

# Stress Detection from social media posts using DistilBERT

Guide: Anesu Nyabadza

Gayathri Chava

# Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
2.1	Research Question . . . . .	3
2.2	Problem Definition . . . . .	3
2.3	Project relevance in the field of Data Analytics . . . . .	4
2.4	Procedure of Ethical Approval . . . . .	4
<b>3</b>	<b>Literature Review</b>	<b>4</b>
3.1	BERT . . . . .	5
3.2	DistilBERT . . . . .	6
3.3	Applications . . . . .	6
3.4	Ethics . . . . .	7
<b>4</b>	<b>Data Preparation and Exploration</b>	<b>8</b>
4.1	Introduction . . . . .	8
4.2	Description of Data . . . . .	8
4.3	Data Cleaning . . . . .	9
4.4	Data Preprocessing . . . . .	10
4.5	Data Visualization . . . . .	10
4.5.1	Distribution of Classes . . . . .	10
4.5.2	Text Length Analysis . . . . .	12
4.5.3	Word Frequency Visualisation . . . . .	15
4.5.4	Text Length Analysis: Stressed and Non-Stressed . . . . .	17
<b>5</b>	<b>Platform, Language, and Architecture</b>	<b>20</b>
<b>6</b>	<b>Results and Discussion</b>	<b>20</b>
<b>7</b>	<b>Conclusion and future work</b>	<b>23</b>

# 1 Abstract

Social media has already become a part of everyone's life these days. People are very much interested in sharing their views, feelings, experiences, and achievements irrespective of their situation. Because of the huge amount of data available on social media, which would be helpful for the researchers to explore this data and make some realistic models which would be very helpful to the public [30]. Stress is becoming a common thing that everyone one is feeling these days. People are sharing their feelings in social media platforms; from this we can get the data and build models which can tell whether the person is feeling stressed or not based on the post they posted on the social media platforms. Data we used in this report were from the publicly available datasets, which were taken from the platforms like Reddit and Twitter. Pre-Trained language model called DistilBERT was used which was 40% smaller, and 60% faster than the BERT with 97% of the results. In this report first we explained about the methodology we used to build this model, and then we explained the relevance tree and data description and how we handled the data. Finally, we discussed the results and the future work. The evaluation metrics we chose for this project are the accuracy and the F1-score. Though accuracy gives the overall correctness of the model, we want to minimize the false alarms and at the same time model is supposed to detect the stressed ones so we choose F1-score as the evaluation metric. Future work includes of dealing with vast amount of data and images and videos so that we can get the deeper understanding of the data and building models with all this data, where health professionals can use those models to help in the early detection of stress and improve mental health outcomes.

**Keywords:** Social media, DistilBERT and stress.

# 2 Introduction

In today's world, social media is becoming a very important part of everyday life. People are using platforms like Reddit and Twitter to share their feelings and emotions with everyone in these platforms. People are interested in their experiences, achievements and sometimes their problems as well without even noticing it. This data can give so much information to the researchers like what they are feeling and how they are experiencing

their lives, which researchers can use to detect the mental state of the person, like whether the person is feeling stressed or not based on the content they posted. Stress is a common problem in everyone's life. Around 65% people did not receive help when they are dealing with these kind of problems [26]. Everyone feels stress sometimes, but some people stress a lot and that will stay long-term as well in some cases, which can affect their mental health, and if this continues, it will also affect their physical health. High stress can lead to trouble sleeping, depression, and sometimes heart issues as well. It is very important to detect stress early and help the affected people before it causes any harm to their lives. Traditionally, to detect or to study stress, researchers used to do surveys and interviews. These methods are very helpful to researchers, but they took lot of time and effort. The other thing we can't be sure whether the data we are getting is real or not, because people can also add false data to those surveys. With the improvement of social media apps and its availability people are posting their feelings in these platforms, where researchers can use this data to detect the mental state of the people without asking them. Natural language processing helps in identifying the text patterns, sentiments in the tweets and posts and identifying the stress related posts. The aim of this dissertation is to detect whether the person is feeling stressed or not based on the posts and tweets they posted on platforms like Reddit and Twitter, which would be helpful for the professionals in health care and the policy makers so that to provide early assistance and propose some new plans to reduce stress in people. By identifying and responding to the effected people helps them to become normal and also reduces the load to the healthcare system .

## 2.1 Research Question

Can we use NLP model like DistilBERT model to detect stress?

## 2.2 Problem Definition

Stress is a feeling that anyone can feel when they don't meet their expectations, which they defined for themselves, irrespective of age and gender[27]. It can affect people mental and physical health, which sometimes leads to serious health problems and affects their productivity. Researchers conducting interviews and surveys to detect stress is the traditional method, which is a time taking and costly process. These days platform like Reddit and Twitter are providing a huge amount of data. Though we have the data

to conduct research, the data we found on these platforms is unstructured and informal which would be a challenging aspect for the traditional machine learning models. This research mainly focuses on using the pre-trained language models like DistilBERT model which can capture the difficult patterns in the text which we cannot see in machine learning models. Therefore, the main goal in this research is to develop a model using DistilBERT model which can predict the stressed posts and at the same time reducing the false alarms.

### **2.3 Project relevance in the field of Data Analytics**

In the current situation there are many problems that can affect the quality of life, in those stress is considered as a key factor in both the psychological and physiological health problems [29]. Social media generates huge amount of unstructured text in the form of posts and tweets every day. Detecting whether the person is feeling stressed or not based on the post they post on platforms like Reddit and Twitter is a real-world problem. This project is very relevant in the field of data analytics because it shows how we can solve real world problems with the help of advanced technologies. We cannot analyse the data generated from social media manually. But with the pre-trained language models like DistilBERT model we can find the patterns in the text data, and predict whether the person is feeling stressed or not. This research shows how data analytics is useful by converting the unfiltered data into useful insights. This can be used to understand human behaviour and support mental health and solving the real-world problems.

### **2.4 Procedure of Ethical Approval**

Ethical approval form was submitted mentioning about the data where I got that data and it was approved.

## **3 Literature Review**

In the recent years we can see the impact of transfer learning approaches in Natural Language Processing (NLP), where pre-trained language models are becoming an important tool in many NLP tasks. Many methods have been developed for detecting stress from the textual data as the input for the model [17]. Bidirectional Encoder Representations

from Transformers (BERT) is one of the pre-trained language models [11]. From the name we can say that this approach captures the context of the text from both the sides. The main problem with the BERT model is they require large computational resources which can be problem while making the real-time models. To overcome the limitations of the BERT model, a model was developed by HuggingFace which is known as DistilBERT which is smaller and lighter version of BERT and provide same results. DistilBERT is 40% smaller, 60% faster than BERT models and maintains the 97% of the results. Because of this advantages it is becoming an important choice while building models for real-time applications.

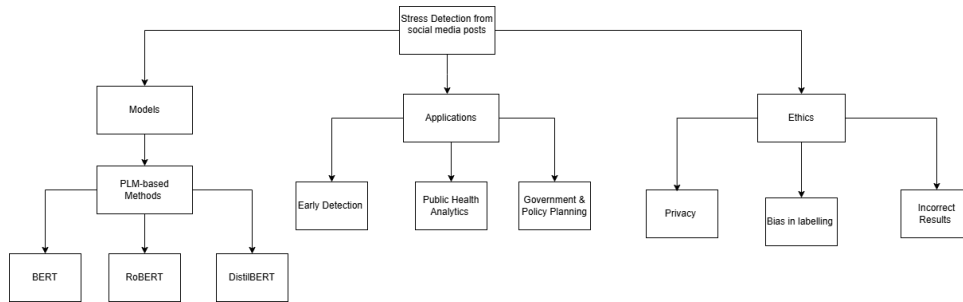


Figure 1: Relevance Tree outlining the literature process

### 3.1 BERT

In 2018 Google developed a transformer-based model called as Bidirectional Encode Representations from Transformers[4]. Compared to the traditional methods, BERT has the capability to read the text in both directions, which makes that the model can understand better [28]. This model uses multi-layer bidirectional transformer encoder. BERT has 12 layers, 768 hidden units and 12 attention heads. This model contains 110 million parameters[3]. BERT model is trained based on the two methods. First one is the masked language modeling which means some words are masked randomly and then the model is trained to predict the hidden words based on the context [22]. The second one is next sentence prediction which helps the model to predict the future sentence based on the previous one [12]. Though all these we have few disadvantages of the BERT model like the computational cost is expensive which makes it less suitable for real world models [5]. Along with that using pretrained models like BERT in transfer learning does not provide the great performance consistently [20].

### 3.2 DistilBERT

DistilBERT maintains the core architecture of the BERT model, but with alternatives like reducing the size and computational requirements. Knowledge distillation is the compression technique which is the core technique behind DistilBERT. This technique trains a smaller (student) model by replicating the knowledge from the large model (teacher), whereas in our case teacher is the BERT model which is large and pre-trained model and the DistilBERT is the student model.

**Model Size:** BERT has 12 layers, where as DistilBERT has 6 layers and the parameters are reduced from 110 million to 66 million [23].

**Training:** DistilBERT is trained using three methods, they are masked language modeling, distillation loss and cosine-distance loss.

**Masked Language Modeling:** BERT model also uses the same modeling. Masked language modeling means certain words in a sentence are randomly masked and the DistilBERT models learns to predict those masked words based on the surrounding words and its context. Like BERT model DistilBERT models learns the language patterns by the masked language modeling [14].

**Distillation loss:** Distillation loss is calculated as the difference between the probability distribution of teacher and the probability distribution of the student [10].

**Cosine-distance loss:** The model is trained to generate the outputs which are masked and which are similar to the teacher model (BERT), to make the similar outputs [16].

#### Advantages of DistilBERT

- The smaller size than BERT makes the DistilBERT more efficient and using less memory resources compared to BERT model [23].
- Achieving the best results like BERT, irrespective of its smaller size make it more popular and suitable for building the real time models.

### 3.3 Applications

In the online world, stress is the common experience that every human being feels [27]. Early detection can helps us in so many ways, as well as in this case early stress detection

will be very helpful, so people won't take any wrong decisions that can take their lives. Researchers building the models which can detect whether the person is feeling stressed or not will be very helpful to the health professionals. With the help of these models doctors can predict the stress levels faced by their patients and treat them as early as possible before it becomes worse. With the correct models people can gain the trust and take the help of those models and also take necessary precautions to overcome stress. In previous days, the scenario is completely different to the current situation. Before, researchers need to spend lot of time in gathering the data. They are supposed to do surveys, interviews to get the data which is very time taking process. Because of the usage of social media these days, we can get lots of data from these platforms. With the help of this researchers can really build the better models, so with that they can help the government to plan campaigns, and to encourage open conversations.

### 3.4 Ethics

Social media is providing large amount of data for the researchers who wants to build the models that can predict whether the person is feeling stressed or not based on the posts they made on the platforms like Reddit and Twitter. During the covid-19 period social media became more famous, where the entire world is shut down, and people started sharing their emotions, daily tasks and everything in the social media and make that as the huge platform for the data. With this we can have the large amount of data from platforms like Reddit and twitter, which gave the researchers to train their models with different types of data. We all know that every coins has two sides, like that though we are getting large data, we need to very careful while dealing with this data. We need to anonymize the data before we use that for training the model. Models are not supposed to tell whether the person is feeling stressed or not based on their location and their name, so we need to anonymize the data. If the models is predicting like that then we need to consider that the model is biased. While training the model we need to very careful and check whether the data is biased or not. We all know that the biased data will lead to incorrect results. Let's say for example if the data is biased based on length then there will be problem when the normal data is feed to the model. If the model decides whether the person is feeling stressed or happy based on the post length they post on social media platforms, then it will give us the incorrect results. Because the length of post may vary



from person to person, like some people share in small length for both the good and bad situations and some people share everything though it is stressful or happy. So we need to be very careful when the data is biased. If the data is biased then we need to monitor the model's performance regularly to avoid biased results.

## 4 Data Preparation and Exploration

### 4.1 Introduction

Data plays a key role in every research. For my research which is finding whether the people are experiencing stress or they are posting happily based on the posts and tweets they posted on social media platforms like Reddit and Twitter I am using the publicly available datasets which were created by Aryan Rastogi [19].

Four datasets are used in this paper, which are Reddit Title, Reddit Combo, Twitter Full and Twitter Non-Advert. In the four datasets each text entry is labelled as 0 or 1, where 0 represents a stress negative text and 1 represents a stress positive text.

### 4.2 Description of Data

The description of the datasets is given as under:

**Reddit Title:** This dataset consists of two columns, which are title and label. Title column contains the titles from both the stress and non-stress related subreddits from Reddit. Label column represents the binary classification of the column which shows whether the title is indicating stress or not.

**Reddit Combo:** This dataset has 4 columns, which are title, body, body title and label. Title column represents the titles from both the stress and non-stress subreddits from Reddit. Body column contains the message shared by the user in the Reddit. Body title column is the combination of both the title and body column. Label is a binary classification column which indicates whether the post related to stress or non-stress.

**Twitter Full:** This dataset has 3 columns, which are text, hashtags and labels. Text column contains the text which they posted in Twitter. Hashtags column contains the hashtags used by the users for the tweets they made in Twitter. Labels column contains values like 0 or 1 where 0 represents non-stress tweet and 1 represents stress related tweet.

**Twitter Non-Advert:** This dataset is the cleaner and improved version of the Twitter full dataset. This dataset has 2 columns, which are text and label. Text column consists tweets of stress and non-stress related tweets. Label column contains the label for the text column with the values of 0 or 1, where 0 represents non-stress tweet and 1 represents the stress related tweet.

### 4.3 Data Cleaning

The data I used for this project is taken from the social media platforms like Reddit and Twitter. In these platforms people often share their feelings, experiences and their adventures in normal text which often includes so many punctuations, numbers and so many common words. In real life it is okay for having all these in our posts, but when we are dealing with the tasks like natural language processing, we need to be very careful with these kinds of data, because they can affect the analysis. So, I explained everything in detail about I handled these types of data below.

**Handling Duplicates and Missing Data in Reddit Title Dataset:** In Reddit title dataset there is no null values when I loaded the dataset, which I got from the publicly available repository. 24 duplicates were found in this dataset. I dropped those duplicate values.

**Handling Duplicates and Missing Data in Reddit Combo Dataset:** In Reddit combo dataset we have four columns, which are title, body, body title and label. In these four columns I found 7 null values in the body column. So instead of filling up some value in those I removed them, because I don't want anything to influence the model. I checked for the duplicate values, and there are no duplicate values in this dataset.

**Handling Duplicates and Missing Data in Twitter Full Dataset:** In Twitter full dataset we have three columns which are text, hashtags and labels. I found 8 null values in the hashtags column and I removed them. I found 373 duplicate values which I removed them, before they bias the model by over representing them. Removal of null and duplicate values make the dataset clean and unique.

**Handling Duplicates and Missing Data in Twitter Non-Advert Dataset:** In Twitter Non-advert dataset, we have two columns which are text and label. As I mentioned earlier, this dataset is the cleaner version of the twitter full dataset, it doesn't have any null values. But it has 79 duplicate values, which I removed them to make the

dataset unique.

## 4.4 Data Preprocessing

After dealing with the missing and duplicate values, we need to focus on the text data in more depth now, as we are doing the natural language processing tasks. For that first I converted the text in the title column into lowercase as part of the text normalization. After that I found there are some numerical values in the dataset. As we are dealing with detecting whether the person is feeling stressed or not based on the social media posts they post on platforms like Reddit and Twitter, our main goal is the text data, where numbers are not important in this. So, removing those numbers makes the data clean and prevent overfitting, and our model mainly focus on the text data. So, I used regular expression to remove numbers from the dataset.

The data we are using for our model is from the social media, which explains there will be some punctuations in the data, which is very common. If we observe carefully, these punctuations don't hold any meaning. We need to remove these punctuations as well for better model performance and to prevent overfitting. After the removal of punctuations and numbers, now the main goal is tokenization of the data. For the tokenization I use the word tokenize method which is a part of the Natural Language Toolkit (NLTK) library. This method divides the sentence into words or tokens. Stopwords are the words that occur very frequently in a sentence, but represents the less meaning. So, I removed the stopwords from the dataset.

## 4.5 Data Visualization

### 4.5.1 Distribution of Classes

By using the seaborn's countplot, class distributions are plotted for all the datasets. Bar plot is a graphical representation which is used to compare different categories, which would allow the user to identify the differences among them [6]. For the Reddit title dataset, from the Figure 2 we can see that both the stressed and non-stressed posts are almost equal which shows that there will be less chances for the biased predictions.

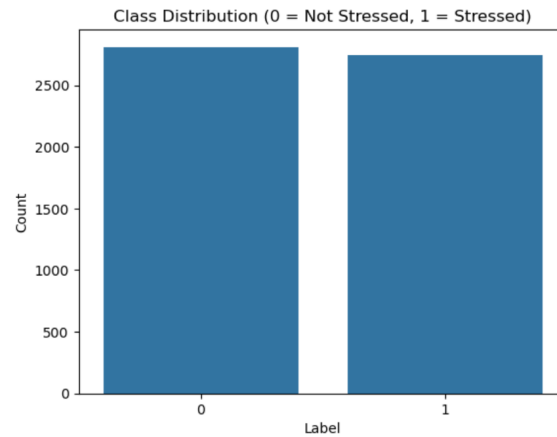


Figure 2: Distribution of reddit titles by class

Bar plot is a graphical representation which is used to compare different categories, which would allow the user to identify the differences among them [6]. In twitter full dataset, from the Figure 3 we can see that this dataset is the well-balanced dataset with 50.9% as non-stressed tweets and the 49.1% as stressed tweets. With the balanced dataset, the model has equal number of tweets from both the classes, which would be helpful in better predictions and less chances of the biased results.

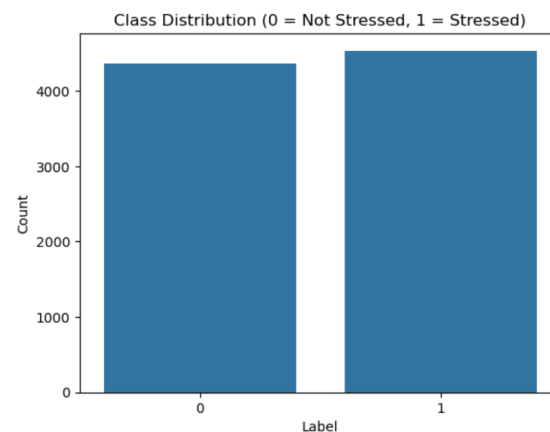


Figure 3: Distribution of twitter posts by class for twitter full dataset

Bar plot is a graphical representation which is used to compare different categories, which would allow the user to identify the differences among them [6]. In Twitter Non-advert dataset, from Figure 4 we can see that we have 62% of non-stressed tweets, and 38% of stressed tweets, which shows that there are a greater number of non-stress tweets.

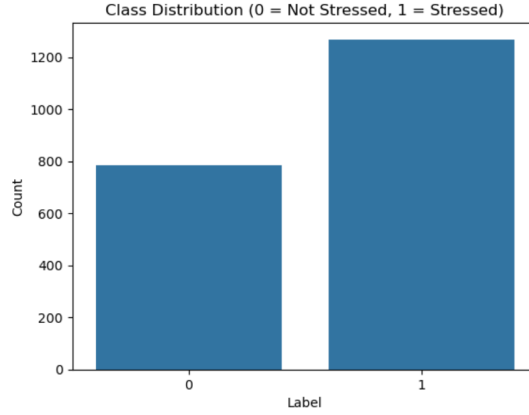


Figure 4: Distribution of twitter posts by class for twitter non-advert dataset

#### 4.5.2 Text Length Analysis

In our Reddit title dataset, we have title column as the text, so I used the len function to found the length of the titles, in our dataset. We used histogram which is a graphical representation that helps in visualizing the data distribution over a period, which will be more helpful in understanding the data and to understand the trends and the patterns of the data [8]. The histogram, Figure 5 is the title length distribution of the Reddit posts. The x-axis shows the number of characters ranging from 0 to 350. The y-axis shows the count of titles. Histogram shows that the most of the titles are in the range of 50-100 characters. From the distribution we can say that is the right-skewed distribution. Short titles are common and rarely we can see the posts with the long titles.

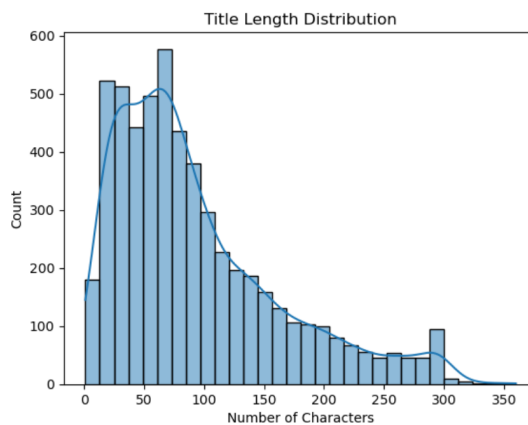


Figure 5: Title Length Distribution for Reddit titles

For the Reddit Combo dataset, we can see the two histograms. We used histogram which is a graphical representation that helps in visualizing the data distribution over

a period, which will be more helpful in understanding the data and to understand the trends and the patterns of the data [8]. The first histogram, Figure 6 is the title length distribution of the Reddit posts. The x-axis shows the number of characters ranging from 0 to 300. The y-axis shows the count of titles. We can see, the number of characters for the title they post on the Reddit is in between 30-60. The histogram for the title column is the right skewed distribution, which says that mostly they are less titles and rarely we can see the large titles.

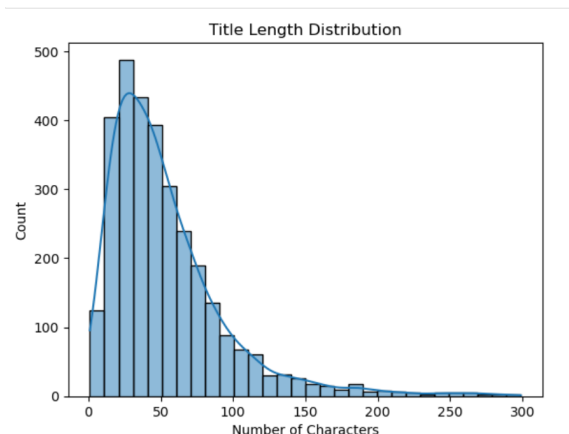


Figure 6: Title length distribution for Reddit posts

The second histogram, Figure 7, is the body length distribution of the Reddit posts. The x-axis shows the number of characters ranging from 0 to 30000. The y-axis shows the count of bodies. In Reddit, the post will have one title and the body. Title briefly explains, the main thing is expressed in the body. So, in general the length of the body will be more than the title length. From the histogram we can say that this is the right-skewed distribution. The average length of the body is 861 characters.

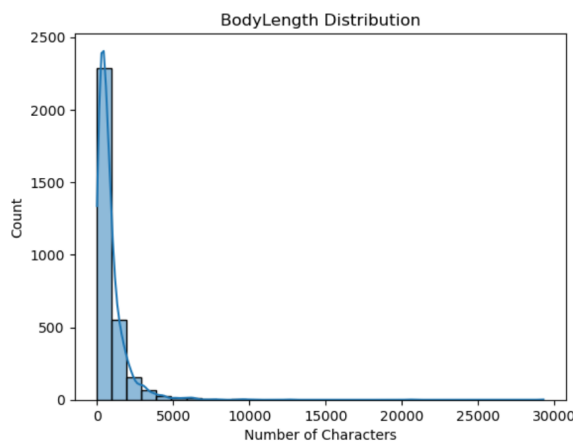


Figure 7: Body length distribution for Reddit posts

The Figure 8, is the text length distribution of the Twitter tweets. We used histogram which is a graphical representation that helps in visualizing the data distribution over a period, which will be more helpful in understanding the data and to understand the trends and the patterns of the data [8]. The x-axis shows the number of characters ranging from 0 to 1000. The y-axis shows the count of texts. The average length of the title in twitter is 181 characters. From the histogram we can see that the distribution is the right skewed, which means that most of the title have the characters in the range between 150-300, and after that the we can see the drop in the frequency. From this we can know that there are only few titles which are more characters, and in general we can say that we see most titles with less characters and only a few titles with more characters.

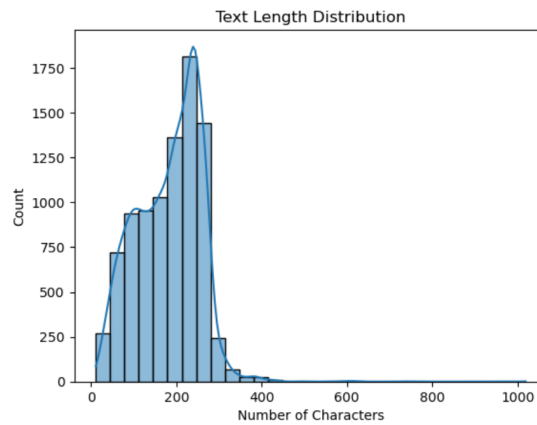


Figure 8: Text length distribution for the Twitter tweets

The below Figure 9, is the histogram, for the text column in the Twitter Non-advert dataset. We used histogram which is a graphical representation that helps in visualizing the data distribution over a period, which will be more helpful in understanding the data and to understand the trends and the patterns of the data [8]. The x-axis shows the number of characters ranging from 0 to 400. The y-axis shows the count of texts. We can see the distribution is the right-skewed one. The average title length for the twitter tweets is 177 characters. From the histogram we can say there are a greater number of texts with the 250 characters. The number of texts with more than 250 characters is gradually decreasing which indicates that, in the twitter we have less lengthy tweets.

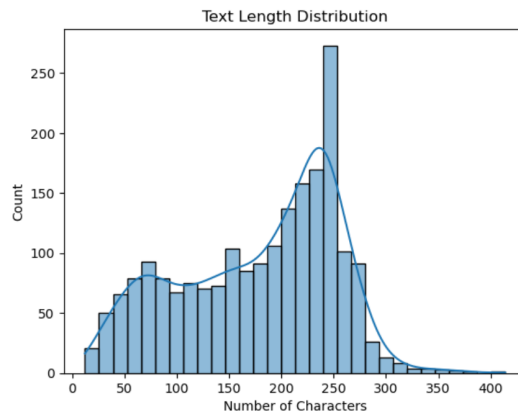


Figure 9: Text length distribution for Twitter Non-advert dataset

#### 4.5.3 Word Frequency Visualisation

When we are dealing with the text data, wordcloud is the one of the visualization techniques. Wordcloud shows the most frequent words in the given text, where the most common word is bigger in size [7]. For our dissertation we are using four datasets and I generated the wordclouds for both the stressed and non-stressed posts in all the datasets. From the below Figure 10 & 11, we can see that words like stress and depression are occurring most frequently in both the Reddit title and Reddit combo stressed posts.



Figure 10: Wordcloud for stressed posts    Figure 11: Wordcloud for stressed posts

Wordcloud shows the most frequent words in the given text, where the most common word is bigger in size [7]. Where as in the non-stressed posts from both these datasets we can words like happy, today and life are repeating more stating that people are sharing their happiness and their positive life which we can see from the below Figures 12 & 13.







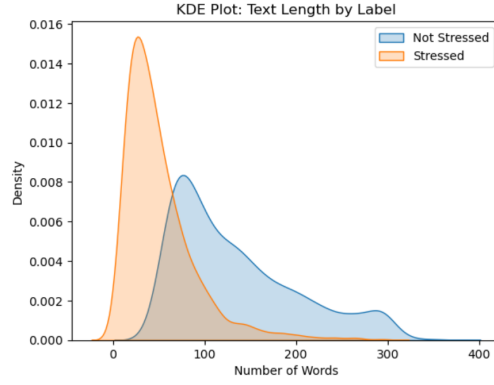


Figure 18: Distribution of the title in the Reddit title dataset

This KDE plot, Figure 19, explains the distribution of the body of the Reddit posts from the Reddit combo dataset. With the help of probability density function Kernel Density Estimation (KDE) plots provide more the best way to visualize the continuous data [13]. The x-axis shows the number of words ranging from 0 to 30000. The y-axis shows the density. Blue curve indicates the non-stress posts body, and orange curve indicates the body part of the stressed posts. We can see the length of the body for both the stressed and non-stressed posts are almost in the same length. While reaching the 5000 words both the curves are going down which is stating that the length of the body part for both the stressed and non-stressed posts are not more than 5000 words.

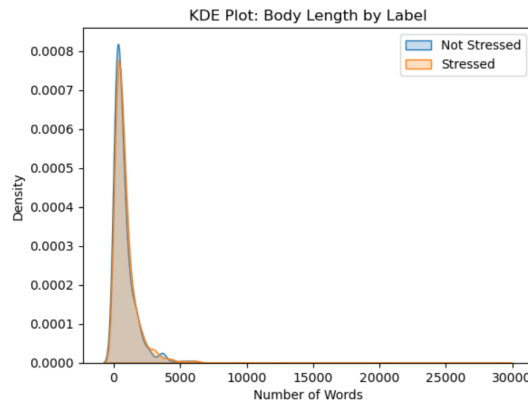


Figure 19: Distribution of the body in the Reddit combo dataset

This KDE plot, Figure 20 shows the text distribution of the twitter tweets from the twitter full dataset. With the help of probability density function Kernel Density Estimation (KDE) plots provide more the best way to visualize the continuous data [13]. The x-axis shows the number of words ranging from 0 to 1000. The y-axis shows the

density. Blue curve indicates the non-stressed tweets text length and the orange curve indicates the stressed tweets text length. From the beginning we can see that there is this peak from 0 to 150 for both the stressed and non-stressed tweets. After 150 again there is peak till 300 words for both the tweets. After that the curve slightly started to going down, which indicates mostly the length of the tweets for both the stressed and non-stressed are lies in between 50 to 200 with mostly 150-200 words.

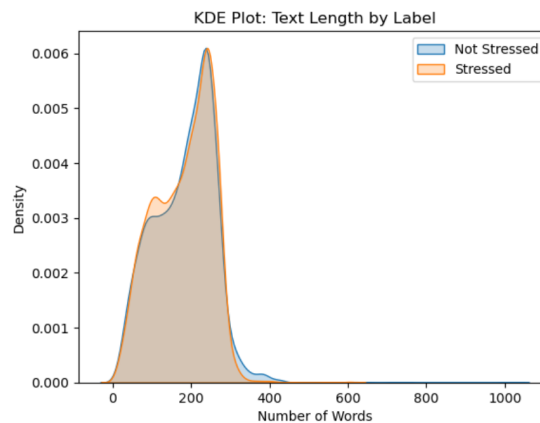


Figure 20: Text length distribution of the twitter full dataset

This KDE plot, Figure 21, shows the text length distribution of the twitter non-advert dataset by label. With the help of probability density function Kernel Density Estimation (KDE) plots provide more the best way to visualize the continuous data [13]. The x-axis shows the number of words ranging from 0 to 400 and y-axis shows the density. We can see the two peaks in this plot. First peak occurs around 150 for both the stressed and non-stressed tweets, slightly more density for the stressed ones. The second occurs around 200-250 for both the tweets, but this time we can see the more density for the non-stressed ones. From this we can say that people are posting their feelings like stressed ones in shorter length and they are expressing their happiness in more words.

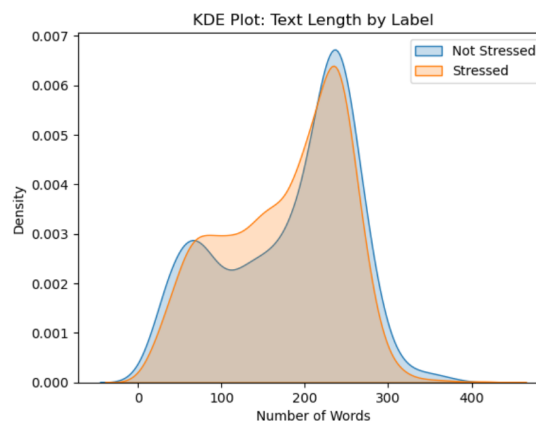


Figure 21: Text length distribution of the twitter non-advert dataset

## 5 Platform, Language, and Architecture

Programming language python is used for this dissertation project. Libraries like seaborn and matplotlib were used for the visualisation like histograms and KDE plots. The main goal of this dissertation is detecting whether the person is feeling stressed or not based on the post they made on the social media, which involves the classification of the text data, which leads to the usage of natural language processing technology. The architecture I used was the DistilBERT model. DistilBERT model is one of the pre-trained language models (PLM). DistilBERT is the faster and efficient version of BERT model. DistilBERT is 40% smaller, 60% faster and maintains 97% of the results of the BERT model.

The platform I choose for performing this project is Google Colab, because of its free access to GPUs, and its pre-installed libraries. For my dissertation Jupyter notebook is not supporting for training the model, so I choose Google Colab because if it's free GPU access.

## 6 Results and Discussion

The methodology I used for detecting whether the post is related to stress or not-stressed which is a binary classification is the DistilBERT model, which we fine-tuned with our all four datasets. Firstly, the dataset was split into 80% training and 20% testing. DistilBERT tokenizer was used to convert the titles into numbers. Each text is divided into sub-word tokens by the tokenizer. All the titles are padded so that all of them will be

having the same length of 512 tokens. The model was trained after this with the 3 epochs which means that the model can see the full training data 3 times, and with a batch size of 32 which states that 32 titles can be processed at once. Confusion matrix was plotted to visualize the performance of the model. With the help of confusion matrix, we can calculate accuracy, precision, recall and f1-score. For this thesis I choose accuracy and f1-score as the evaluation metrics. We all know that accuracy tell how many were correctly identified. The main reason for choosing f1-score is we are dealing with the people mental health aspects, so in this case we need to correctly identify the stressed ones and at the same time we need to minimize the wrong predictions like predicting as stressed for the non-stressed individual, so that people will trust the model.

**True Positive:** The actual value is positive, and the model predicts a positive value [2].

**True Negative:** The actual value is negative, and the model predicts a negative value [9].

**False Positive:** The actual value is negative, and the model predicts a positive value [15].

**False Negative:** The actual value is positive, and the model predicts a negative value [25].

### Accuracy

Accuracy is defined as the how many predictions were actually correct out of all the predictions made by the model [24].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

### Precision

Precision is defined as the out of all the positive predictions made by the model, how many were actually correct [18]. In our case it is how many predicted stressed posts are actually stressed posts.

$$\text{Precision} = \frac{TP}{TP + FP}$$

### Recall

Recall is defined as the how many actual positives cases are correctly predicted by the model [21]. In our thesis it is like how many of the actually stressed posts were correctly

identified by the model.


$$\mathbf{Recall} = \frac{TP}{TP + FN}$$

### F1-Score


F1-score is the harmonic mean of the precision and recall [1]. It gives the balance between the correctly detecting stressed posts and reducing the false alarms.

$$\mathbf{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

From the below table we can see the accuracy and f1-score for all the datasets.



Dataset	Accuracy	F1-Score
Reddit Title	94.8	94.6
Reddit Combo Text	92.5	95.8
Reddit Combo Body title	95.8	97.4
Twitter Full Text	87.9	88
Twitter Full Hashtags	75	78
Twitter Non-Advert	85.8	88.4



First we will discuss about the results for the Reddit title dataset. From the above table we can see that our DistilBERT model had achieved 94.8% accuracy and 94.6 f1-score for the reddit title dataset. Here we trained the model with the titles of the posts they made on the Reddit platform. From this we can say that like both the evaluation metrics had got the best ones which means though the text length is small the model is trained well in detecting the overall stressed and at the same time predicting the right stressed ones by minimizing the false alarms.

For the Reddit Combo dataset, when the model is trained with the title only it is yielding the lesser results in accuracy when compared to the titles from the Reddit title dataset. Though the accuracy is less, we can see the progress in the F1-score which is higher than the previous one, indicating that only with the title itself the model can detect the stressed ones and at the same time can reduce the false alarms as well.

Here comes the main part, which is training the model with both the body and the title of the post, which was posted by different people on the Reddit platform. When the model is trained with the complete post we can see the model performance which is improved compared to the results that the model achieves when it is trained and tested

with titles itself. One thing we need to understand carefully is the increasing the F1-score, which states that the model can detect the stressed persons and along with that in the same time it can decrease the false predictions when it is trained with both the title and the body. This makes sense, when the model get to know about the full post it can learn the entire context and gives the result, which makes it more suitable for real-world applications.

Now we will discuss the results for the Twitter datasets. For the Twitter full dataset, the model is trained with the text and it achieves the 87.9% accuracy and 88% f1-score, which means though the accuracy is little bit small compared to the f1-score, the model is predicting the correct ones and also reducing the false alarms from the texts itself. When the model is trained with the hashtags of the twitter full dataset, scores for both the evaluation metrics is less. This is making sense completely, we can predict whether the person is feeling happy or not from the tiles, and body, but when it comes to hashtags it would be little difficult compared to all those, because they are very small and sometimes people use different hashtags which doesn't completely fit into the situation. So making predictions with just the hashtags is achieving the accuracy and f1-score less compared to all others. Finally twitter non-advert dataset with the text itself achieved the 85.8 % accuracy and 88.4% f1-score. From all the results we can say that we have less information from the Twitter when compared to Reddit, which makes the great chances for detecting stress from its posts, because of their lengthy and descriptive posts.

## 7 Conclusion and future work

In this dissertation we used the DistilBERT model which is 60% faster than BERT to detect whether the person is feeling stressed or not based on the posts they made on the social media platforms. For our project we choose two platforms called Reddit and Twitter. After data preprocessing and data visualization,, our model was trained on these datasets and achieved the best results. The future work includes of working with the larger datasets from the multiple platforms so that we can have texts from different people which enables to the deep understanding of the data. Along with the text data, working with the images and videos in this concept like detecting whether they are feeling stressed or not would be more helpful. In this thesis, throughout the process we checked



whether the person is feeling stressed or not, but it would be better if they include stages like low, medium and high like multi model instead of binary classification, because we all know that these things like stress starts from slow and after ignoring all the times it will become worse in the end, so it would be more helpful for the people and the health professionals if they have models which can detect the stress from the basic level. Ethical considerations and privacy issues should be handled carefully in the future like building the models that can explore techniques like anonymization, to make sure that the final output is safe, responsible and the most important thing unbiased one. Developing the real models that would gain trust in the people and at the same time the mental health professionals can use these models while treating their patients.

## References

- [1] S. Afifa and R. Kumar. Development of f1-score as a robust metric for classification tasks. *International Journal of Computer Science and Information Technology*, 14(2):77–85, 2022.
- [2] Aniruddha Bhandari. Confusion matrix in machine learning, 2020. [Accessed 10 September 2025].
- [3] Wikipedia contributors. Bert (language model) - wikipedia. [https://en.wikipedia.org/wiki/BERT\\_\(language\\_model\)](https://en.wikipedia.org/wiki/BERT_(language_model)), 2023.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.
- [5] Ali Gardazi. Challenges in deploying bert for real-time applications. *Journal of Machine Learning Applications*, 12(1):101–115, 2025.
- [6] GeeksforGeeks. Bar plot in matplotlib, 2025. [Accessed 10 September 2025].
- [7] GeeksforGeeks. Generating word cloud in python, 2025. [Accessed 10 September 2025].
- [8] GeeksforGeeks. Histogram - definition, types, graph, and examples, 2025. [Accessed 10 September 2025].
- [9] GeeksforGeeks. Understanding the confusion matrix in machine learning, 2025. [Accessed 10 September 2025].
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distillation loss for knowledge transfer in deep learning models. *Neural Information Processing Systems*, 2015.
- [11] Wei Hu and Lin Zhang. Accelerating nlp tasks with pre-trained language models. *International Journal of Artificial Intelligence*, 15(2):45–60, 2023.
- [12] Mandar Joshi, Danqi Liu, Omer Levy, and Luke Zettlemoyer. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

- [13] Janvi Kumari. Kernel density estimation (kde) plots, 2025. [Accessed 10 September 2025].
- [14] Xinyu Li and Yue Wang. Masked language modeling techniques for nlp pre-training. *Journal of Natural Language Engineering*, 27(5):507–520, 2021.
- [15] Jacob Murel and Eda Kavlakoglu. What is a confusion matrix?, 2024. [Accessed 10 September 2025].
- [16] Ramesh Nair and Ananya Gupta. Evaluating distilbert for stress detection from social media. *Journal of Computational Social Science*, 6(2):75–89, 2024.
- [17] Suja Sreeith Panicker and Prakasam Gayathri. A survey of machine learning techniques in physiology based mental stress detection systems. *Biocybernetics and Biomedical Engineering*, 39(2):444–469, 2019.
- [18] David M.W. Powers. Evaluation metrics in machine learning: precision, recall, f1-score. *Journal of Machine Learning Research*, 21:1–36, 2020.
- [19] Aryan Rastogi. Stress detection from social media using nlp techniques. *Journal of Social Computing*, 8(3):123–135, 2022.
- [20] Marco Rehbein and Ping Li. Improving bert performance for low-resource domains. *Artificial Intelligence Review*, 53:4567–4582, 2020.
- [21] Lucas Ribeiro and Ana Fernandes. Near real-time stress detection using nlp models. *Applied Artificial Intelligence*, 39(1):112–130, 2025.
- [22] S-Logix. Deep contextual word embedding model for semantic similarity. <https://slogix.in/machine-learning/deep-contextual-word-embedding-models-for-semantic-similarity/>, 2025. Accessed: 09 September 2025.
- [23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [24] V. Sathyanarayanan. Confusion matrix and evaluation metrics in machine learning. *Data Science Review*, 9(1):22–35, 2024.

- 
- [25] DataCamp Team. What is a confusion matrix in machine learning?, 2024. [Accessed 10 September 2025].
- [26] The World Mental Health Institute (WMHI). Why 65% of people won't get help if they have a mental problem, 2025. [Accessed 9 September 2025].
- [27] Romeo Turcan and Kathleen McKeown. Dreddit: Early detection of stress on social media. *Proceedings of ACL*, pages 123–132, 2019.
- [28] Analytics Vidhya. Understanding bert for nlp. <https://www.analyticsvidhya.com/blog/2019/06/understanding-bert-nlp/>, 2019.
- [29] Xiangxuan Wan and Li Tian. User stress detection using social media text: A novel machine learning approach. *International Journal of Computers Communications Control*, 19(5), 2024.
- [30] M. Zhuang, D. Cheng, X. Lu, and X. Tan. Postgraduate psychological stress detection from social media using bert-fused model. *PloS One*, 19(10):e0312264, 2024.