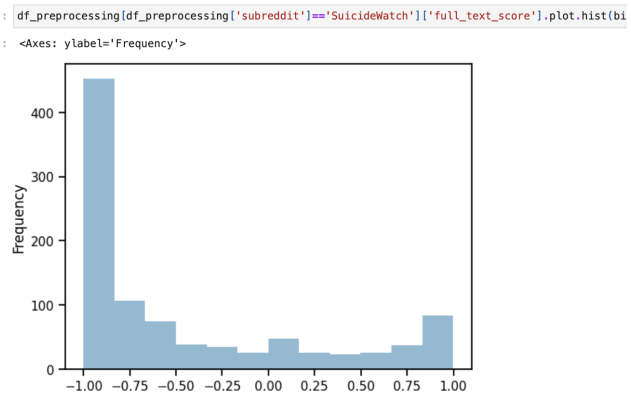
```
----  depression  ----

Words with highest TF-IDF score

         word   tf       idf      tfidf
0        life  782  1.866093  41.811753
1        time  673  1.944618  35.104992
2      people  539  2.171289  32.503732
3         day  521  2.181946  31.076752
4       thing  550  2.133156  30.213537
5      friend  485  2.263776  29.749695
6  depression  359  2.307348  27.727715
7        year  469  2.225749  26.635652
8         bad  359  2.369999  26.053195
9        good  351  2.400635  23.884599
```

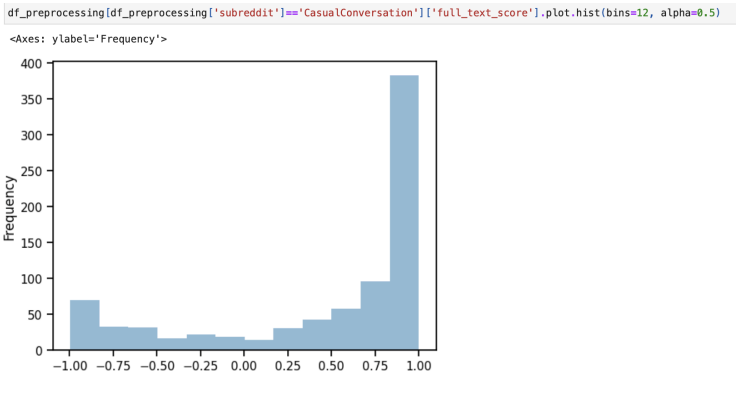
```
Words with lowest TF-IDF score

             word  tf       idf     tfidf
4492         sand   1  7.123589  0.073119
4493      scatter   1  7.123589  0.073119
4494      sighing   1  7.123589  0.073119
4495         slim   1  7.123589  0.073119
4496   straightup   1  7.123589  0.073119
4497         twat   1  7.123589  0.073119
4498       typing   1  7.123589  0.073119
4499    wasteland   1  7.123589  0.073119
4500   watercolor   1  7.123589  0.073119
4501   zombielike   1  7.123589  0.073119
```

The TF-IDF assessment guided feature selection, guaranteeing that models focused on words of greatest contextual relevance.

5. **Sentiment Assessment:** Emotional tone scores were computed to evaluate the feelings expressed in posts across subreddits.

- Posts from r/CasualConversation showed an overall positive sentiment, aligning with the typical, cheerful tone of conversations.
- Subreddits like r/SuicideWatch and r/depression exhibited negative sentiment scores, illustrating the intensity and hopelessness conveyed in these discussions.
- This analysis offered a numerical viewpoint on the emotional tone of posts, reinforcing the hypothesis that sentiment differs markedly between subreddits.





4.2 Author Topic Modeling

What does Author Topic Modeling mean?
Author Topic Modeling (ATM) is a type of topic modeling that links each author (or source) to a distribution of topics reflecting the content they create. In contrast to conventional topic modeling that detects themes throughout all documents, ATM includes authorship to reveal trends in how people address various subjects. This method is especially useful for datasets such as Reddit, where users share content in various subreddits.

Application to the Dataset:
ATM was employed to examine the distribution of topics among subreddits while considering authorship. This assisted in revealing themes specific to certain subreddits and recognizing overlaps in conversations.

1. **Prevailing Themes Throughout Subreddits:**
   – Posts in r/bipolar frequently centered on therapy and medication, showcasing organized treatment approaches for this condition.
   – In r/Anxiety, contained posts where users looked for urgent assistance, frequently mentioning "panic attacks" and "fear of future events."
   – r/CasualConversation served as a control group, with overarching discussions not focused on mental health prevailing in this subreddit.

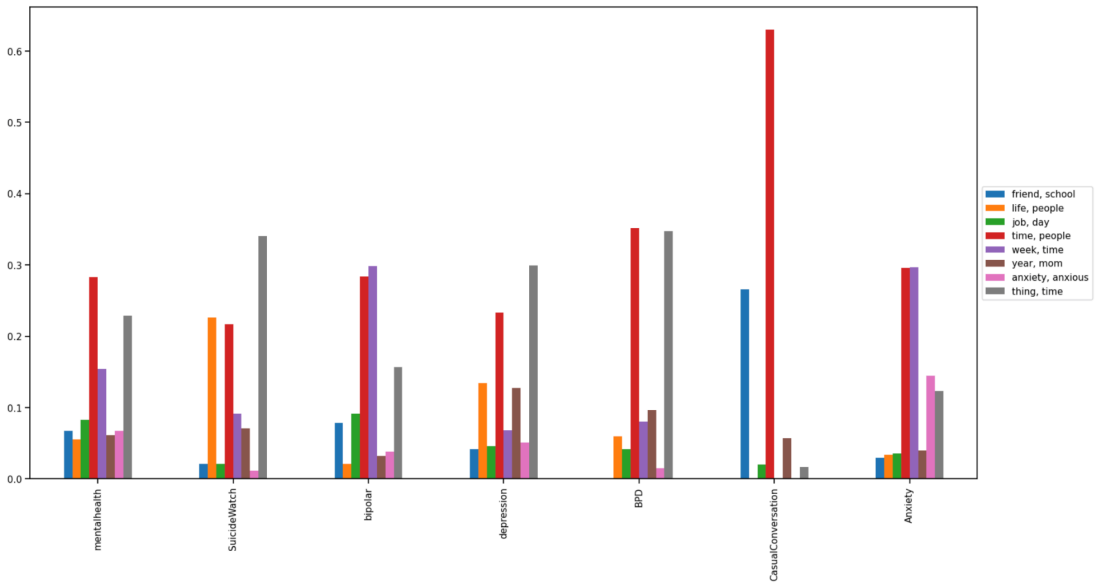
– Subjects in r/SuicideWatch shared similarities with those in r/depression but displayed a heightened emphasis on crisis circumstances and thoughts of suicide.

2. **Topic Coherence:** ATM indicated that r/depression and r/SuicideWatch exhibited notable linguistic similarities, which complicated the task of differentiating between them using topic modeling by itself. Metrics for coherence, including UMass and CV, were employed to assess topic quality, while preprocessing techniques like lemmatization enhanced these results.

3. **Significance**: By connecting authors to particular subjects, ATM underscored how individuals conveyed similar worries across different subreddits. For instance, writers who discussed depression in r/depression frequently also expressed similar worries in r/Anxiety. This interrelationship highlighted the necessity for detailed models that consider coexisting mental health disorders.

4. **Challenges**: The primary challenge with ATM was maintaining adequate coherence in topics while not compromising on granularity. Moreover, Reddit's chaotic and unorganized data necessitated significant preprocessing to obtain valuable outcomes.

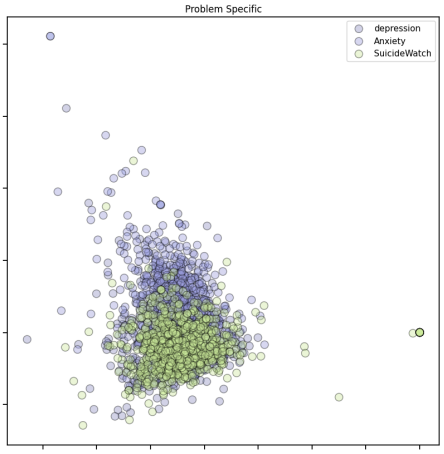
ATM enhanced the analysis, allowing for the recognition of user-specific trends and offering a more comprehensive insight into the dataset.



4.3 Dimensionality Reduction

What is Dimensionality Reduction?
Dimensionality reduction comprises techniques focused on reducing the count of features (or dimensions) in a dataset while retaining as much essential information as feasible. In high-dimensional data contexts such as text, these methods facilitate visualization and modeling. Techniques such as Truncated Singular Value Decomposition (SVD), t-SNE, and UMAP convert data into reduced-dimensional spaces, facilitating the recognition of patterns and clusters.

1. **Techniques Used**:
   - **Truncated Singular Value Decomposition (SVD)**: This technique was utilized on the TF-IDF matrix to condense the dataset into two dimensions for visualization purposes. SVD was selected due to its effectiveness in managing sparse matrices.
   - Other potential techniques such as t-SNE and UMAP were evaluated but not used because of computational limitations.
2. **Findings**:
   - **Distinct Clusters**: Subreddits such as r/CasualConversation created clear clusters, signifying their distinct vocabulary not associated with mental health.
   - **Overlap in Related Subreddits**: Subreddits such as r/depression and r/SuicideWatch demonstrated considerable overlap in the diminished space. This overlap emphasized common linguistic trends, indicating the strong link between depression and suicidal thoughts.
   - **Separation of Anxiety**: The posts from r/Anxiety were notably different, defined by words such as "panic" and "fear," differentiating them from other mental health conversations.
3. **Challenges and Solutions**:
   - The intersection of r/depression and r/SuicideWatch created difficulties for classification tasks. To tackle this, contributions from these subreddits were combined under a unified "depression" tag for baseline models, aiming to differentiate them in subsequent versions.
   - Labels with noise in other subreddits were addressed by meticulous preprocessing and dimensionality reduction, leading to enhanced clustering quality.
4. **Benefits of Dimensionality Reduction**:
   - Streamlined visualization enabled straightforward understanding of subreddit connections.
   - The process underscored the efficiency of preprocessing methods such as TF-IDF vectorization in identifying significant features.



5. Model Development

5.1 Baseline Models with Reduced Labels

The initial phase of the project simplifies the classification task by consolidating certain labels to create a reduced label set. Specifically, the subreddits ‘r/depression’ and ‘r/SuicideWatch’ are grouped together under ‘Depression’. This reorganization is vital because a lot of the conversations had within the r/depression subreddit have mentions of people’s lower points, times when the posts mention suicide attempts, or trauma related to it, making it more practical to classify posts into three broader categories. Subreddits like r/CasualConversations also deals with several topics – some anxiety, depression, suicide related, making it hard to categorize each post individually. By focusing on this simplified structure, the model development process becomes more tractable, interpretable in subsequent stages, and as a great baseline to work off of.

5.1.1 The Most Frequent Label Classifier

For the baseline models, two distinct approaches were used to establish ground level benchmarks. The *Most Frequent Label Classifier* is a naive model that assigns the most *common* label, ‘Depression’ to all input texts. It simply identifies the most common label in the training dataset and assigns this label to all test instances. This approach makes use of label frequency only and does not utilize any features or context from the text itself. It operates on the assumption that if a label is predominant in the dataset, it will likely be the correct prediction for any given input. Despite its simplicity and lack of ‘intuition’, this model provides a great comparison point to measure the improvement introduced by other more advanced methodologies. Its primary strength lies in its simplicity and speed, requiring minimal computation. However, it ignores deeper textual content and context, offering only a statistical baseline. Its accuracy mostly only depends on dataset imbalance: in a highly imbalanced dataset, it might perform deceptively well. Any comparative analysis needs a simple, more ‘basic’ benchmark to place deeper and more intuitive models against, and this helps form that. The Most Frequent Label Classifier is a simple statistical benchmark, useful for gauging improvement.

5.1.2 The Logistic Regression Model

The second baseline uses Logistic Regression, trained on features extracted using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. TF-IDF converts text data into numerical representations based on the importance of each term in the document relative to the corpus. For the scope of this project, TF-IDF was applied at both the post level (where each post is treated as an individual unit), and at the document level (where the contextual relationships across entire subreddits are captured). Despite Logistic regression being limited to linear decision boundaries, it is able to act as an effective classifier for these feature vectors and is able to capture term-level correlations very well, and is computationally efficient.

While this method can capture direct correlations between specific terms and labels to an extent, it sometimes struggles with nuanced or sequential patterns in posts, especially posts with temporal dependencies or deeper contextual knowledge, which limits its understanding of nuanced or evolving contexts. This, and the most-frequent-label classifier really highlights the need for advanced models capable of handling these types of complexities.

The baseline models, while foundational, are limited in their scope and capability. The Most Frequent Label Classifier relies mostly on label distribution and does not account for the text's features, resulting in simplistic predictions. Logistic Regression with TF-IDF provides a more refined approach compared to the Label Classifier, by incorporating term-level features but is constrained due to its inability to model sequential, long-term or contextual dependencies.

5.2 Advanced Models – Structured Predictions

To address the limitations of baseline models, advanced methodologies were explored, focusing on structured predictions that explicitly account for sequential and temporal dependencies in the data. Both advanced models mentioned below, and employed in the project, use a precursor allocation system to assign pseudo-tags to the dataset, to help give it more structure. The model starts with the generation of labeled sequences using Latent Dirichlet Allocation (LDA), which assigns pseudo-labels to text tokens based on their thematic content.
The pseudo labels assigned:
‘A’ – Mental health-related content
‘B’ – Non-mental-health-related content
‘O’ – Neutral or ambiguous terms.
For example, in the sentence “*I feel very anxious today*,” the word “*anxious*” might be tagged as ‘A’, indicating its relevance to mental health. This process removes the need for extensive, laborsome manual annotation and lets the model work on labeled sequences.

5.2.1 The Structured Perceptron Model

The Structured Perceptron model is designed to handle sequence labeling tasks, making it suitable for mental health-related text data that often exhibits contextual dependencies. It marks a shift toward sequential and structured predictions. Unlike independent classifiers, it considers the relationships between tokens in a sequence, making it ideal for text requiring contextual understanding.

Once the sequences are prepared using LDA, the perceptron model uses a combination of feature functions and weighted parameters to score sequences. The model iteratively updates these weights during training using a Viterbi decoding approach (Dynamic Programming) and a Greedy approach, which identifies the most likely sequence of labels at each step. When the predicted sequence matches the correct sequence, the model’s score is boosted, while every wrong sequence leads to a penalty in the score. While the Viterbi approach solves a subproblem to find the best score, the greedy approach aims to find the best sequence at each step instead of having to evaluate the whole problem and coming back to it. This iterative training makes sure that the perceptron captures dependencies between words and learns to assign labels in context,

rather than treating tokens independently like the baselines do. The structured perceptron is particularly adept at identifying patterns in nuanced mental health data, such as distinguishing between explicit and implicit references to mental states. For example, phrases like "*I'm not feeling myself lately*" may require context-aware interpretation, which the perceptron's sequence-based approach can handle effectively.
The Structured Perceptron introduces the ability to capture dependencies within text sequences. By leveraging pseudo-labels and structured prediction techniques, it excels at tasks requiring an understanding of word relationships within sentences. The only qualm is its reliance on manual or semi-automated labeling processes and limited capacity for long-term dependency modeling is resource-exhaustive and restricts its application to more complex data.

5.2.2 LSTM (Long Short-Term Memory)

The Long Short-Term Memory (LSTM) model represents a more deep-set sophisticated approach in its deep learning leverage, meant for text data with long-term dependencies. Mental health-related posts often contain extended narratives where the sentiment or context evolves across the text. For instance, a post might start with neutral descriptions and gradually reveal worsening mental states or a progression of repressed trauma manifesting itself in future behaviors. LSTMs are very well-suited for this scenario due to their ability to retain relevant information across long sequences while mitigating issues like vanishing gradients that might occur in traditional Recurrent Neural Networks (RNNs).

In this model, text sequences are first tokenized and padded to a fixed length, determined during the data exploration phase. Each token is then converted into a dense vector representation using an embedding layer, which can capture semantic relationships between words. For instance, the embeddings for "anxious" and "nervous" would be closer in the vector space compared to say "pizza", reflecting their similar meanings. These embeddings serve as inputs to the LSTM layers, where the model processes the sequence while retaining essential information through its cell state and gate mechanisms (input/add, forget, and output gates).

The output from the LSTM is passed through a dense layer with a softmax activation function to predict a label (A, B, or O) for each token in the sequence. The model is trained using categorical Cross-entropy Loss, a standard objective for multi-class classification problems, and optimization is performed using Backpropagation through Time (BPTT). This procedure makes sure that the model learns to identify not only local patterns in the training set, but also overarching trends in the text, such as worsening mental health over time.

LSTMs offer the bonus of maintaining long-term dependencies across longer sequences, making them particularly effective for analyzing lengthy posts or discussions in subreddits. The LSTM model outperforms the other three approaches in maintaining information across extended sequences and dynamically learning from previously provided context and 'intuition'. Its ability to process lengthy narratives and identify subtle patterns makes it a robust choice for mental health text classification. But this complexity comes at the cost of higher computational demands and the need for a larger processed and labeled dataset to prevent it from overfitting.

6. Results and Evaluation

We evaluated the performance of our models using multiple evaluation metrics, including accuracy, precision, recall, and F1-score. The models were tested across different training sizes and configurations, including variations in feature sets and representations such as TF-IDF at post-level and document-level

Baseline Models:

1. **Most Frequent Label Classifier**

The Most Frequent Label Classifier served as a baseline model for categorizing Reddit posts related to mental health. As anticipated, the model did not perform well on classes with lesser representation, like Anxiety and CasualConversation, where precision, recall, and F1-score were all measured at 0.00. This suggests that the model completely failed to forecast these categories. Nevertheless, it displayed satisfactory performance in the depression category, attaining a precision of 0.49 and a recall of 1.00, leading to an F1-score of 0.66. The overall accuracy was 0.49, indicating the classifier's inclination towards the dominant class (depression). The macro average F1-score of 0.22 and weighted average F1-score of 0.32 emphasize the model's failure to generalize across every category. This performance highlights the necessity for more sophisticated models that can address class imbalance and recognize subtle differences among mental health conditions.

	precision	recall	f1-score	support
Anxiety	0.00	0.00	0.00	204
CasualConversation	0.00	0.00	0.00	175
depression	0.49	1.00	0.66	360
accuracy			0.49	739
macro avg	0.16	0.33	0.22	739
weighted avg	0.24	0.49	0.32	739

2. Logistic Regression

– with TF-IDF at Post-Level

The Post-Level TF-IDF method greatly enhanced classification performance in comparison to the baseline model. The model attained a total accuracy of 0.78, showing consistent performance across every class. The Anxiety class attained the highest precision of 0.91 and an F1-score of 0.81, reflecting the model's capability to accurately recognize posts associated with anxiety. The depression category exhibited impressive outcomes, achieving a recall of 0.93 and an F1-score of 0.82, highlighting the model's ability to identify the dominant class effectively. The CasualConversation class, despite underperforming compared to other categories, attained a commendable precision of 0.75 and an F1-score of 0.61, indicating minor difficulties in differentiating non-mental health posts from associated categories. The macro average F1-score

of 0.75 and a weighted average of 0.77 suggest that the model effectively balances its performance among all classes. These findings validate that TF-IDF feature extraction at the post level successfully captures term significance and enhances classification results, surpassing the baseline classifier.

	precision	recall	f1-score	support
Anxiety	0.91	0.73	0.81	204
CasualConversation	0.75	0.52	0.61	175
depression	0.73	0.93	0.82	360
accuracy			0.78	739
macro avg	0.80	0.72	0.75	739
weighted avg	0.79	0.78	0.77	739

– with TF-IDF at Document-Level

The Document-Level TF-IDF method demonstrated modest performance enhancements over the baseline but was inferior to the post-level representation. The total accuracy obtained was 0.59, suggesting difficulties in accurately categorizing posts into all groups. The depression class achieved the highest performance, recording a recall of 0.99 and an F1-score of 0.71, highlighting the model's tendency to favor the majority class. Nonetheless, the Anxiety category, although attaining a high precision of 0.94, exhibited a significantly lower recall of 0.39, leading to an F1-score of 0.55. This underscores the model's inclination to overlook a considerable number of anxiety-related posts. The CasualConversation class exhibited weak performance, achieving a precision of 0.50 and a recall of merely 0.01, resulting in an F1-score of 0.01, which shows challenges in identifying non-mental health posts. The macro average F1-score of 0.42 and the weighted average F1-score of 0.50 indicate uneven performance among different classes. These findings indicate that although document-level TF-IDF identifies certain overarching trends, it has difficulty differentiating subtleties between similar categories, particularly for less represented classes.

	precision	recall	f1-score	support
Anxiety	0.94	0.39	0.55	204
CasualConversation	0.50	0.01	0.01	175
depression	0.55	0.99	0.71	360
accuracy			0.59	739
macro avg	0.66	0.46	0.42	739
weighted avg	0.65	0.59	0.50	739

Advanced Models

1. Structured Perceptron

The Structured Perceptron model showed capacity to identify and categorize various classes by utilizing contextual relationships in sequential data. The accuracy of training and development steadily heightened over several iterations, demonstrating the model's ability to learn from an expansive feature space.

In its predictions, the model effectively distinguished between sentences that convey positive feelings and those suggesting possible mental health issues. For instance, statements like "My life is so fantastic!" were categorized as not related to mental health, whereas phrases like "The medication isn't effective" were accurately recognized as part of the mental health category. This emphasizes the Structured Perceptron's ability to detect subtle patterns in text and to assign suitable labels according to context.

The findings indicate that the Structured Perceptron can manage sequential data, rendering it appropriate for intricate text classification projects, especially in the mental health field.

```
sp = StructuredPerceptron()
inference_method = 'greedy'
%time sp.fit(instances_train, instances_test, iterations=10, inference=inference_method)

sp.save('model_greedy.pickle')
```

```
................................................1000
................................................2000
................................................3000
.........................   314965 features
        Training accuracy: 0.91

        Development accuracy: 0.88

................................................1000
................................................2000
................................................3000
.........................   320913 features
        Training accuracy: 0.92

        Development accuracy: 0.89

................................................1000
................................................2000
................................................3000
.........................   323672 features
        Training accuracy: 0.92

        Development accuracy: 0.88

................................................1000
................................................2000
................................................3000
.........................   325370 features
        Training accuracy: 0.93

        Development accuracy: 0.88

................................................1000
................................................2000
................................................3000
.........................   326442 features
        Training accuracy: 0.93

        Development accuracy: 0.86

................................................1000
................................................2000
................................................3000
.........................   327104 features
        Training accuracy: 0.94
```

```
[ ] print(sp.predict('My life is so amazing!'.split(), method='greedy'))
```

```
['0', 'B', '0', '0', '0']
```

```
[ ] print(sp.predict('The medication is not working'.split(), method='greedy'))
```

```
['0', '0', '0', 'A', '0']
```

LSTM

Although the LSTM model began with low training and validation accuracy during the early epochs, it showed steady enhancements as the training continued. This demonstrates the LSTM's capability to recognize patterns across time, particularly for subtle sequences where context affects classification results.

In comparison to the Structured Perceptron, the LSTM model excelled in identifying and categorizing mental health-related terminology within sequential data. For instance, in the input sentence "I feel sad," the term "sad" was accurately recognized as part of the mental health category ('A'), whereas other neutral words were classified as unrelated to mental health ('0'). This underscores the LSTM's enhanced capability to recognize emotional signals and preserve contextual connections throughout token sequences.

The LSTM's ability to manage longer dependencies and its enhanced classification precision render it the top-performing model for this job. Though accuracy stays relatively low in initial epochs, additional fine-tuning and expanded datasets can greatly improve its performance, reinforcing its position as a strong option for sequential mental health text classification.

```
Epoch 1/5
180/180 ━━━━━━━━━━━━━━━━━━━━ 406s 2s/step - accuracy: 0.0576 - loss: 0.6735 - val_accuracy: 0.0589 - val_loss: 0.1661
Epoch 2/5
180/180 ━━━━━━━━━━━━━━━━━━━━ 431s 2s/step - accuracy: 0.0577 - loss: 0.1557 - val_accuracy: 0.0584 - val_loss: 0.1581
Epoch 3/5
180/180 ━━━━━━━━━━━━━━━━━━━━ 392s 2s/step - accuracy: 0.0594 - loss: 0.1413 - val_accuracy: 0.0587 - val_loss: 0.1522
Epoch 4/5
180/180 ━━━━━━━━━━━━━━━━━━━━ 445s 2s/step - accuracy: 0.0583 - loss: 0.1404 - val_accuracy: 0.0588 - val_loss: 0.1546
Epoch 5/5
180/180 ━━━━━━━━━━━━━━━━━━━━ 438s 2s/step - accuracy: 0.0589 - loss: 0.1315 - val_accuracy: 0.0601 - val_loss: 0.1547
```

```
Enter sentences separated by commas: i, am, sad
[(['i'], ['0']), (['am'], ['0']), (['sad'], ['A'])]
```

7. Conclusions

This initiative demonstrates how Natural Language Processing can address significant real-world issues, especially in the area of mental health. Through the examination of Reddit posts from subreddits dedicated to anxiety, depression, and general discussions, we successfully categorized various mental health topics. Sophisticated methods such as TF-IDF, structured perceptron models, and LSTM networks were instrumental in identifying the patterns and subtleties of these online conversations. Importantly, the project also revealed indicators of severity, like posts indicating suicidal thoughts, highlighting its potential for significant impact.

The procedure involved multiple stages, such as data preprocessing, exploratory analysis, and model creation, to guarantee that the insights obtained were solid and practical. Although baseline models served as a straightforward initial approach, the more sophisticated models delivered greater insights by revealing connections and context within the text. This careful mixture of methods underscores the importance of merging simple and advanced techniques to tackle intricate issues.

Aside from the technical results, this work has practical consequences. It offers a structure for comprehending how individuals discuss mental health on the internet, which may assist researchers and professionals in identifying early indicators and facilitating intervention efforts. Additionally, by enhancing comprehension of mental health discussions, this initiative can help in building safer, more compassionate online environments.

Certainly, there is potential for development. Broadening the dataset to encompass a wider range of perspectives, correcting disparities among categories, and partnering with mental health experts could enhance subsequent versions of this project. Investigating advanced technologies such as transformer-based models could aid in identifying subtle differences and long-term relationships in language, thereby enhancing accuracy and dependability.

In conclusion, this project has taken an important step in bridging technology and mental health advocacy. By using NLP to analyze sensitive and impactful topics, it paves the way for tools that are not only effective but also deeply considerate of the people they aim to serve.

8. Future Scope

This initiative presents multiple thrilling opportunities for upcoming investigation and utilization:

Model Improvement: Integrating sophisticated architectures such as transformers (e.g., BERT, GPT) might boost the model's capability to recognize intricate linguistic nuances and extended dependencies. This would enhance classification precision and the capability to manage more intricate datasets.

Partnerships with Mental Health Specialists: Working alongside professionals in psychology and psychiatry may enhance the models, ensuring they meet actual needs and ethical standards. Experts could likewise assist in confirming high-risk signals pinpointed by the system.

Enlarging the Dataset: Incorporating posts from a wider variety of subreddits and different language groups would enhance the model's inclusivity and flexibility. This may assist in tackling biases and enhance its effectiveness among various groups.

Practical Application: Implementing the model on university mental health platforms could offer early identification resources for counseling services on campus. For instance, a platform might evaluate anonymous student submissions to detect at-risk students and direct them to available support services. This integration would have a tangible and immediate effect on promoting mental well-being on campuses.

Applicability Beyond Reddit: This method might be applied to examine conversations on different platforms such as Twitter, Discord, or mental health forums, allowing the approach to be flexible across multiple digital environments.

Investigating Multimodal Data: Merging text analysis with various data formats, like audio (from podcasts) or video transcripts, could provide a more comprehensive perspective on mental health trends and behaviors, enhancing the model’s effectiveness even more.

The project has the capacity to evolve from a proof of concept to a broadly useful and influential resource in both academic and practical settings.

9. References

M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in Reddit social media forum,” *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

G. Gkotsis, T. Oellrich, S. Velupillai, R. Dobson, and R. Dutta, “Characterization of mental health conditions in social media using informed deep learning,” *Scientific Reports*, vol. 7, no. 45141, 2017.

M. De Choudhury, S. Counts, and E. Horvitz, “Language use on social media reveals an individual’s mental health status,” in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, Paris, France, 2013, pp. 353–362.

D. Yang, R. Kraut, and J. Lehmann, “Mental health and language on social media: Exploring depression and anxiety,” in *Proc. ACM Web Sci. Conf.*, Troy, NY, USA, 2017, pp. 1–10.

S. C. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, “Detecting depression and mental illness on social media: An integrative review,” *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, 2021.

P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V. Nguyen, and J. Boyd-Graber, “Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter,” in *Proc. NAACL-HLT Conf.*, Denver, CO, USA, 2015, pp. 99–107.

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

X. Sun, Y. Yuan, and X. Zhou, “Multi-task learning for mental health using LSTM networks,” in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Tokyo, Japan, 2017, pp. 1–10.

G. Shen, X. Feng, Y. Liu, and H. Zhang, “Multi-label classification of mental illness in social media posts using LSTM,” in *Proc. Int. Conf. Artif. Intell.*, 2019, pp. 347–353.

M. Kumar, A. Deshpande, and S. Kulkarni, “Toward balanced training data for mental health classification on Reddit,” in *Proc. ACL Workshop Comput. Linguist. Clin. Psychol.*, Seattle, WA, USA, 2020, pp. 182–193.