

# Exploratory Data Analysis

## Early Detection of Mental Health Conditions on Twitter: An NLP-Based Approach

By Anvita Reddy, Pavan Kumar Settem

### Table of Contents:

<b>Objective of the Project:</b>	<b>1</b>
<b>Data Collection Process:</b>	<b>1</b>
<b>Data Filtering and Preprocessing:</b>	<b>3</b>
<b>Data Quality Using MarkovML Platform:</b>	<b>3</b>
1. Data Split	3
2. Duplicate Rows	4
3. The Term Frequency Bar Chart :	5
4. Handling Missing Values:	6
5. Word Clouds Generated:	7
6. Keywords Analysis RAKE:	8
7. Sentiment Analysis:	9
<b>Data Relevance:</b>	<b>11</b>
<b>Data Size:</b>	<b>11</b>
<b>Ethical Considerations</b>	<b>11</b>

# Exploratory Data Analysis:

## Objective of the Project:

The objective of the project is to develop an NLP-based system that utilizes Twitter data to analyze language patterns, sentiment, and contextual clues in tweets, with the aim of identifying individuals with mental health conditions such as depression, anxiety, or stress. The project focuses on early identification of individuals at risk.

## Data Collection Process:

Initially planned to Scrape data using the Twitter's official API, but it was discontinued<sup>[1]</sup> So instead we looked for resources online and came across a few:

### 1. Depression Detection Using Twitter Data

<https://github.com/swcwang/depression-detection>

Collection of 3200 rows consisting of:

- 2357 tweets belonging to depressive hashtags like #depressed #depression #loneliness #hopelessness<sup>1</sup>.
- 843 of general tweets contain other emotions, such as joy, love, surprise, as well as some emotionally neutral tweets.

### 2. Detecting Depression in Tweets

<https://github.com/viritaromero/Detecting-Depression-in-Tweets>

Collection of 10314 tweets consisting of:

- 8000 positive(no-depressive) tweets from Sentiment140 dataset with polarity as 4, [Sentiment140\[2\]](#)
- 2314 tweets from scraped from Twitter

### 3. Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study

[https://github.com/aryan-r22/Stress-Detection\\_Social-Media-Articles](https://github.com/aryan-r22/Stress-Detection_Social-Media-Articles)

Collection of 2051 tweets:

- A total of 1268 tweets not related to mental health were collection with as a part of covid dataset collection and a sentiment analysis was run, where tweets under categories of 'joy' were selected
- 783 tweets using SVM and Naive Bayes to get tweets related to mental health

# Data Filtering and Preprocessing:

As part of the Pre-processing, we performed the following steps:

## Step 1:

- Removing Emoticons
- Removing URLs
- Removing unicode Characters
- Removing hashtag and mentions
- Removing contractions and stopwords
- Removing special characters
- Conversion to lowercase

Original Tweet	Cleaned Tweet
hi @JohnDoe ! I would suggest music of my electro project *** <a href="http://bit.ly/12KoF0">http://bit.ly/12KoF0</a> *** free download & have fun cheers	hi suggest music electro project free download fun cheers

**All these conversions help in maintaining the data uniformity and noise reduction, therefore making the relevant data available for the model to make meaningful patterns from**

## Step 2:

### Considering only English Tweets

Storing only English texts ensures a consistent language framework, simplifying the data preprocessing and model training processes, while also allowing to leverage existing English-language resources and pre-trained models, optimizing performance and efficiency.

## Step 3:

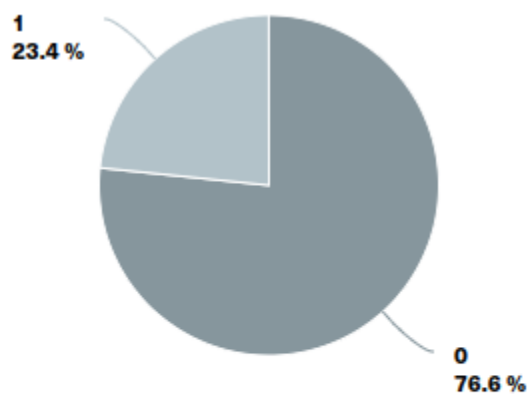
### Balancing all Datasets Sourced

Storing equal tweets belonging to both classes in a binary classification project ensures a balanced dataset, preventing class imbalance issues during model training, and helps the model learn from an equitable representation of each class, leading to more accurate and unbiased predictions.

# Data Quality Using MarkovML Platform:

## 1. Data Split

We had earlier uploaded our previous Data onto MarkovML which showed:



Classes

● 0 ● 1

Which now looks like an almost equal split:



Classes

● 0 ● 1

## 2. Duplicate Rows

Initially the Data had a lot of duplicates which are now erased to ensure the authenticity in the data

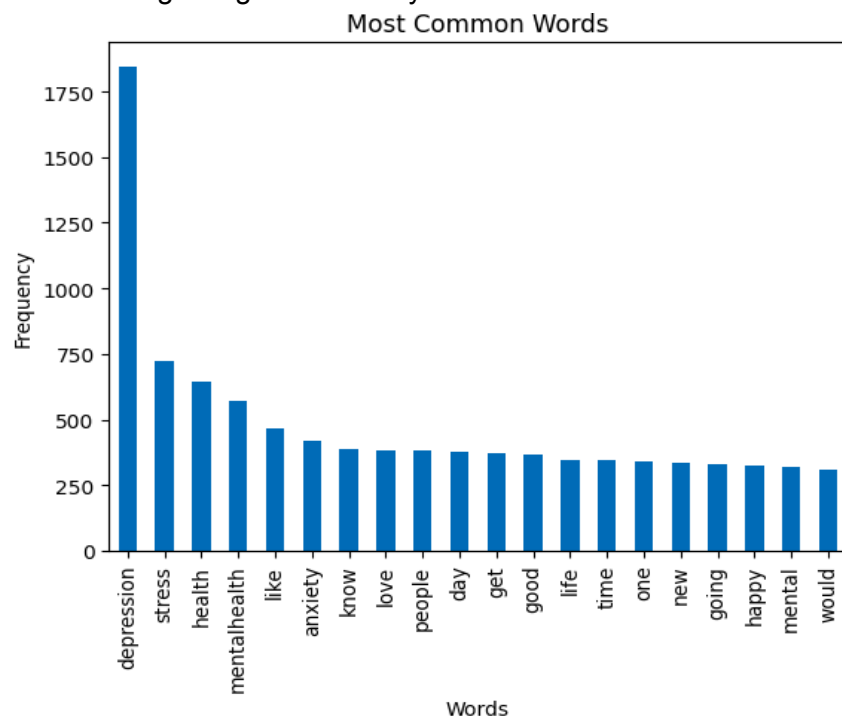
Data summary	
Num Features	2
Num Rows	7,876
Missing Cells	0
Percent Missing Cell	0
Dup Rows	142

**AFTER CLEANING:**

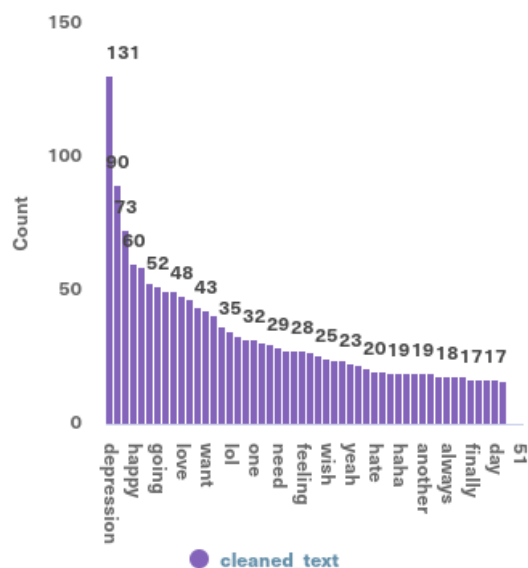
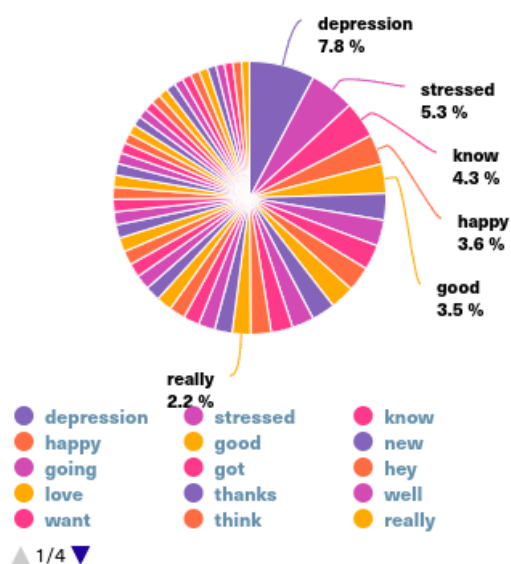
Data summary	
Num Features	2
Num Rows	6,588
Missing Cells	0
Percent Missing Cell	0
Dup Rows	0

### 3. The Term Frequency Bar Chart :

It provides insights into the most frequent words or terms present in the text data, helping to identify key trends, popular topics, and potential anomalies, which aids in understanding the dataset and guiding further analysis.



From MarkovML:



#### 4. Handling Missing Values:

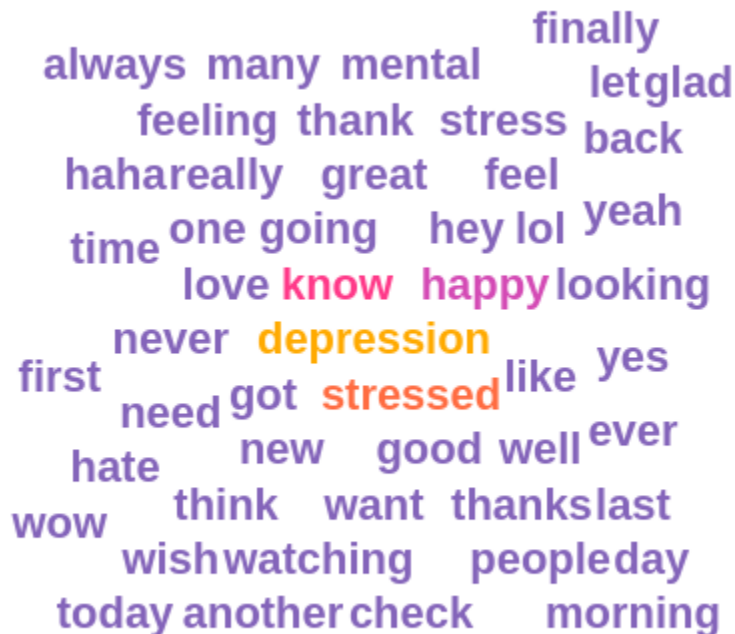
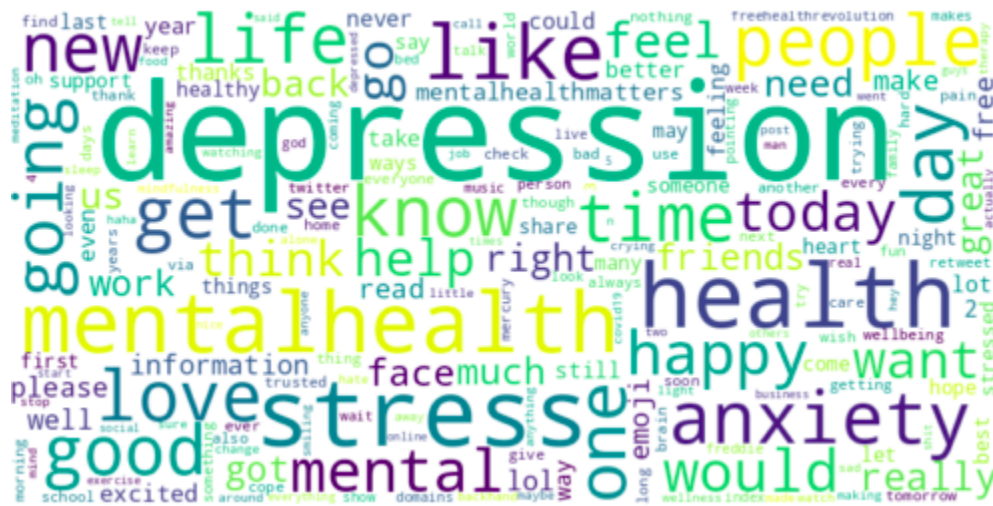
We made sure to manually categorize tweets for which a label was missing because it allows for a comprehensive examination of the dataset without the need for imputation or dealing with data gaps, leading to more reliable and accurate insights.



#### 5. Word Clouds Generated:

Word clouds are helpful in this context as they visually represent the most frequently occurring words in mental health-related tweets, allowing for quick identification of prevalent topics and themes.

## Word Cloud



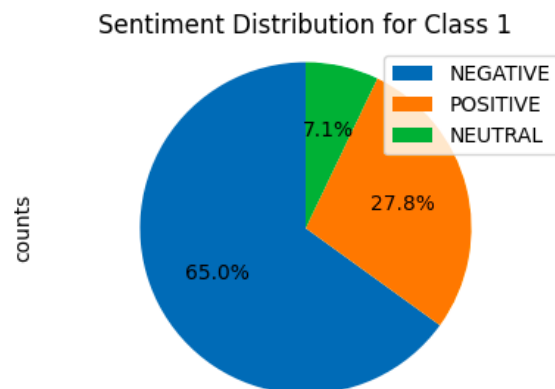
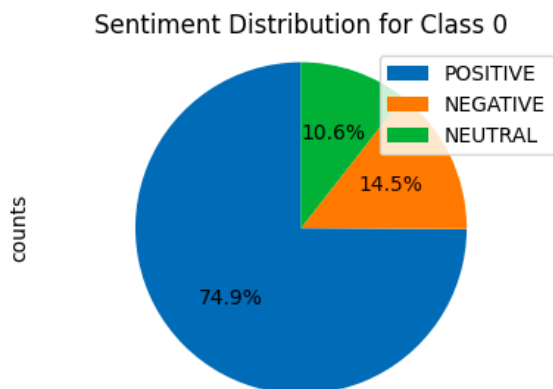
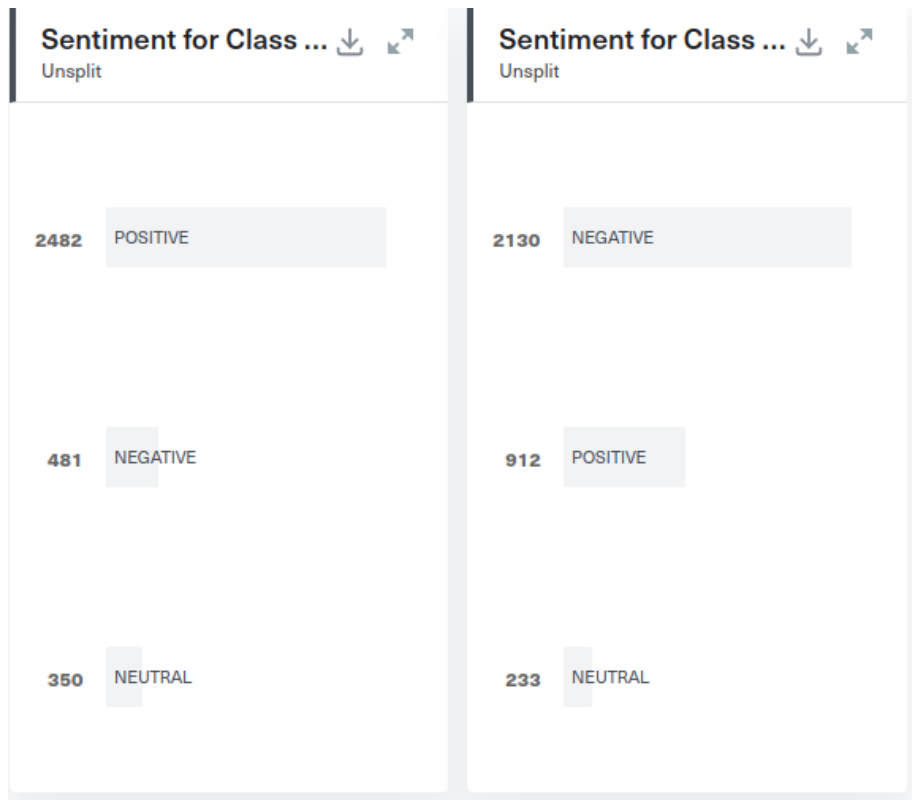
## 6. Keywords Analysis RAKE:

RAKE (Rapid Automatic Keyword Extraction) is valuable for mental health tweet analysis as it identifies essential phrases, helps understand main topics, reduces data complexity,



Sentiment analysis is useful in the context of mental health tweet analysis as it helps determine the emotional tone expressed in the tweets, enabling the identification of sentiments related to


mental health issues, such as stress, anxiety, or depression, providing valuable insights into the overall emotional well-being of users and the prevalence of different sentiments in mental health discussions. The Visualization looks as one would expect



## Data Relevance:

The exploratory data analysis (EDA) conducted on the dataset showcases the prevalence of mental health-related keywords, sentiments, and key phrases, confirming its relevance for the given context of analyzing mental health disorders using tweets and providing valuable insights for further analysis and model training.

## Data Size:

cleaned_tweets_team_2	Type	Rows	Segments	 LABEL TRUST ESTIMATE ⓘ
	Text	6,588	Not split	

The size of the dataset demonstrates that the size is substantial enough to capture a diverse range of mental health discussions on social media, justifying its relevance for the given context of analyzing mental health disorders using tweets and ensuring sufficient data for meaningful insights and model training.

## Ethical Considerations

To ensure privacy and protect user identities in the tweets, hashtags and mentions were removed through a preprocessing step during data cleaning. This involved identifying and extracting hashtags (words or phrases preceded by the '#' symbol) and mentions (usernames preceded by the '@' symbol) from the tweet texts, and subsequently deleting them from the data. By removing hashtags and mentions, the sensitive information related to specific users or topics is masked, preserving user anonymity and complying with privacy guidelines.