***A PROJECT ON***

# “MENTAL HEALTH SENTIMENT ANALYSIS”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYTICS



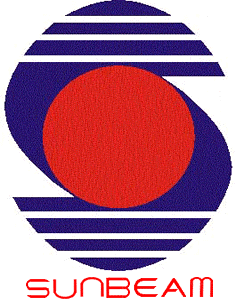
**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

Submitted By:

Sana Hungund (86900)

Anushka Kute (87111)

**Mr.Nitin Kudale Mrs.Manisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘Walmart Stores Sales Prediction’ is done by Sana Hungund & Anushka Kute in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date:

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Sana Hungund

DBDA August 2024 Batch, SIIT Pune

Anushka Kute

DBDA August 2024 Batch,

SIIT Pune

.

**TABLE OF CONTENTS**

1. **Introduction**
   1. Introduction And Objectives
   2. Why this problem needs To be Solved?
   3. Dataset Information

## Problem Definition and Algorithm

* 1. Problem Definition
  2. Algorithm Definition

## Experimental Evaluation

* 1. Methodology/Model
  2. Exploratory Data Analysis

## Results And Discussion

1. **GUI**
2. **GitHub link**

## 7.Future Work And Conclusion

* 1. Future Work
  2. Conclusion
     1. **Introduction**
        1. **Introduction And Objectives:**

Mental health issues such as depression and anxiety have been on the rise, making it crucial to develop systems that can automatically detect early signs of mental distress. This project leverages BERT (Bidirectional Encoder Representations from Transformers) to classify text statements into different mental health categories.

Objectives:

* To develop an NLP-based model capable of classifying mental health-related text.
* To train the model using a BERT-based deep learning approach.
* To deploy the model using Streamlit for user-friendly interaction.
* To improve prediction accuracy using data pre-processing and oversampling techniques.
* To integrate the system with real-world applications for mental health support.

## Why this problem needs To be Solved?

Early detection of mental health issues is essential for timely intervention. By automating sentiment classification, individuals in distress can be directed to appropriate mental health resources, potentially preventing severe outcomes. Many people hesitate to seek professional help, and an automated tool can act as an early screening mechanism.

## Dataset Information.

The dataset used for this project consists of text statements labeled with their corresponding mental health statuses. The data has undergone pre-processing, including cleaning, stopword removal, and tokenization. To ensurep better classification, data balancing techniques such as oversampling were implemented.

## Combined \_Data.csv :

## It has three columns.

## **Index :** The index number of the row

## **Statement :** (Text Column)

## Contains user-generated text or sentences expressing thoughts, feelings, or concerns. Examples include questions, confessions, or emotional expressions. Some entries contain short phrases, while others are longer sentences.

## **Status :** (Categorical Column)

## Represents the classification or sentiment label for each statement. In your example, all values are labeled as "Anxiety". This column likely comes from a sentiment analysis or mental health classification model.

## The dataset has 52,681 rows, it is a large-scale sentiment dataset, and is split into 80% training and 20% testing for BERT-based mental health sentiment analysis model.

## Problem Definition and Algorithm:

* + - 1. **Problem Definition**

Given a dataset containing text statements related to mental health, the goal is to classify each statement into different mental health categories viz, normal, anxiety, depression,bipolar,stress and suicidal; using BERT for sequence classification. The project focuses on building a system that can analyze the sentiment in text and predict a possible mental health condition.

## Algorithm Definition

## The model is implemented using BERT (bert-base-uncased) with fine-tuning. The approach includes:

1. Data Preprocessing: Removing stopwords, special characters, and unnecessary text elements. Stopwords like "the," "and," and "is" are removed, while punctuation is also eliminated to focus on meaningful words.
2. Tokenization: Using BERT’s tokenizer to convert text into subword tokens, ensuring better handling of rare words.
3. Data Augmentation: Generating synthetic text samples to balance dataset distribution. This prevents class imbalance issues, where certain categories might have fewer examples.
4. Model Training: Fine-tuning the BERT model for classification using a transformer-based approach. The model is trained with a learning rate of 2e-5, batch size of 16, and optimized using AdamW optimizer.
5. Evaluation: Using key metrics like accuracy, precision, recall, and F1-score to assess performance. A confusion matrix is used to visualize classification errors.

**BERT**

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning-based NLP model developed by Google in 2018. It is designed to understand the context of words in a sentence by considering both previous and next words, making it bidirectional. Unlike traditional NLP models that analyze text sequentially, BERT enables models to understand the full meaning of a sentence by analyzing words in relation to each other.

BERT is pre-trained on vast amounts of text data, including books, Wikipedia, and other large corpora, and can be fine-tuned for specific tasks such as text classification, sentiment analysis, and question answering.

### How BERT Works?

BERT uses a transformer-based architecture that consists of self-attention mechanisms to weigh the importance of different words in a sentence. The key processes in BERT include:

1. Tokenization: BERT splits sentences into subwords using WordPiece tokenization. This helps in handling out-of-vocabulary words by breaking them into smaller known units.
2. Masked Language Model (MLM): During pre-training, BERT randomly masks words in a sentence and predicts them based on surrounding words. This allows the model to learn bidirectional context.
3. Next Sentence Prediction (NSP): The model is trained to determine whether one sentence follows another in the original text, improving its understanding of sentence relationships.

### Applications of BERT

* Text Classification: Categorizing text into predefined classes, such as sentiment analysis and spam detection.
* Named Entity Recognition (NER): Identifying names, locations, and other important entities in text.
* Question Answering: Used in search engines and chatbots to extract precise answers from large documents.
* Machine Translation: Enhancing the accuracy of language translation models.
* Text Summarization: Generating concise summaries from large documents or articles.

## 3. Experimental Evaluation:

**3.1 Methodology:**

The methodology for this project consists of multiple steps that involve **data collection, preprocessing, model training, evaluation, and deployment**. The structured approach ensures that the sentiment classification task is optimized for accuracy and interpretability.

### ****Step 1: Data Collection****

The dataset consists of mental health-related text statements labeled into different sentiment categories (e.g., Normal, Anxiety, Depression). The data was collected from various online sources and preprocessed to remove irrelevant information.

### ****Step 2: Data Preprocessing****

To enhance model performance, data preprocessing was conducted as follows:

**Text Cleaning:** Removal of special characters, numbers, and extra whitespace.

**Stopword Removal:** Eliminating common words (e.g., "the," "is") that do not contribute to sentiment analysis.

**Tokenization:** Converting sentences into word tokens using **BERT’s tokenizer**.

**Data Balancing:** Addressing class imbalances using **Random OverSampling** to ensure equal representation of each sentiment category.

## Loading in raw data

data=pd.read\_csv("/content sample\_data/Combined Data.csv")

# Drop Unnamed column

data = data.drop(columns=['Unnamed: 0'])

data.dropna(inplace=True)

## Pre - processing:

stop\_words = set(stopwords.words('english'))

def clean\_statement(statement):

statement = statement.lower()

statement = re.sub(r'[^\w\s]', '', statement)

statement = re.sub(r'\d+', '', statement)

words = statement.split()

words = [word for word in words if word not in stop\_words]

cleaned\_statement = ' '.join(words)

return cleaned\_statement

data['statement'] = data['statement'].apply(clean\_statement)

Purpose: Cleans a text column by removing punctuation, numbers, and stopwords.

Key Techniques: Lowercasing, regex-based cleaning, list comprehension.

Final Result: Text is simplified and cleaned, making it better for NLP tasks like sentiment analysis.

Resampling Data

from imblearn.over\_sampling import RandomOverSampler

import pandas as pd

ros=RandomOverSampler(sampling\_strategy='auto', random\_state=42)

X = data.drop(columns=['status'])

y = data['status']

X\_resampled,y\_resampled = ros.fit\_resample(X, y)

data = pd.concat([X\_resampled, y\_resampled], axis=1)

print(data['status'].value\_counts())

Purpose: Balances an imbalanced dataset using random oversampling.

1. Drops the target column (status) from features.
2. Uses RandomOverSampler to increase the minority class samples.
3. Merges the resampled dataset into a new data DataFrame.
4. Verifies the new class distribution.

The dataset now has an equal number of samples in each class, improving performance for ML models.

### ****Step 3: Feature Engineering and Tokenization****

Text statements were **tokenized** using the **BERT tokenizer**, which breaks words into meaningful subwords.

Tokenized data was converted into numerical format for input into the deep learning model.

### ****Step 4: Model Training and Fine-Tuning****

The **BERT-base-uncased** model was fine-tuned using labeled sentiment data.

**Training Parameters:**

Learning Rate: **2e-5**

Batch Size: **16**

Optimizer: **AdamW**

Epochs: **5**

Loss Function: **Cross-Entropy Loss**

The training process was conducted using **PyTorch and Hugging Face’s Trainer API**.

### ****Step 5: Model Evaluation****

To assess model performance, key evaluation metrics were used:

**Accuracy**: Measures the percentage of correctly predicted labels.

**Precision**: Evaluates how many of the predicted positive cases are actually correct.

**Recall**: Measures how well the model identifies all relevant instances.

**F1-Score**: Harmonic mean of precision and recall for a balanced performance measure.

**Confusion Matrix**: Visualizes classification performance and identifies misclassification errors.

### ****Step 6: Deployment and User Interface (UI)****

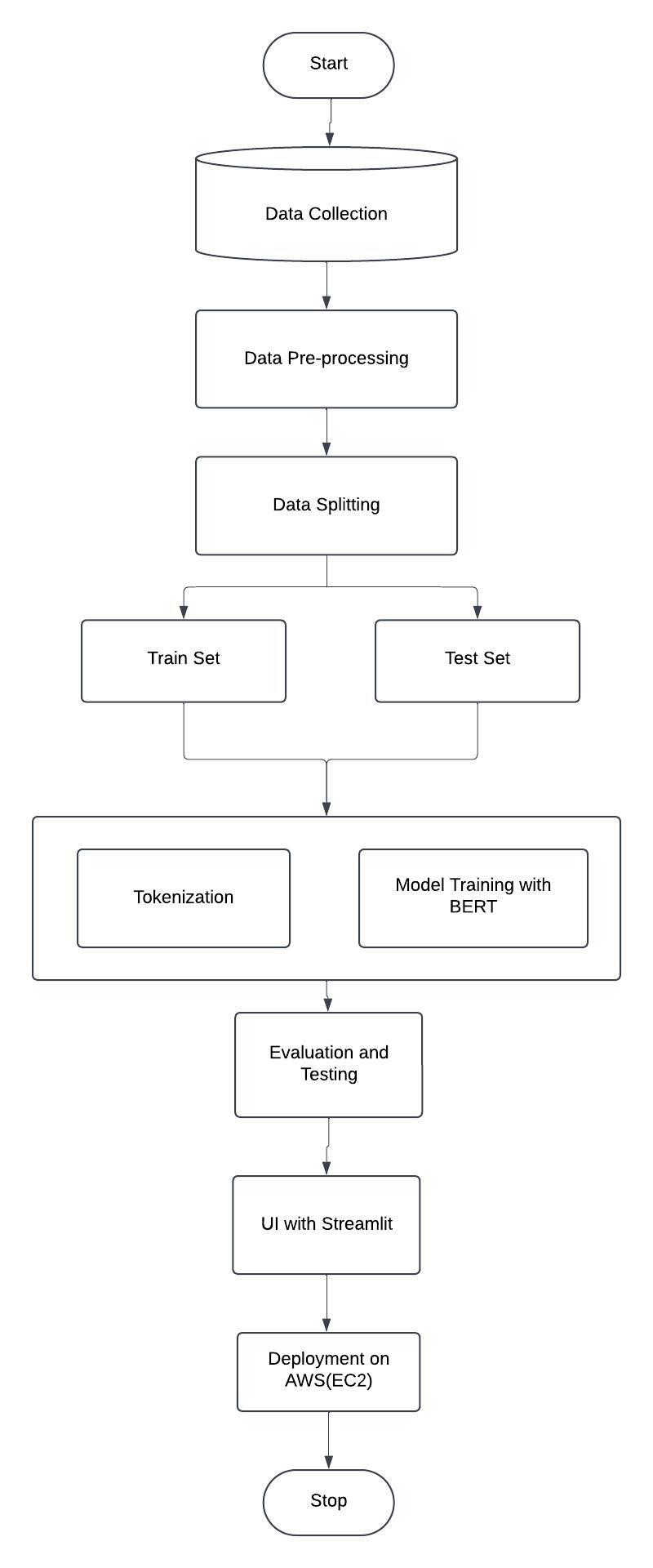
The trained model was deployed using **Streamlit**, allowing real-time sentiment classification.

The UI provides an input field for users to enter mental health statements and receive sentiment predictions instantly.

The model was deployed on **AWS EC2**, making it accessible via a web interface.

By following this methodology, the system ensures an **accurate, scalable, and user-friendly** solution for mental health sentiment analysis.

**Flow Diagram :**

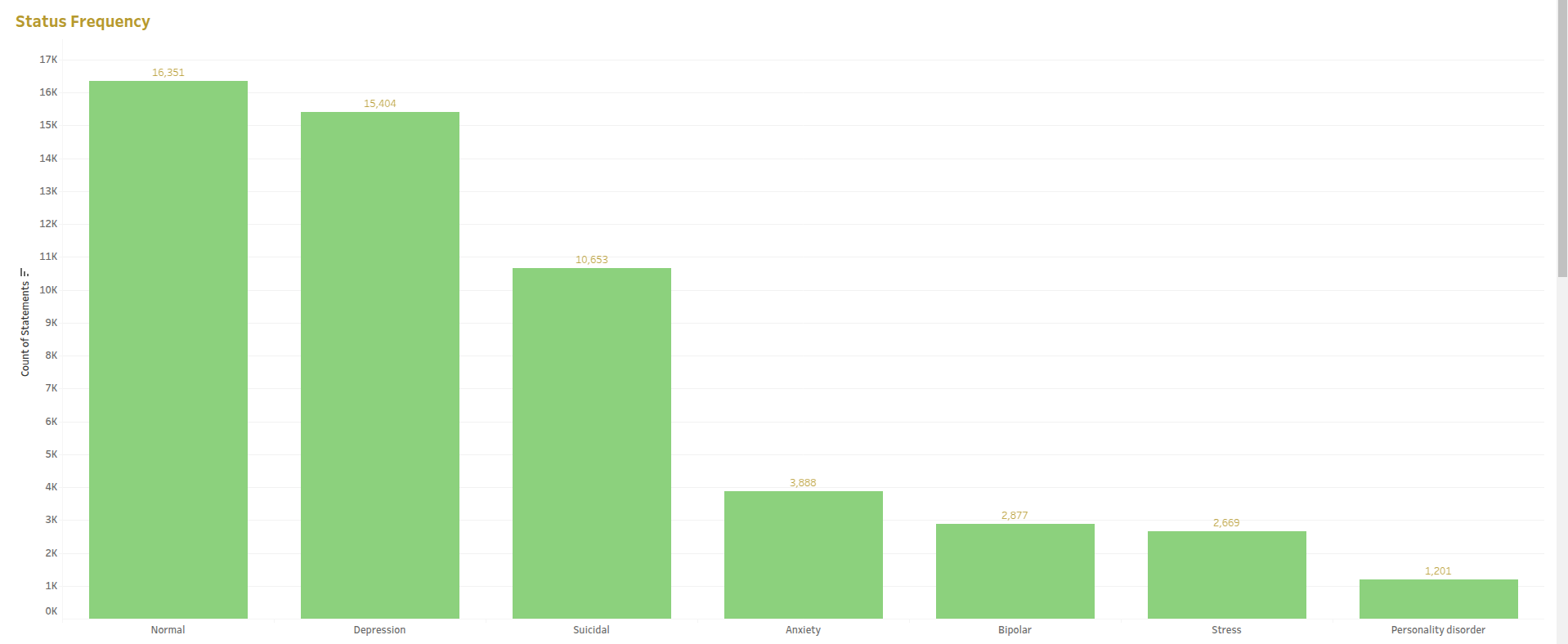


## **3.2 Exploratory Data Analysis**

## **Data Visualization and Insights**

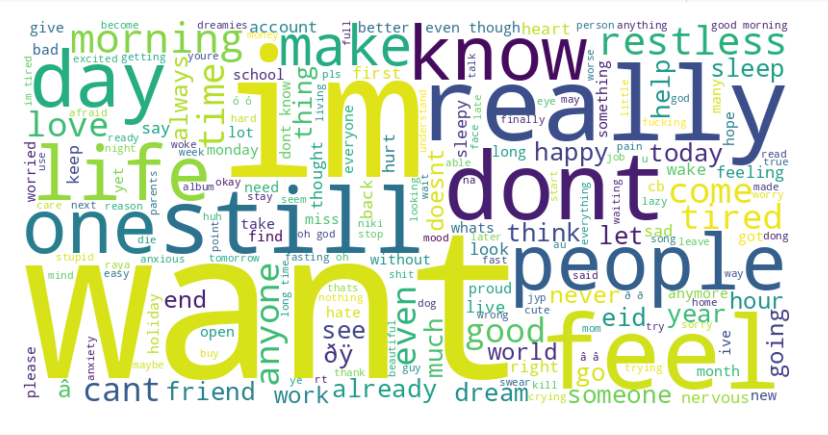
To better understand the distribution of mental health sentiments in our dataset, we created various visualizations using Tableau. The key findings are:

### 1. Sentiment Frequency Distribution

A bar chart was generated to show the frequency of different mental health sentiments (e.g., Normal, Anxiety, Depression, etc.). This helped in identifying any class imbalances, which were later addressed using oversampling techniques.

### 2. Word Cloud of Common Phrases

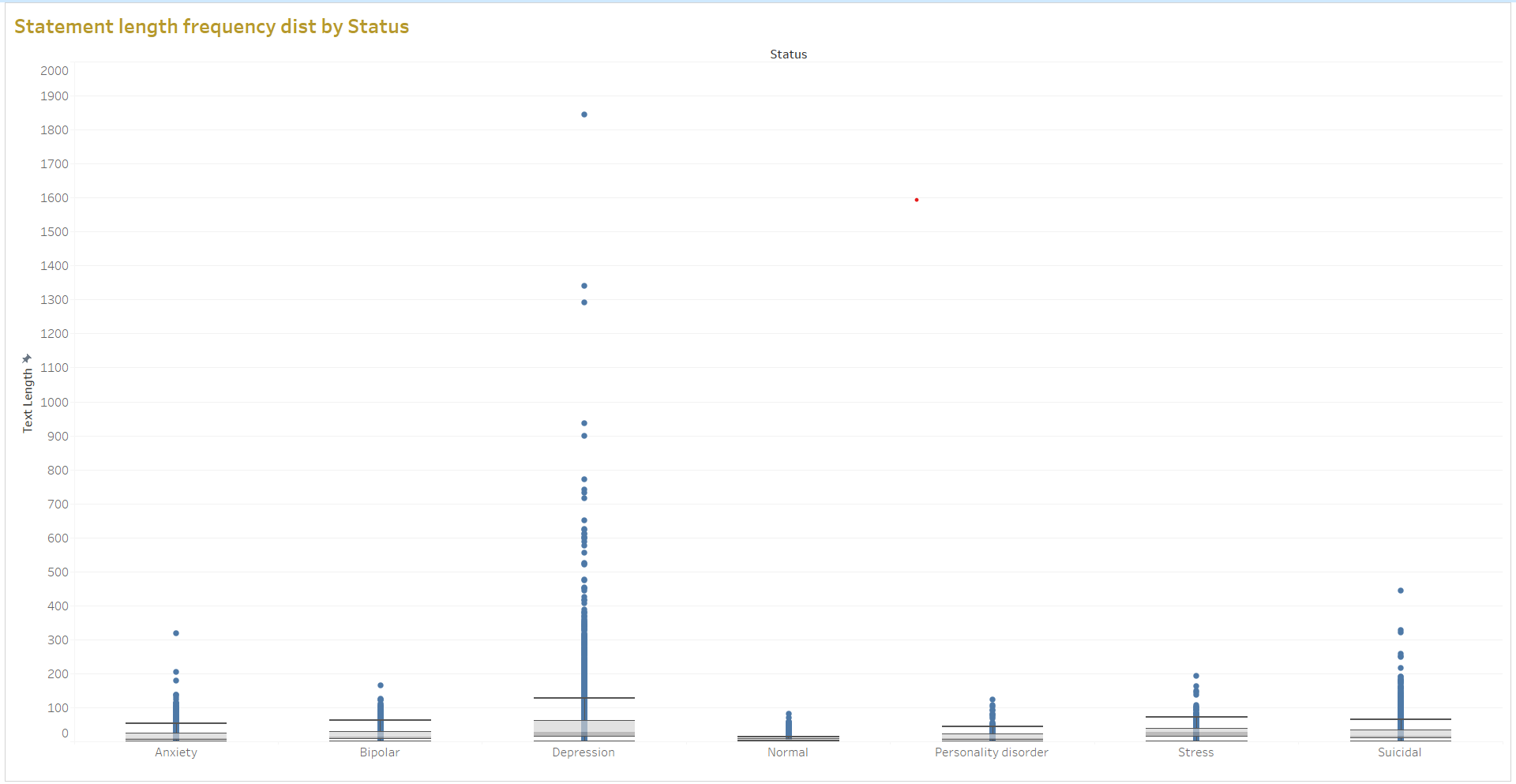
A word cloud was created to visualize the most frequently occurring words in different sentiment categories. Words such as "stress," "sadness," "happiness," and "fear" were prominently observed, providing insight into common expressions associated with mental health states.



### 

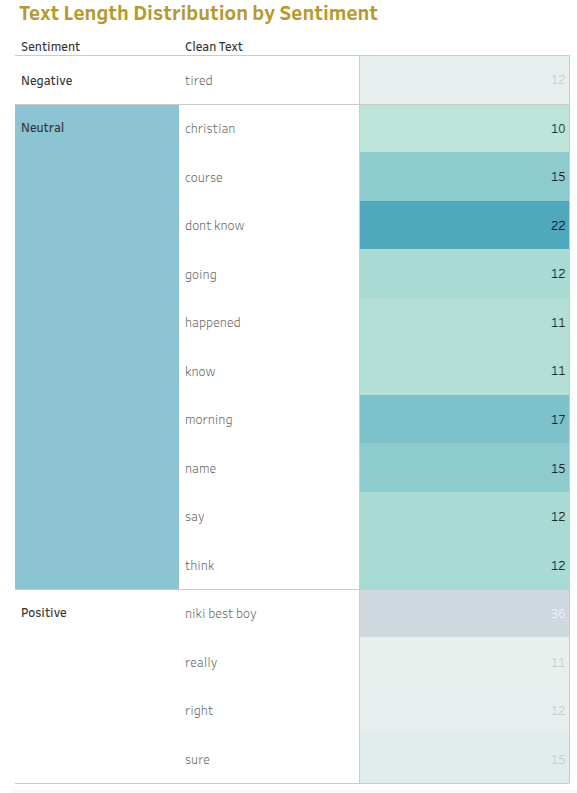
### 3. Statement Length Frequency Distribution by Status

A whisker plot was plotted to analyze the distribution of statement lengths across different sentiment categories. This visualization helped in understanding whether longer or shorter statements were more prevalent in specific sentiment classes.



### 4. Text Length Distribution by Sentiment

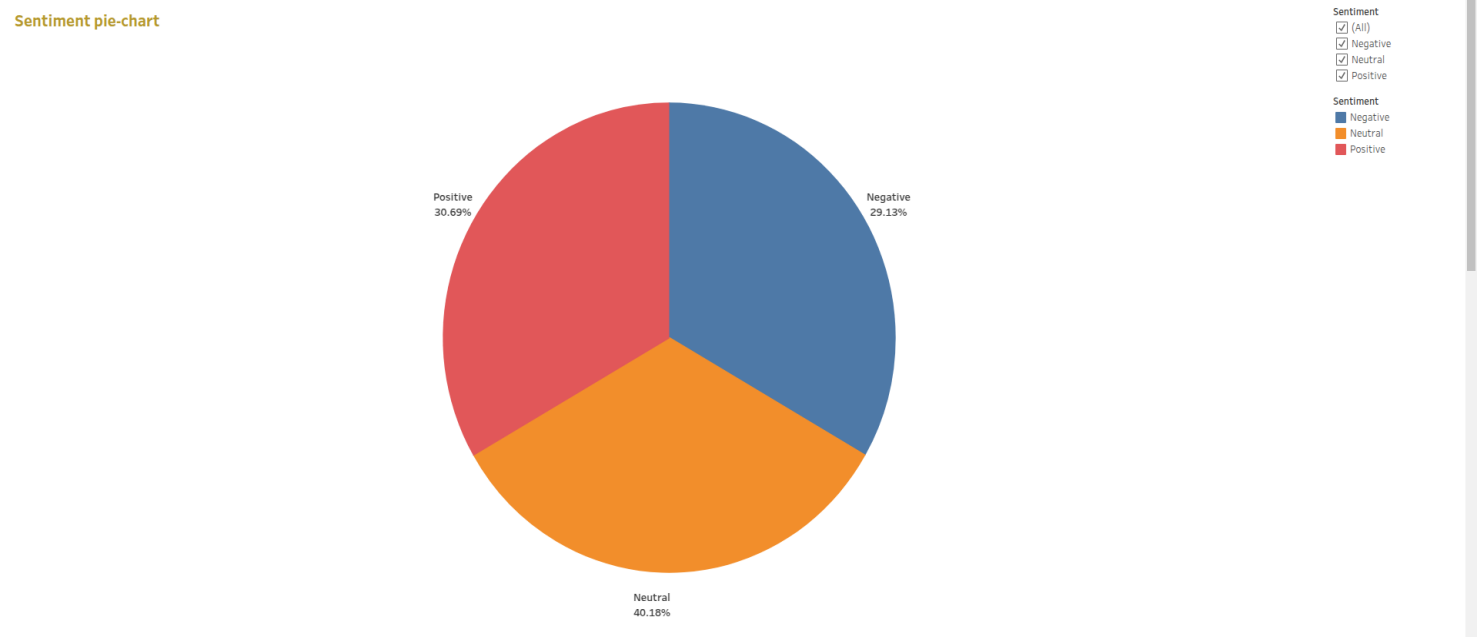
Text table summarizes the average, minimum, and maximum text lengths for each sentiment category. It shows that positive sentiments have longer text on average, while negative or neutral sentiments tend to be shorter. The distribution highlights variability, with some outliers in text length for each category. This helps identify how sentiment influences text verbosity in dataset.



### 5. Percentage Distribution of Sentiments

A pie chart displayed the proportion of each sentiment category in the dataset. This helped in understanding whether the dataset had an even distribution of labels or if one category dominated over others.

These visualizations provided valuable insights that were used to improve model performance and ensure better interpretability of results.



## 4. Results and discussion:

After training the BERT-based model, the results were evaluated using various metrics:

The confusion matrix indicates that the model successfully differentiates between different mental health categories, with minimal misclassification. The model shows strong generalization capability due to BERT’s contextual understanding.

**Classification Report :**

print(classification\_report(test\_labels,predicted\_labels,target\_names=label\_encoder.classes\_))

|  |
| --- |
| Class precision recall f1-score support |
|  |
| Anxiety 0.96 0.98 0.97 532 |
| Bipolar 0.97 0.99 0.98 559 |
| Depression 0.87 0.59 0.70 557 |
| Normal 0.93 0.92 0.92 610 |
| Personality disorder 0.99 1.00 1.00 581 |
| Stress 0.96 1.00 0.98 570 |
| Suicidal 0.73 0.90 0.81 598 |
|  |
| accuracy 0.91 4007 |
| macro avg 0.92 0.91 0.91 4007 |
| weighted avg 0.91 0.91 0.91 4007 |

**Confusion Matrix**

cm = confusion\_matrix(test\_labels, predicted\_labels)

plt.figure(figsize=(10, 7))

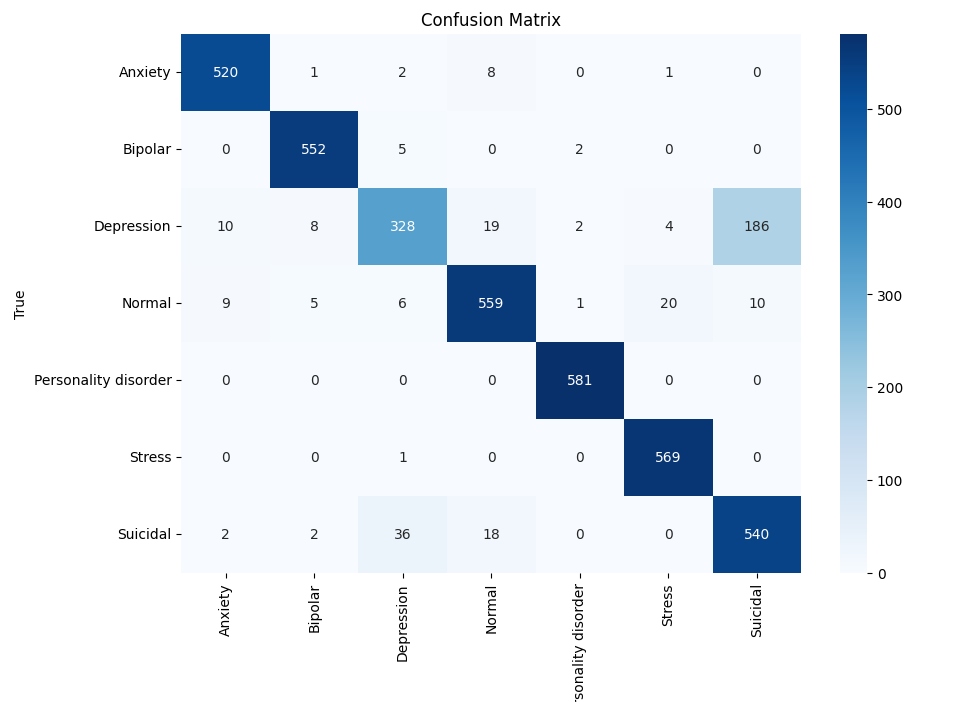
sns.heatmap(cm,annot=True,fmt='d',cmap='Blues',xticklabels=label\_encoder.classes\_, yticklabels=label\_encoder.classes\_)

plt.xlabel('Predicted')

plt.ylabel('True')

plt.title('Confusion Matrix')

plt.show()



## 5. GUI:

The Graphical User Interface (GUI) for this project was developed using Streamlit, a powerful and lightweight framework for building interactive web applications in Python.

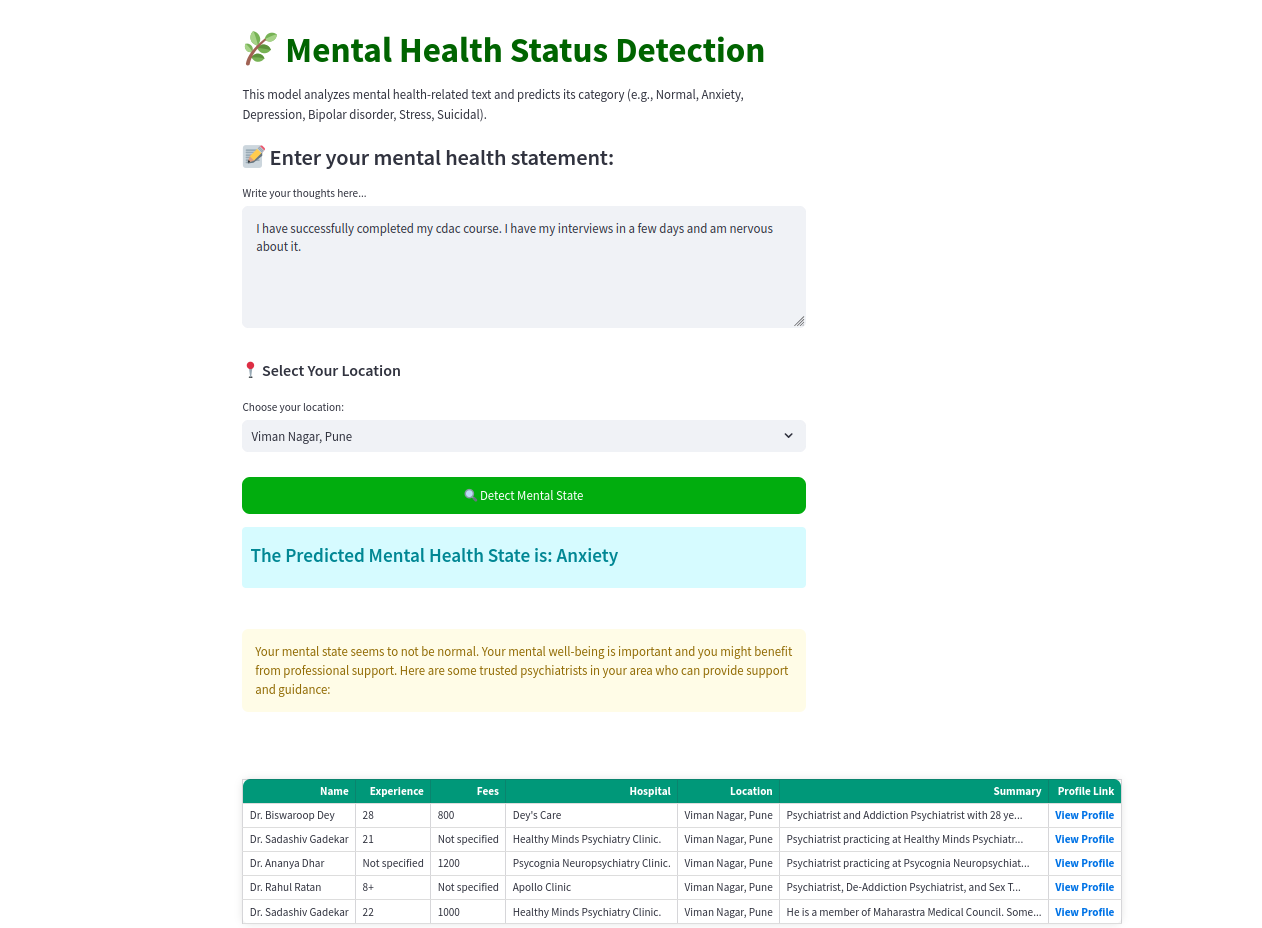
**Features of the GUI:**

User Input: Users can enter their mental health-related statements in a text box.

Real-time Prediction: Upon clicking the "Detect Mental State" button, the model processes the input and classifies it into one of the predefined mental health categories (e.g., Normal, Anxiety, Depression).

Interactive Design: The interface is intuitive, with a responsive layout that adapts to various screen sizes.

Deployment Ready: The GUI can be deployed on cloud platforms like AWS EC2, making it accessible from anywhere.



**Implementation Details:**

Framework Used: Streamlit

Model Integration: The BERT model, fine-tuned on mental health sentiment data, is loaded into the application for inference.

Preprocessing: Stopword removal, special character elimination, and tokenization are applied before model inference.

Deployment: The application is packaged and deployed on AWS EC2 for real-time user access.

This Streamlit-based GUI ensures a seamless and user-friendly experience for analyzing mental health sentiments from text input.

**6.GitHubLink:**

**ML Project** : <https://github.com/Sana-Hungund/Sana_CDAC_Project.git>

## BDA project : https://github.com/Sana-Hungund/BDA\_Project\_Repository.git

## 7.Future work And Conclusion 7.1 Future Work:

1. Model Improvement:
   * Experiment with more advanced transformer models (e.g., RoBERTa, DistilBERT) to improve classification accuracy and model robustness.
   * Implement domain-specific fine-tuning for better context understanding related to mental health.
   * Explore using multi-modal data (text + voice or text + image) to create a more comprehensive mental health detection system.
2. Model Deployment:
   * Deploy the model to a production environment (e.g., AWS Lambda, or a more scalable EC2 setup) to handle large-scale real-time mental health data and improve user accessibility.
   * Develop mobile applications or browser extensions to make the tool more widely available for users seeking instant mental health assessments.
3. Real-time Sentiment Analysis:

* Integrate real-time data processing for users to track their mental health state continuously through daily inputs, providing more personalized insights and recommendations.Add integration with social media platforms or forums to monitor the mental health state of users in larger-scale, real-time environments.

4. Personalized Mental Health Suggestions:

* Implement a recommendation system that offers suggestions for coping strategies, therapy options, or nearby mental health professionals based on the user’s predicted mental state.
* Develop a system that provides alerts or notifications for users who display increasing signs of anxiety or depression, prompting them to seek professional help.

## 7.2 Conclusion:

## The Mental Health Status Detection Tool has made significant progress in enabling real-time sentiment analysis to identify mental health states like anxiety, depression, and normal behavior from text. By leveraging BERT-based transformers, the tool achieves high accuracy in classifying text inputs into various mental health categories. The approach of cleaning the text, coupled with the oversampling of imbalanced classes, has led to a balanced and fair model performance.

* Key takeaways:

1. High Accuracy: The model achieves reliable predictions for identifying mental health states, providing users with insights based on their input text.
2. Scalability: The tool can be further enhanced to accommodate larger datasets and more diverse inputs, ensuring it remains useful across different platforms.
3. Practical Implications: By offering real-time detection, the project aids in identifying mental health conditions early, potentially offering timely interventions and recommendations for help.

The future directions for this project point toward continuous improvement in model performance, user accessibility, and personalizations, making it a valuable tool in the field of mental health monitoring and support.