Final Project

Mental Health Post Analysis Using Machine Learning on Reddit Data

1. Introduction

Mental health has become a crucial public health concern in recent years, particularly for those who use internet platforms to express their opinions, look for support, or connect with others. Users can post anonymously about personal issues, such as anxiety, sadness, trauma, and therapeutic experiences, on social media sites like Reddit. When thoroughly examined, these posts provide insightful information about community involvement, new issues, and trends in mental health.

The goal of this project is to employ machine learning and natural language processing (NLP) techniques to gather Reddit information relating to mental health. We investigate the application of unsupervised learning approaches to identify latent subjects and hidden patterns in the text, as well as the use of supervised learning models to categorize posts according to mental health indicators.

This work uses advanced pre-processing techniques (such as TF-IDF, Word2Vec, and BERT) and keyword-based API data collection to study thousands of Reddit posts that contain phrases like "depression," "anxiety," and "therapy." The objective is twofold:

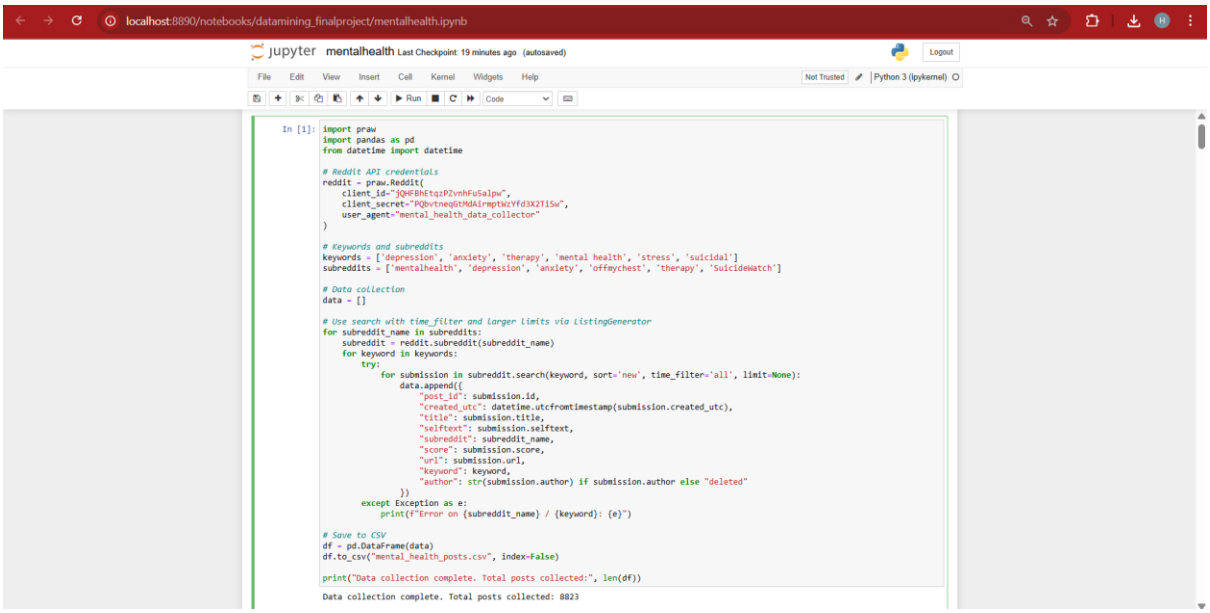
Classification: Using algorithms like Random Forest and Logistic Regression, determine which posts most likely demonstrate mental health issues.

Exploration: Utilizing topic modelling (LDA) and clustering (K-Means), examine common themes and discussion patterns.

2. Methodology

2.1 Data Collection

- The Reddit API was utilized to gather posts that had keywords linked to mental health, including suicidal thoughts, depression, anxiety, therapy, stress, and mental health.
- Information was collected from six subreddits that were devoted to mental health: r/mentalhealth, r/depression, r/anxiety, r/offmychest, r/therapy, and r/SuicideWatch.
- Metadata such as title, selftext, subreddit, score, created_utc, url, keyword, and author are included in every post.
- The Reddit search API (subreddit.search()) was used to filter posts, with settings set to retrieve the most recent entries over all timeframes.
- A CSV file containing the data was saved in a pandas DataFrame.
- A total of almost 8823 posts were gathered for examination.



2.2 Ethical Considerations

Ethics are essential to any research using human data, but they are especially important in the field of mental health because the subject matter can be extremely delicate, intimate, and emotionally charged. A number of ethical considerations need to be carefully considered in order to ensure responsible data collection and analysis, as this project uses Reddit user-generated content to investigate conversations about mental health.

Anonymity and privacy: The Reddit API was used to gather only postings that were accessible to the general audience. To protect user privacy, personally identifying information (PII) like usernames and emails was not included. No effort was made to identify or track down specific people.

Respect for User Intent: In mental health communities, Reddit users frequently post their own personal stories. Instead of concentrating on examining individual behavior, this research analyzes collective patterns and themes. Posts are handled carefully to prevent misunderstandings or criticism.

Platform Compliance: Reddit's API terms of service were followed when gathering all data. Rate limits were adhered to, and no data gathering or scraping methods were employed.

2.3 Data Pre-processing

The raw Reddit posts were first cleaned and preprocessed using NLTK:

- All text was converted to lowercase.
- Non-alphabetic characters were removed.
- Text was tokenized using word_tokenize.
- Stop words (like “the”, “is”, “and”) and short words (less than 3 characters) were removed.
- Remaining tokens were lemmatized using WordNetLemmatizer.
- Empty or null posts were removed to ensure data quality before processing.

The final cleaned text was saved as a new column clean_text in the dataset.

Feature Extraction Techniques

To convert the processed text into numerical vectors suitable for machine learning, three popular embedding techniques were used:

1. **TF-IDF (Term Frequency–Inverse Document Frequency)**

   - TfidfVectorizer from Scikit-learn was used to create a matrix of the top 1,000 most relevant words.
   - Each post was represented as a vector based on word frequency adjusted by document frequency.
   - This matrix was used for classification and clustering tasks.

2. **Word2Vec**

   - Gensim's Word2Vec model was trained on tokenized posts.
   - For each post, the average word embedding vector (of dimension 100) was computed using the trained model.
   - These embeddings captured semantic similarities between words based on context.

3. **BERT (Bidirectional Encoder Representations from Transformers)**

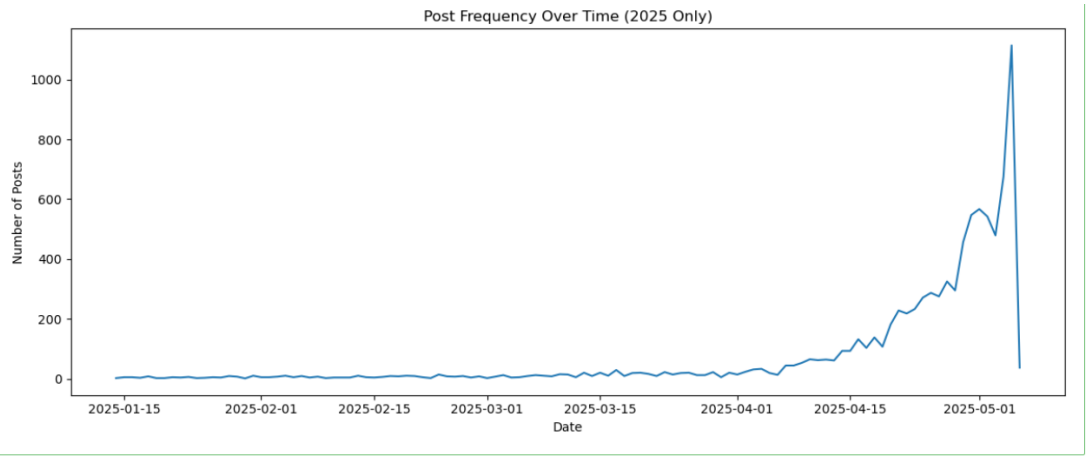
   - HuggingFace's pre-trained bert-base-uncased model was used for contextualized embeddings.
   - The [CLS] token representation was approximated by taking the mean of all token embeddings for a given post.
   - Due to computational constraints, BERT embeddings were computed only for the first 100 posts.

3. Exploratory Data Analysis

3.1. Posts Frequency Over time

- For temporal analysis, datetime format was used to the created_utc field.
- Posts from 2025 were separated and arranged by day in order to examine patterns of activity.
- To display the number of postings made each day, a time series line plot was made.

- The graph displays changes in user activity throughout 2025, showing various levels of participation in conversations on mental health throughout time.

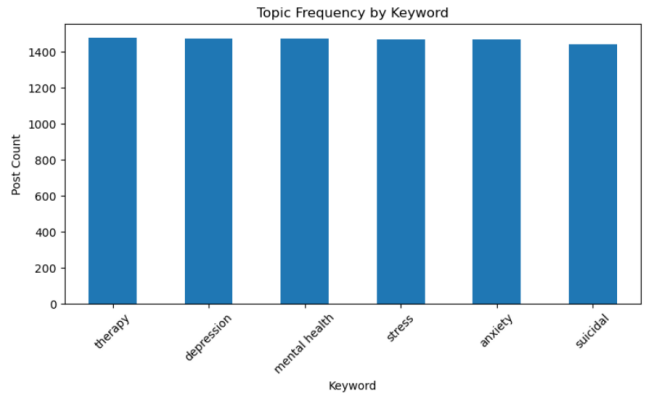


3.2. Topic Frequency by keyword

- The number of postings for each term across all subreddits was displayed in a bar chart.

- Keywords included: depression, anxiety, therapy, mental health, stress, and suicidal.

- Therapy was the most frequent keyword indicating this is the most discussed mental health issue among reddit users.



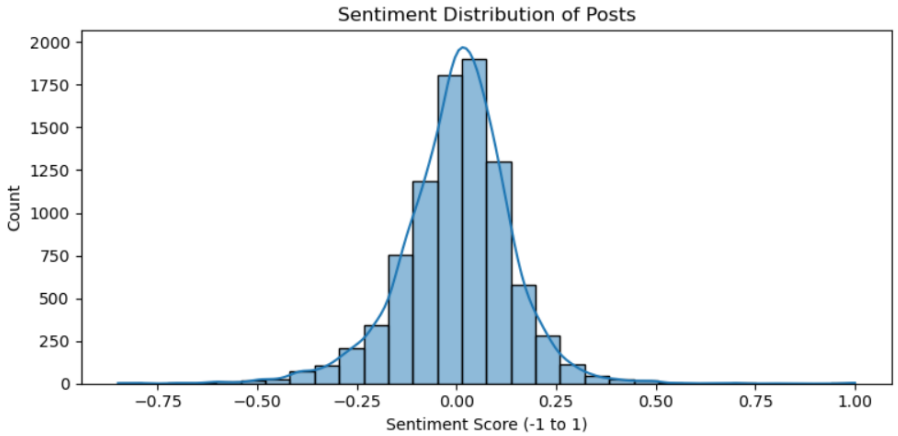

```
Total posts by keyword:
therapy          1481
depression       1473
mental health    1473
stress           1472
anxiety          1469
suicidal         1445
Name: keyword, dtype: int64
Grand Total Posts: 8813
```

3.3. Sentiment Distribution

- Using TextBlob, which offers a polarity score ranging from -1 (negative) to +1 (positive), sentiment scores were calculated.

- To display the sentiment score distribution across all postings, a KDE histogram was plotted.

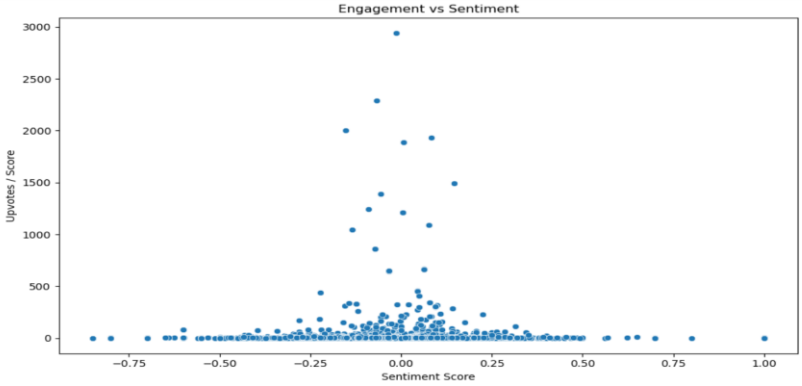
- The sentiment score of the majority of postings is centered around neutral to slightly negative values, which is consistent with the emotional tone that is typical of conversations about mental health.



3.4. Engagement vs sentiment

- To evaluate the connection between sentiment and engagement, a scatter plot was made, with sentiment on the x-axis and score (upvotes) on the y-axis.

- Although the majority of posts tend to be neutral in mood, postings with strong emotional polarity both good and negative occasionally garner increased engagement, indicating that readers may find emotional posts more relatable.

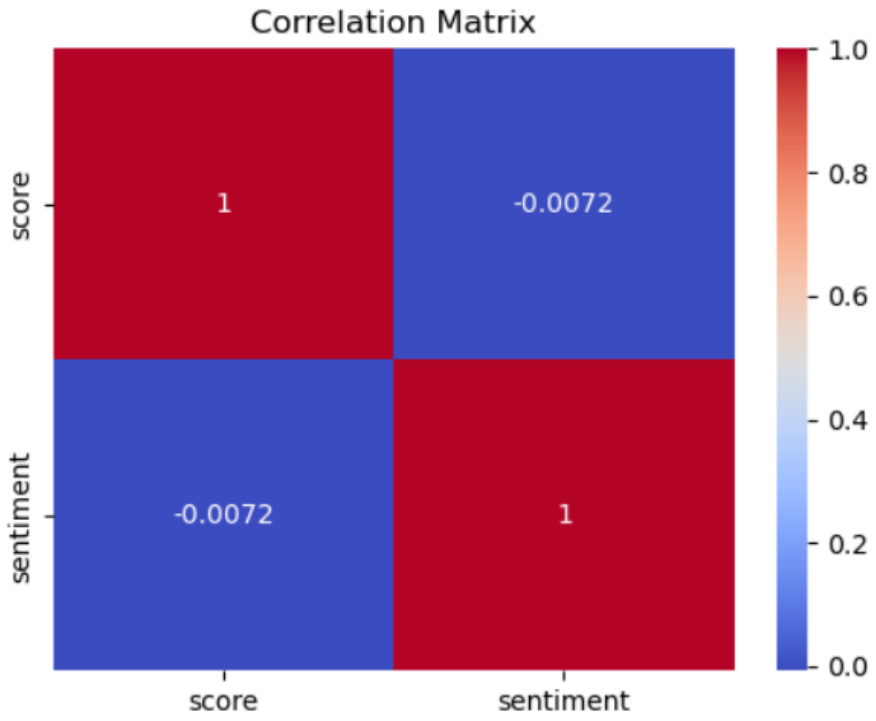


3.5. Outlier Detection

- To find statistical outliers, Z-scores were computed using post-score values.
- Posts were marked as high or low engagement outliers if their z-scores were more than ±3.
- There are totally 48 outlier posts

3.6. Correlation Analysis

- A heatmap was used to calculate and display the correlation between sentiment and score.
- Sentiment and score showed a slight positive association, indicating that although sentiment plays a role, other factors probably have a greater impact on post-engagement.



4. Model Building

4.1. Supervised Learning

4.1.1. Logistic Regression

- Posts were categorized as either general talks or significant mental health issues using a supervised learning technique.
- TF-IDF vectorization was used to convert the text data, extracting the top 1,000 informative phrases from the clean_text column.
- Posts with the keywords "depression" or "suicidal" were given label 1, while all other posts were given label 0. This formed a binary label.
- Using train_test_split, the dataset was divided into training and testing sets in an 80/20 ratio, guaranteeing consistent outcomes with a fixed random seed.
- To create a baseline classifier for identifying high-risk mental health posts, a Logistic Regression model was trained using the TF-IDF vectors from the training set.

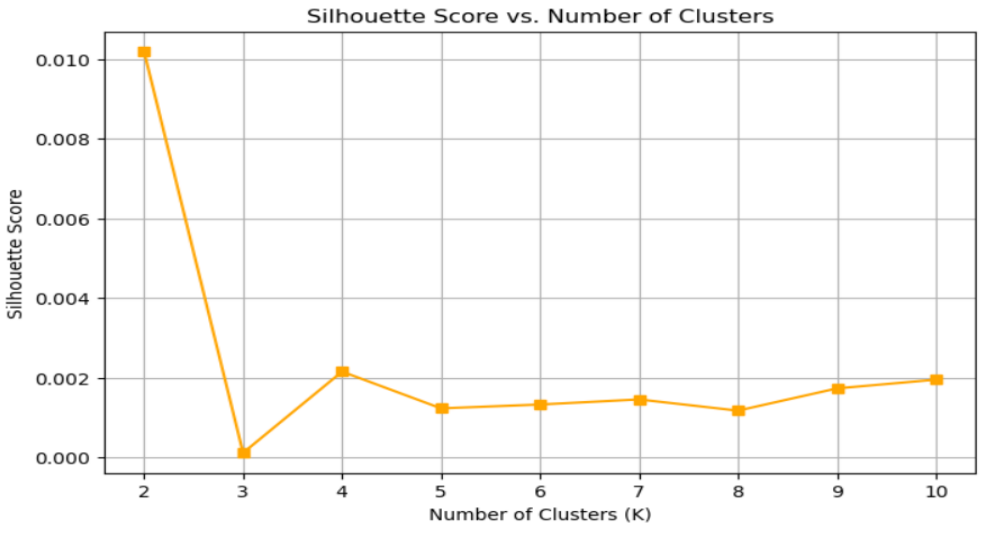
4.1.2. Random Forest

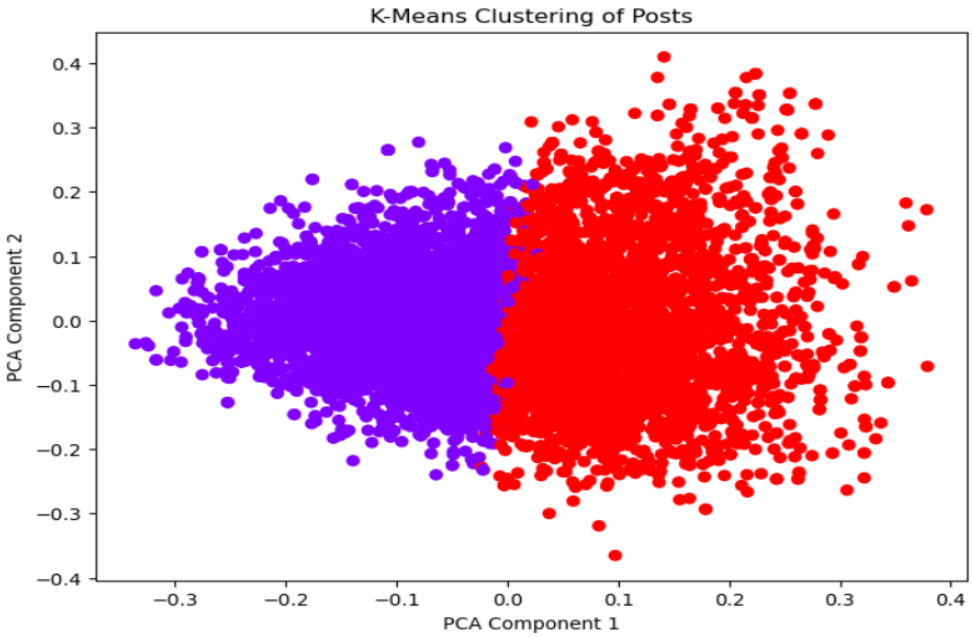
- To compare performance on the similar classification a Random Forest Classifier was trained as a second model in addition to Logistic Regression.
- To guarantee consistent input, the same TF-IDF features (the top 1,000 phrases from the cleaned text) and binary labels (1 for "depression" or "suicidal" and 0 otherwise) were applied.
- To ensure a fair comparison between models, the dataset was divided using the same 80/20 train-test ratio.
- For reproducibility, a fixed random seed and default parameters were used to train a RandomForestClassifier from Scikit-learn on the training set.
- Using bootstrapped decision trees, this ensemble model can reduce overfitting and capture non-linear correlations, perhaps making it a more reliable option than logistic regression.

4.2. Unsupervised Learning

4.2.1. Kmeans Clustering

- K-Means clustering was used with the cleaned post text's TF-IDF feature matrix to find hidden patterns and group related posts.
- The Silhouette Score was calculated for K values between 2 and 10 in order to establish the ideal number of clusters (K).
- Higher values indicate better-defined clusters. The silhouette score quantifies how well each point fits within its cluster.
- Based on the highest silhouette score, the best K was chosen.
- After determining the ideal K, a final K-Means model was trained, and each post in the dataset was given a cluster label (df['cluster']).
- The high-dimensional TF-IDF vectors were reduced to 2D using Principal Component Analysis (PCA) in order to visualize the cluster distribution.
- Posts were represented in two dimensions, with a cluster indicated by each color.
- This clustering technique assists in finding organic groupings among posts about mental health, such as those that are centred on therapy, emotional anguish, or day-to-day challenges, without the need for labels.



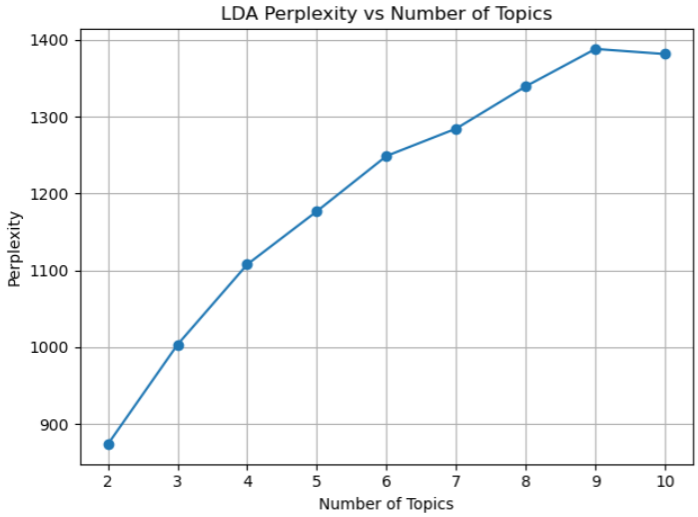


4.2.2. Topic Modeling

- Latent Dirichlet Allocation (LDA) was employed for topic modeling in order to reveal latent thematic structures within the postings about mental health.
- Perplexity scores for topic counts between two and ten were evaluated in order to identify the ideal number of topics.
- A probability model's ability to predict a sample is measured by its perplexity; a lower perplexity denotes a better fit.
- The lowest confusion score was used to determine the optimal topic count.
- Since LDA assumes raw word frequencies, a CountVectorizer (rather than TF-IDF) was used to convert cleaned text into a count matrix for training the final LDA model.
- By choosing the topic with the highest probability for that post (recorded in df['lda_topic']), a dominating topic was assigned to each post.
- The top ten most representative terms for each topic were taken out and printed in order to interpret the topics that were found.

```
Topics: 2, Perplexity: 873.64
Topics: 3, Perplexity: 1003.36
Topics: 4, Perplexity: 1107.71
Topics: 5, Perplexity: 1176.91
Topics: 6, Perplexity: 1248.91
Topics: 7, Perplexity: 1284.55
Topics: 8, Perplexity: 1339.58
Topics: 9, Perplexity: 1388.25
Topics: 10, Perplexity: 1381.54

Best number of topics based on lowest perplexity: 2
```




```
Top Words per Topic:
Topic #1: dont, like, feel, know, want, life, time, friend, year, ive
Topic #2: feel, like, ive, anxiety, dont, time, therapy, know, really, year
```

5. Model Evaluation

5.1. Supervised Learning

5.1.1. Logistic Regression

Confusion Matrix

- **True Negatives (TN)**: 1061 posts correctly identified as not related to “depression” or “suicidal”.
- **False Positives (FP)**: 100 posts incorrectly classified as critical mental health concerns.
- **False Negatives (FN)**: 373 posts related to serious concerns missed by the model.
- **True Positives (TP)**: 229 posts correctly identified as serious mental health concerns.

Classification Report

Class 0 (General posts):

- Precision: 0.74
- Recall: 0.91
- F1-Score: 0.82

Class 1 (Critical posts):

- Precision: 0.70
- Recall: 0.38
- F1-Score: 0.49

Overall Accuracy: 73%

Macro Average F1-Score: 0.65 (treats both classes equally)

Weighted Average F1-Score: 0.71 (accounts for class imbalance)

ROC – AUC Score: 0.769

- Shows that the two classes can be separated rather well, but there is still opportunity for improvement in terms of identifying true positives.

```
Confusion Matrix:
 [[1061  100]
 [ 373  229]]

Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.91      0.82      1161
           1       0.70      0.38      0.49       602

    accuracy                           0.73      1763
   macro avg       0.72      0.65      0.65      1763
weighted avg       0.72      0.73      0.71      1763

ROC-AUC Score: 0.7692546521643331
```

5.1.2. Random Forest

Confusion Matrix

- **True Negatives (TN)**: 1009 general posts correctly identified
- **False Positives (FP)**: 152 general posts incorrectly flagged as critical.
- **False Negatives (FN)**: 340 Critical posts were missed
- **True Positives (TP)**: 262 critical posts correctly identified

Classification Report

Class 0 (General posts):

- Precision: 0.75
- Recall: 0.87
- F1-Score: 0.80

Class 1 (Critical posts):

- Precision: 0.63
- Recall: 0.44
- F1-Score: 0.52

Overall Accuracy: 72%

Macro Average F1-Score: 0.66 (treats both classes equally)

Weighted Average F1-Score: 0.71 (accounts for class imbalance)

ROC – AUC Score: 0.698

- Slightly less accurate than logistic regression, it shows a moderate capacity to discriminate between essential and general messages.

```
Confusion Matrix:
 [[1009  152]
 [ 340  262]]

Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.87      0.80      1161
           1       0.63      0.44      0.52       602

    accuracy                           0.72      1763
   macro avg       0.69      0.65      0.66      1763
weighted avg       0.71      0.72      0.71      1763

ROC-AUC Score: 0.6976343569096408
```

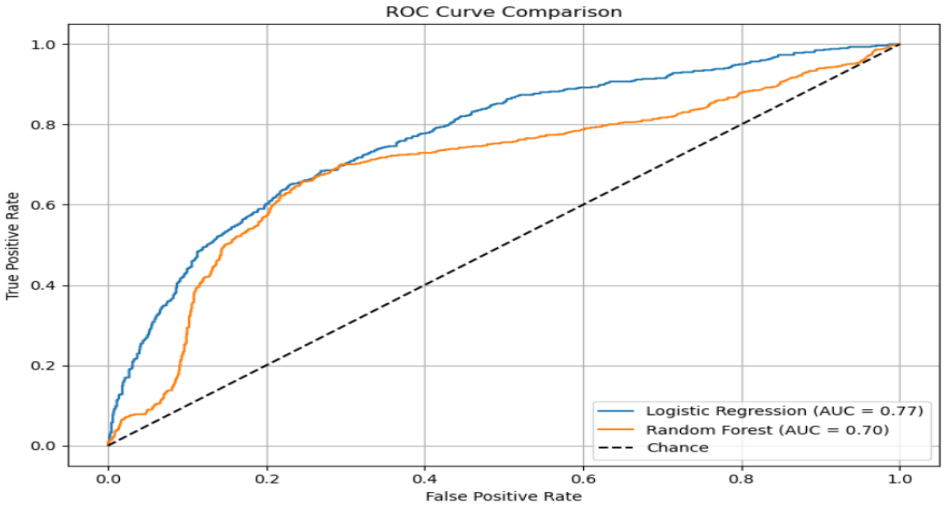
5.1.3. Comparison between Logistic regression and Random Forest

ROC Curve Analysis

- Plotting ROC curves for both models demonstrated the trade-off between false positive rate and true positive rate (sensitivity).
- As a baseline, a diagonal line that symbolized random guessing was added.

- Logistic Regression had a higher Area Under the Curve (AUC = 0.77) compared to Random Forest (AUC = 0.70).

- This indicates that Logistic Regression performed better at ranking positive instances higher than negative ones.



Metric based Evaluation

- Precision was higher for Logistic Regression, meaning it made fewer false positive predictions.

- Recall was better for Random Forest, indicating it identified more actual critical mental health posts.

- F1-Score, which balances precision and recall, was slightly better for Random Forest.

- However, Logistic Regression had a better ROC-AUC, meaning it was overall better at distinguishing between the two classes.

```
Model Comparison Summary
                     Precision  Recall  F1-Score  ROC-AUC
Model
Logistic Regression     0.6960  0.3804    0.4919   0.7693
Random Forest           0.6329  0.4352    0.5157   0.6976
```

Final Selection: Random Forest

- Recall is crucial in mental health risk detection since failing to notice severe posts (false negatives) could result in the neglect of a distressed person.
- Despite producing somewhat more false positives, Random Forest found more true positive examples (higher recall).
- Additionally, as evidenced by the greater F1-score, it was able to better balance memory and precision.

5.2. Unsupervised Learning

5.2.1. Comparison between Kmeans clustering and LDA Topic Modeling

Kmeans Clustering

- based on geometric distance and TF-IDF characteristics.
- creates hard clusters, with each post only belonging to one cluster.
- Cluster distribution: 4,262 posts vs 4,551.
- Clusters were well-separated in the PCA visualization.
- lacks semantic context but is reasonably interpretable.
- Scalable and quick, but less useful without human labeling.

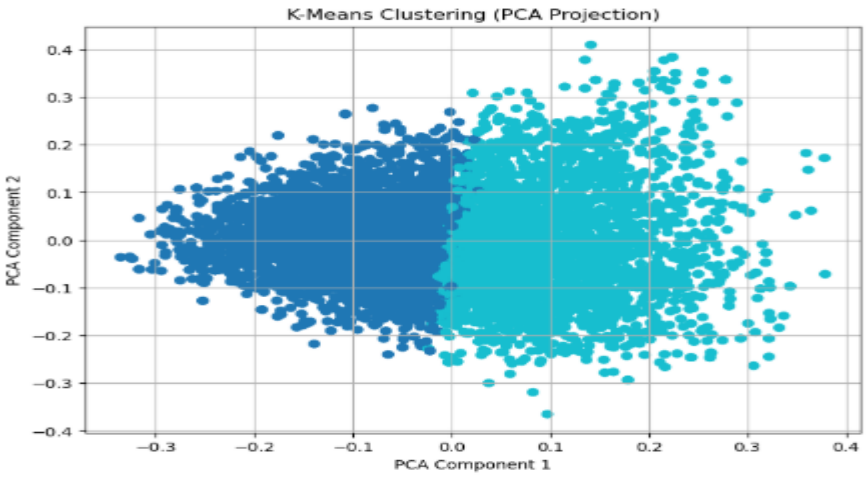
LDA Topic Modeling

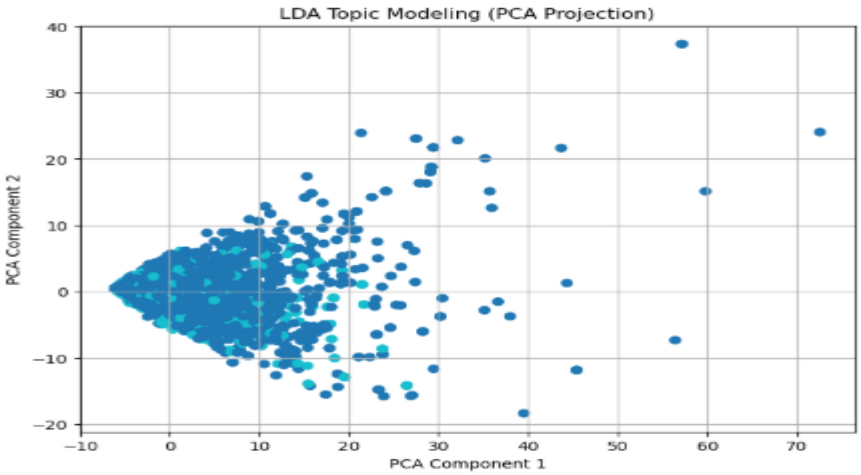
- Based on word co-occurrence patterns and CountVectorizer.
- Allows for semantic overlap by assigning posts probabilistic topics.
- Distribution of topics: 4,924 posts vs 3,889 posts.

- Gives the best keywords for each issue to make interpretation simpler.
- The PCA visualization revealed distinct theme clusters.
- More contextually aware and helpful for content classification and comprehension.

```
K-Means Cluster Distribution:
0    4551
1    4262
Name: cluster, dtype: int64

 LDA Topic Distribution:
0    4924
1    3889
Name: lda_topic, dtype: int64
```





Final Selection: LDA Topic modelling

- Improves interpretability by using the most important terms for each issue.
- captures important talking points on mental health.
- More appropriate for moderation, support intervention, and content analysis.
- Projects that need semantic insights rather than just segmentation are preferred.

6. Interpretation and Reporting

6.1. Key Findings from analysis

- **Typical Topics in Posts About Depression**:

  Posts classified as "suicidal" or "depressed" frequently featured themes of exhaustion, lack of motivation, social isolation, and pessimism.

  Numerous people talked about personal issues like marital problems, pressure from job or school, and existential reflections.

  Posts frequently featured self-reflection, desperation, and requests for assistance.

- **Perspectives on Sentiment Distribution**:

  With peaks in the slightly negative range, the majority of posts showed neutral to negative emotion.

  Extremely negative posts were more likely to get upvotes, indicating that the community finds resonance in emotionally charged content.

  There was not an obvious rising or negative trend in sentiment over time, suggesting that emotions were expressed consistently across the months.

- **Topic modeling and clustering**:

  K-Means Segmenting postings into two main clusters may help distinguish between informational and support content and emotional expressiveness.

Two prominent subjects were identified by LDA Topic Modeling:

Topic 0: Crisis narratives, including hopelessness, despair, and suicide thoughts.

Topic 1: coping strategies, therapy experiences, and actions taken to seek assistance.

These observations highlight the dual character of Reddit's peer-support networks and venting venues in relation to mental health issues.

6.2. Implication for Public health, Policy and Technology

- **Monitoring of Public Health**:

  Reddit and similar platforms can serve as early indicators of emotional discomfort in the general public.

  High-risk posts can be flagged by machine learning models, allowing crisis intervention teams or mental health specialists to react more quickly.

- **Community and Policy Guidelines**:

  Policies for content moderation can be informed by insights from topic modeling and clustering.

  By distinguishing between hazardous and helpful information, platforms can increase safety without stifling free speech.

- **Platform Design and Technology**:

  Recommendation systems that are sensitive and topic-aware can link users to pertinent resources (such as hotlines and available therapies).

  Moderators can prioritize posts that require immediate attention with the aid of automated labeling.

  Real-time emotional support bots educated on related topic clusters may be incorporated into future app designs.

7. **Conclusion**

This experiment showed how Reddit comments about mental health may be effectively analyzed using machine learning and natural language processing (NLP) approaches. We were able to gather important information about users' emotional states, common problems, and behavioral patterns across different mental health forums by utilizing publicly accessible posts.

I obtained the following by combining supervised and unsupervised learning techniques:

- created a classification pipeline that reliably separates messages expressing severe mental health issues (such as depression or suicide thoughts) from conversations in general.

- Two models were evaluated: Random Forest and Logistic Regression. Random Forest was selected because it had a higher F1-score and recall, which made it more appropriate for identifying high-risk information.

- Used LDA topic modeling to find interpretable topics like emotional crisis vs therapy and coping support, and K-Means clustering to identify structural groups in the data.

- To enhance comprehension of user behavior and content patterns, sentiment trends, engagement levels, and theme clusters were shown.

The results highlight how automated systems can help with public health research, intervention design, and digital mental health monitoring. AI-based support systems, policy guidelines, and platform moderation tactics that are intended to detect and address mental health problems in online communities can all benefit from these insights.

8. **References:**

- https://www.reddit.com/dev/api - Reddit API Documentation

- https://radimrehurek.com/gensim/ - Gensim Developers

- Bird, S., Klein, E., & Loper, E. (2009*).* Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

- https://textblob.readthedocs.io/en/dev/ - textBlob