Unveiling Mental Health Insights through Social Media Analysis

DATS 6312 – Natural Language Processing

Team 4: Anjali Mudgal, Sunisha Harish, Udbhav Kush

Table of Contents

1. Introduction

In 2020, the World Health Organization (WHO) cast a stark light on the global prevalence of mental disorders, revealing that nearly 1 billion people, constituting just over one in ten of the world's population, grapple with the complexities of mental health. Closer to home, the United States witnessed a striking statistic in 2021, where more than one in five adults—equivalent to a staggering 57.8 million individuals—were reported to be living with a mental health condition. Amidst these staggering numbers, a profound societal shift has unfolded, as a significant majority now turns to social media as their primary conduit for communication. This transformation positions social media platforms not only as hubs for connection but also as unique arenas for gauging and understanding the subtle nuances of mental health indicators in an increasingly interconnected world. As our digital landscapes evolve, the intersection between mental health and social media becomes an ever more critical frontier for exploration and analysis.

Figure 1: Mental Health Project

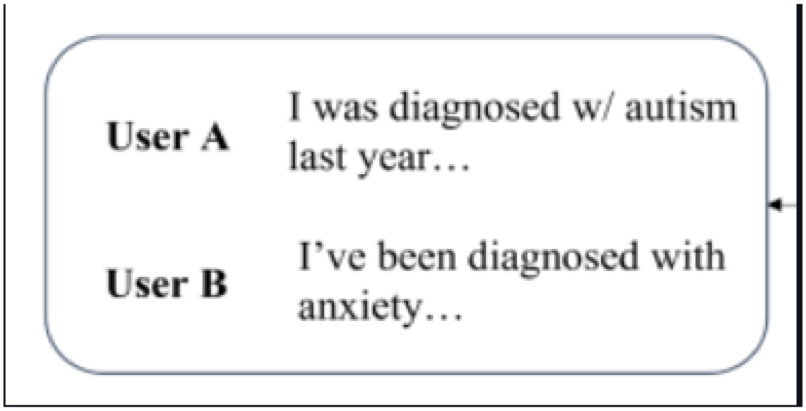
2. Dataset

2.1 Dataset Used

Upon formulating our problem statement, we embarked on a comprehensive literature review to investigate the various solutions that have been applied to address this challenge. During our exploration, we encountered a research paper available at (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10315126) that aligns closely with our specified requirements. Recognizing the compatibility of their data with our research objectives, we opted to leverage the data presented in this research paper as a valuable foundation for our own investigation.

The dataset utilized in our study was meticulously curated by the author of the paper through the following methodology:

- Users and posts, encompassing both English and Spanish content, were extracted from Twitter utilizing its application programming interface (API). Tweets obtained originated from users who openly shared their diagnoses, covering the timeframe from September 1st, 2020, to August 31st, 2021.
- After the initial collection of public diagnosis tweets, the most recent tweets (up to 3200) from each user were retrieved using the Twitter API. Subsequently, the dataset underwent refinement by discarding retweets and non-target language tweets. The posting period was confined to a span of 5 years, calculated from the first to the last tweet within the dataset.
- To establish a comparative basis, these tweets were then matched with a control group of users based on similar criteria, including the number of tweets and the posting period. This rigorous data collection and filtering process forms the foundation of our investigation, ensuring a comprehensive and focused analysis.



User A I was diagnosed w/ autism last year…

User B I've been diagnosed with anxiety…

Figure 2 : Sample Tweets

The tweets were classified into 10 different classes based on the mental health disorder. The classes were: EATING DISORDER, SCHIZOPHRENIA, OCD, PTSD, ANXIETY, BIPOLAR, AUTISM, DEPRESSION, ADHD, CONTROL.

Control Group was the non diagnosed group of users.

2.2 Splits

We did a standard split of training-test and dev. The training set was used for training the model and the test set was used to make decisions while training the model. The dev set was untouched and only evaluated on right before the project presentation.

3. Modeling

3.1 Transformers BERT

BERT is an open source machine learning framework for natural language processing (NLP). The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

BERT, which stands for Bidirectional Encoder Representations, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. (In NLP, this process is called attention.)

The reason we choose to use transformers and specifically BERT is because:

- Bert is because it is designed to understand the context in which words appear in a sentence, which is crucial for accurately interpreting the often nuanced and context-dependent language used in tweets related to mental health.

- BERT is pre-trained on a massive amount of diverse textual data. This pre-training allows it to learn general language patterns.

- BERT supports multiple languages, which is essential for analyzing tweets in different languages.

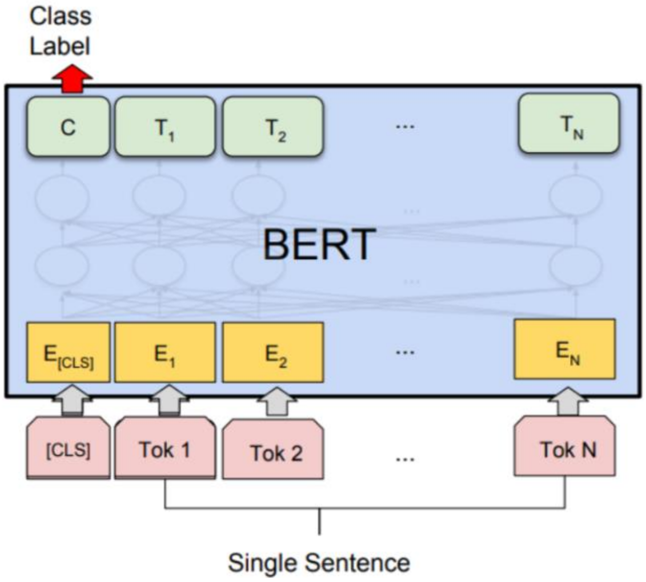


Figure 3: Bert Single Sentence Classfication Task Architecture

3.2 Training the model for classification

We initially started by training off our data with all the 20 million samples we had. Later we downsampled our data as our model was biased towards the Control class. For each split the sampling is done by taking the disorder group with the least data and sampling the rest of the groups to contain the same data points in order to avoid bias.

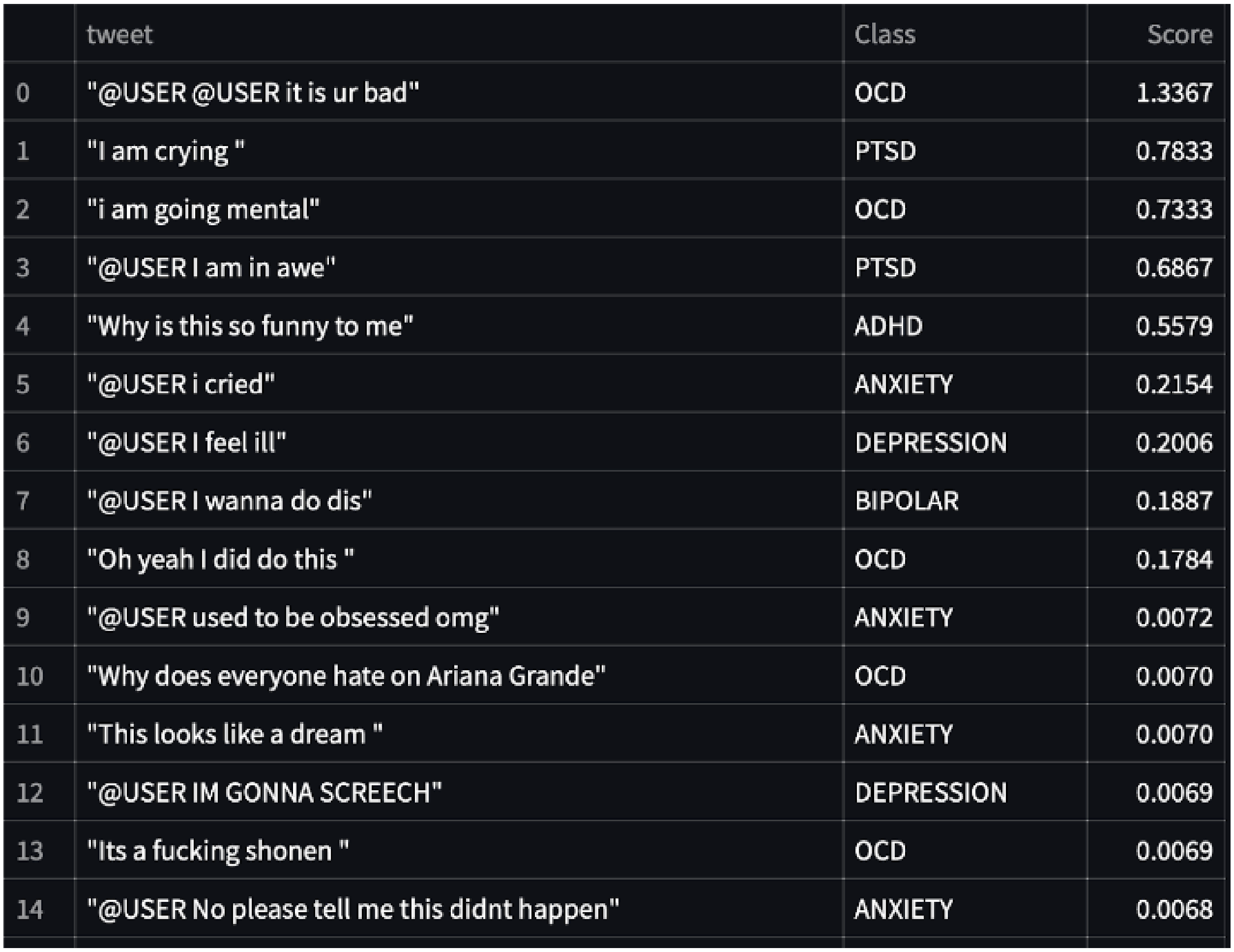
We used the pre trained model on both the English and Spanish dataset.

For training the English model, we finetuned the bert-base-uncased model for BertForSequenceClassification. For this, we achieved an accuracy of 0.65 and an F1 score of 0.62.

For training the Spanish model, we finetuned the dccuchile/distilbert-base-spanish-uncased-finetuned-xnli for AutoModelForSequence Classification. For this, we achieved an accuracy of 0.61 and an F1 score of 0.59.

The results were not that great because usually different disorder group tweets tend to have the same kind of context and structure.

For example, if we look at the figure below “I am crying” and “@USER i cried” mean the same but the same kind of tweets has been tweet by users from the different disorder group.



	tweet	Class	Score
0	"@USER @USER it is ur bad"	OCD	1.3367
1	"I am crying "	PTSD	0.7833
2	"i am going mental"	OCD	0.7333
3	"@USER I am in awe"	PTSD	0.6867
4	"Why is this so funny to me"	ADHD	0.5579
5	"@USER i cried"	ANXIETY	0.2154
6	"@USER I feel ill"	DEPRESSION	0.2006
7	"@USER I wanna do dis"	BIPOLAR	0.1887
8	"Oh yeah I did do this "	OCD	0.1784
9	"@USER used to be obsessed omg"	ANXIETY	0.0072
10	"Why does everyone hate on Ariana Grande"	OCD	0.0070
11	"This looks like a dream "	ANXIETY	0.0070
12	"@USER IM GONNA SCREECH"	DEPRESSION	0.0069
13	"Its a fucking shonen "	OCD	0.0069
14	"@USER No please tell me this didnt happen"	ANXIETY	0.0068

Figure 4: Tweets of users suffering from a mental disorder

3.3 Finding the similarities

To foster a sense of connection and reassure the user that they are not alone, we aim to provide them with tweets that closely resonate with their own experiences if they have been diagnosed with a particular disease. This personalized approach seeks to offer a supportive and relatable online environment, emphasizing shared sentiments and commonality in the face of health challenges.

3.3.1 Combining BERT embedding with the BM25 scores

Given that the Twitter dataset is derived from tweets by individuals openly acknowledging diagnoses such as 'ADHD,' 'ANXIETY,' 'AUTISM,' 'BIPOLAR,' 'CONTROL,' 'DEPRESSION,' 'EATING DISORDER,' 'OCD,' 'PTSD,' and 'SCHIZOPHRENIA,' it's noteworthy that individuals with different diagnosed disorders often share tweets that contain content similar to each other, which was one of the main reason our classified model was not performing well.

In a few research papers, we noticed that people usually combined BERT embedding similarities with some keywords information. Since our dataset is very sensitive to the punctuations and the words used by the user in his tweet, we tokenized the tweets using TweetTokenizer and calculated the bm25 similarity scores.

We would be combining this information to the closest 40-50 tweets that we get from the HNSW index.

We calculated the scores as follows:

1. Created BERT embeddings from the model that we fine-tuned using BertForSequenceClassification.
2. For all the tweets by removing the classification layer and getting the output bert embedding of CLS for each tweet.
3. These embeddings were used to create HNSWFlat indexes from FAISS . Later, the embeddings were added to the clusters.
4. Whenever the user enters a tweet, we calculate the BERT embedding for that user and search for the top 20 closest tweets in the HNSW. (It returns the distance, which we are using to get the similarity)
5. Using these tweets we are finding the Okapi BM25 score for each tweet with the query tweet.
6. The combined score of each embedding with these 20 tweets would be 0.8*(similarity from HNSW bert embedding matching) + 0.2*(bm25 similarity)
7. We now get the top similar tweets, we would use the unique classes in these tweets to do a zero shot classification to get the maximum probable class in which the tweet can fall into.

8. After getting the top class, we are displaying the tweets in the top 20 selection from hnsw which falls into this category based on the combined similarity score.

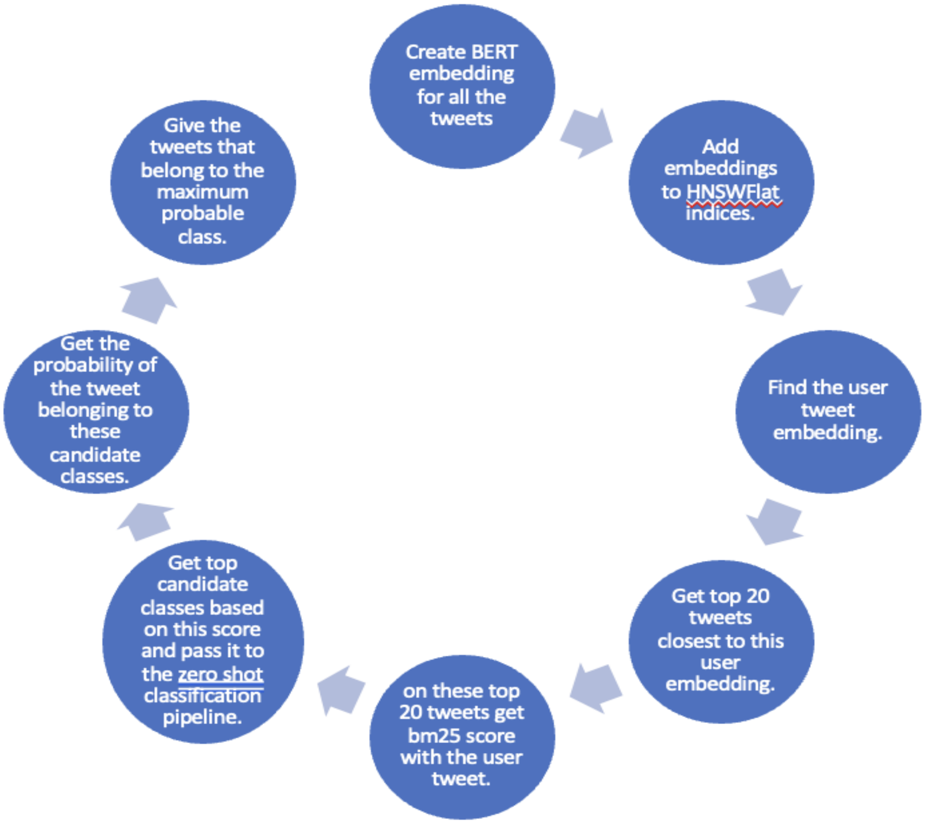
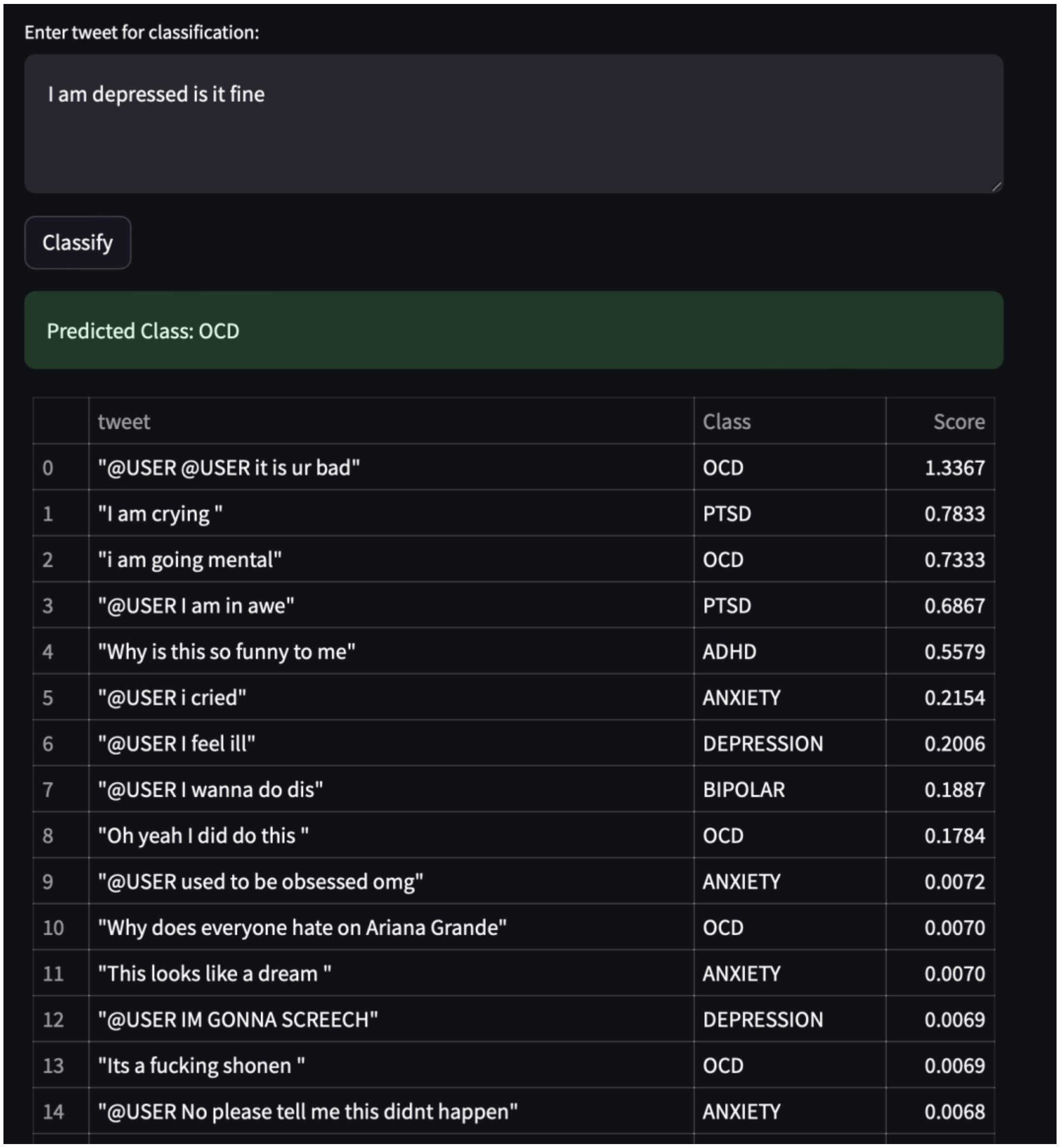

Create BERT embedding for all the tweets
Add embeddings to HNSWFlat indices.
Find the user tweet embedding.
Get top 20 tweets closest to this user embedding.
on these top 20 tweets get bm25 score with the user tweet.
Get top candidate classes based on this score and pass it to the zero shot classification pipeline.
Get the probability of the tweet belonging to these candidate classes.
Give the tweets that belong to the maximum probable class.

Figure 5 : Finding similar tweets architecture

4. App

We've developed an application where users can input a tweet in English and the system categorizes it into either a disorder group or a control group. And similar tweets are also shown to the user to let them know that they are not alone.

We also have a zero shot classifier that gives us the probabilities of the tweet belonging to the candidate classes. Similar tweets relating to the class with the highest importance are showcased.



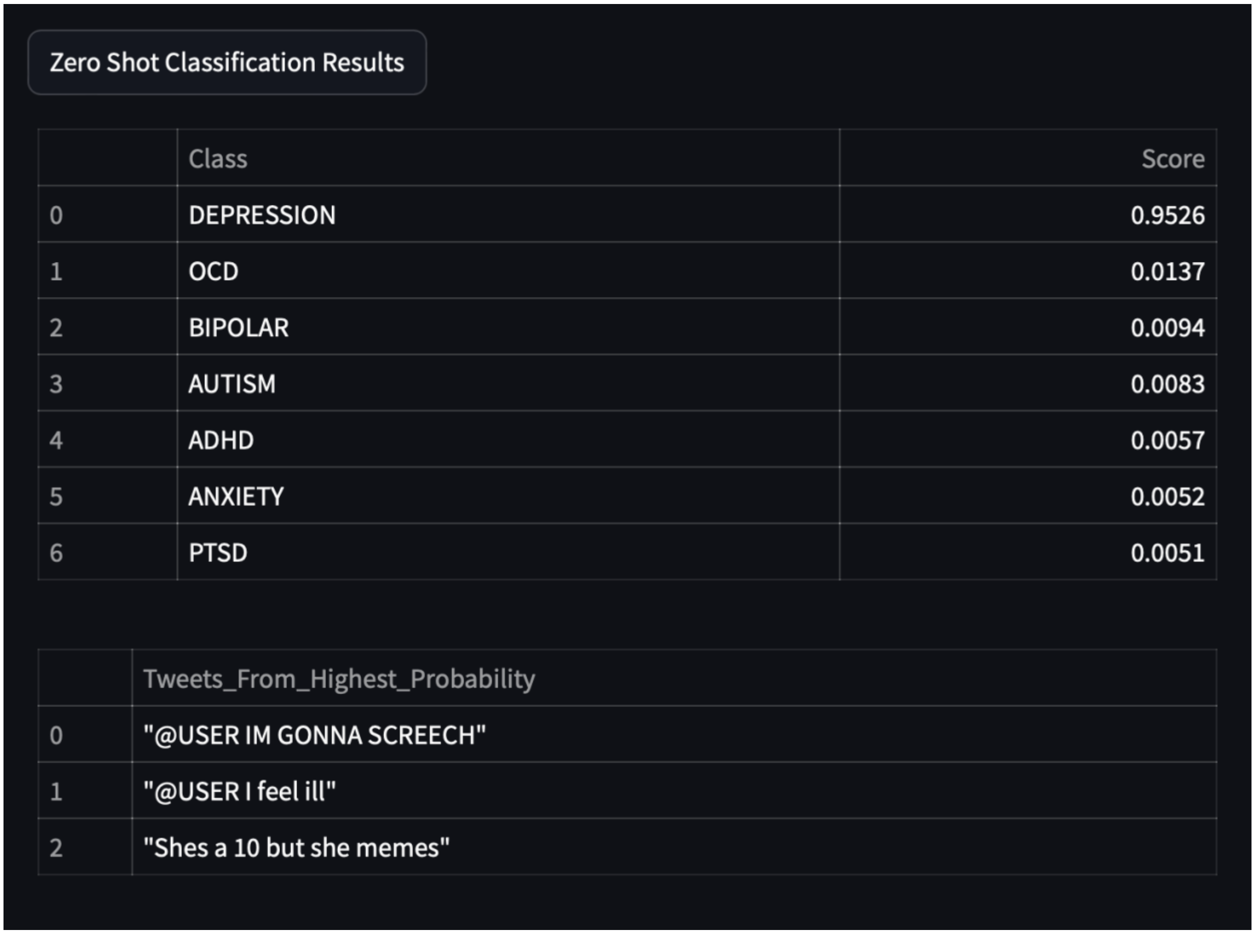
Enter tweet for classification:

I am depressed is it fine

Predicted Class: OCD

	tweet	Class	Score
0	"@USER @USER it is ur bad"	OCD	1.3367
1	"I am crying "	PTSD	0.7833
2	"i am going mental"	OCD	0.7333
3	"@USER I am in awe"	PTSD	0.6867
4	"Why is this so funny to me"	ADHD	0.5579
5	"@USER i cried"	ANXIETY	0.2154
6	"@USER I feel ill"	DEPRESSION	0.2006
7	"@USER I wanna do dis"	BIPOLAR	0.1887
8	"Oh yeah I did do this "	OCD	0.1784
9	"@USER used to be obsessed omg"	ANXIETY	0.0072
10	"Why does everyone hate on Ariana Grande"	OCD	0.0070
11	"This looks like a dream "	ANXIETY	0.0070
12	"@USER IM GONNA SCREECH"	DEPRESSION	0.0069
13	"Its a fucking shonen "	OCD	0.0069
14	"@USER No please tell me this didnt happen"	ANXIETY	0.0068

Figure 6: Combined similarity score : BERT+BM25



Zero Shot Classification Results

	Class	Score
0	DEPRESSION	0.9526
1	OCD	0.0137
2	BIPOLAR	0.0094
3	AUTISM	0.0083
4	ADHD	0.0057
5	ANXIETY	0.0052
6	PTSD	0.0051

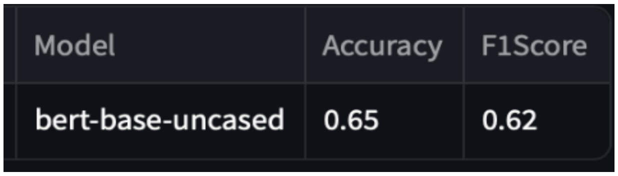
	Tweets_From_Highest_Probability
0	"@USER IM GONNA SCREECH"
1	"@USER I feel ill"
2	"Shes a 10 but she memes"

Figure 7: Zero shot classification results

5. Results

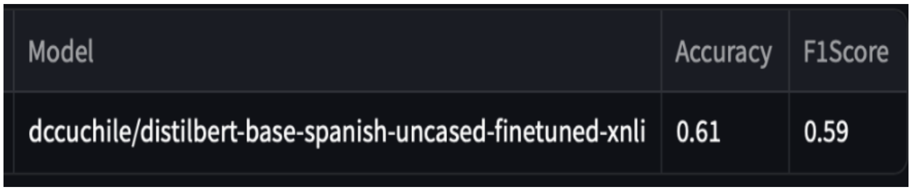
Achieving higher accuracy in our models posed a challenge due to the inherent complexity of tweets related to anxiety and depression. The overlap in content between individuals with diagnosed anxiety or depression and those without made it challenging to discern and classify with utmost precision. As a result, our top-performing models reflect the intricate nature of mental health discussions on social media.

- For English dataset



Model	Accuracy	F1Score
bert-base-uncased	0.65	0.62

- For Spanish dataset



Model	Accuracy	F1Score
dccuchile/distilbert-base-spanish-uncased-finetuned-xnli	0.61	0.59

6. Conclusion

In conclusion, our project wasn’t able to successfully classify English language and Spanish language tweets using the BERT based classification. One main reason is similar context in different classes.

We learned that passing keyword information can help with text analysis where types of words, punctuations used also matter.

We also understood the vector based databases and how HNSW or similar vector based similarity algorithms can help improve the real time search results.

7. Further Improvements

- The development of a mental health assistance chatbot to provide necessary support for the user proved to be challenging for us. However, we plan to try to incorporate this in the future.

- We can also try implementing the langdetect model to detect the language in which a tweet was posted and classify it accordingly.

- We could try langdetect model to detect the language in which the tweet was posted and then use the classification model accordingly.

References:

- *Dccuchile/bert-base-Spanish-WWM-uncased · hugging face*. dccuchile/bert-base-spanish-wwm-uncased · Hugging Face. (n.d.). https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
- *Bert-base-uncased · hugging face*. bert-base-uncased · Hugging Face. (n.d.-a). https://huggingface.co/bert-base-uncased
- M. E. Villa-Pérez, L. A. Trejo, M. B. Moin and E. Stroulia, "Extracting Mental Health Indicators from English and Spanish Social Media: A Machine Learning Approach," in IEEE Access, doi: 10.1109/ACCESS.2023.3332289.
- Mental Illness Classification on social media texts using Deep Learning ... (n.d.). https://www.researchgate.net/publication/361757484_Mental_Illness_Classification_on_Social_Media_Texts_using_Deep_Learning_and_Transfer_Learning
- Nayak, P. (2020, July 6). *Mental Health Classification using Distilbert and supervised machine learning*. Medium. https://nayakpplaban.medium.com/mental-health-classification-using-distilbert-and-supervised-machine-learning-d4b0c536f7b2