Classifying Social Media Users Before and After Depression Diagnosis via their Language Usage: A Dataset and Study

Falwah Alhamed[1,2], Julia Ive[3], Lucia Specia[1]
[1] Department of Computing, Imperial College London, London, UK
[2] King Abdulaziz City for Science and Technology(KACST), Riyadh, Saudi Arabia
[3] Queen Mary University of London, London, UK
{f.alhamed20,l.specia}@imperial.ac.uk, j.ive@qmul.ac.uk

Abstract

Mental illness can significantly impact individuals' quality of life. Analysing social media data to uncover potential mental health issues in individuals via their posts is a popular research direction. However, most studies focus on the classification of users suffering from depression versus healthy users, or on the detection of suicidal thoughts. In this paper, we instead aim to understand and model linguistic changes that occur when users transition from a healthy to an unhealthy state. Addressing this gap could lead to better approaches for earlier depression detection when signs are not as obvious as in cases of severe depression or suicidal ideation. In order to achieve this goal, we have collected the first dataset of textual posts by the same users before and after reportedly being diagnosed with depression. We then use this data to build multiple predictive models (based on SVM, Random Forests, BERT, RoBERTa, MentalBERT, GPT-3, GPT-3.5, Bard, and Alpaca) for the task of classifying user posts. Transformer-based models achieved the best performance, while large language models used off-the-shelf proved less effective as they produced random guesses (GPT and Bard) or hallucinations (Alpaca).

Keywords: Dataset, Lexicon, Mental Health, Depression classification, Social Media, X, Twitter, BERT, LLM, GPT, NLP

1. Introduction

According to the United Kingdom National Health Service (NHS), one in four people experiences mental health problems (McManus et al., 2009). Mental illnesses can negatively influence the quality of life as it is considered one of the causes of years lived with disability, and it is related to high suicide rates (Association, 2013). Understanding the patterns of mental episodes or shifts from one mental state to another is important, not only to help a particular user but also for other users with similar symptoms. Research in mental health monitoring using Natural Language Processing (NLP) is usually carried out using electronic health records and standardized diagnostic questionnaires (Kim et al., 2020; Pradier et al., 2021; Mesbah et al., 2021; de Oliveira et al., 2021).

An alternative approach looks at social media data, which is abundant and diverse. A large number of users on social media often share updates about their daily lives, including moods and feelings. When it comes to research on applying NLP techniques to social media data, it tends to focus on classifying the presence vs absence of depression (Boinepelli et al., 2022; Chancellor and De Choudhury, 2020), or detecting the transition from depression to suicide ideation (De Choudhury et al., 2016; Gong et al., 2019; Matero et al., 2019; Sawhney et al., 2020). To the best of our knowledge, there is no study aiming to understand users' language before and after the diagnosis of depression using NLP on social media. In this paper, we aim to understand whether we can use social media as a data source to understand the linguistic characteristics of users potentially suffering from depression posts on X (formerly known as Twitter). The goal is to surface meaningful patterns to understand linguistic shifts from healthy to depressed states. Our key contributions can be summarized as follows:

- The first English dataset of textual posts by the same users before and after reportedly being diagnosed with depression[1]
- A lexicon for finding posts with symptoms of depression[2]
- Empirical work comparing multiple predictive models (based on SVM, Random Forests, BERT, RoBERTa, GPT-3, GPT-3.5, Bard, and Alpaca) built using our dataset for the task of classifying user posts as before and after depression.

2. Related Work

There is a growing body of literature that investigates mental health in social media. A few surveys analyze these studies in more detail, such as Chancellor and De Choudhury (2020) and Skaik

[1] The dataset can be found at: https://github.com/falwah-alhamed/Depression_Tweets
[2] The Lexicon can be found at: https://github.com/falwah-alhamed/Depression_Tweets

LREC-COLING 2024, pages 3250–3260
20-25 May, 2024.

and Inkpen (2021). To summarise, some studies use the analysis of user activity — such as social connections, number of likes, and number of posts (Shen et al., 2017; Peng et al., 2019) — to detect mental illnesses in social media. Other studies use NLP to understand the meaning behind words and derive textual characteristics (Cacheda et al., 2019; Shen et al., 2017; Joshi et al., 2018; Verma et al., 2020; Jamil et al., 2017). Most of the studies were conducted to understand linguistic cues and distinguish users with depression from control users. For example, a recent study by Boinepelli et al. (2022) aimed to detect depression at the user-level. They used RoBERTa to train a model to classify users as people suffering from depression or control users. While it is reasonable to train the model using data from people suffering from depression and control users, defining control users solely by the absence of explicit mentions of depression in their posts may not reliably indicate their mental health status. Therefore, a more reliable approach is needed when identifying control users in such research studies. A recent study conducted by Santos et al. (2020) used a similar approach in defining users suffering from depression by reporting being diagnosed with depression or taking antidepressant medications on Brazilian Twitter. They extracted the dates of these Portuguese posts with reports and pulled the data before and after that date, it is worth noting that starting taking antidepressants does not necessarily mean the start of the disease as some patients are put on therapy sessions for months to years before prescribing antidepressants, this might introduced error to their dataset. Yet, an English dataset of social media posts of **the same users** prior to and post depression diagnosis is still lacking.

Some studies leverage lexicons to support depression detection in social media. Borba de Souza et al. (2022) used 59 terms dictionary built based on DSM-5 depression and anxiety diagnosis manual besides word embeddings. They used it for a multiclass classifier for depression, anxiety, and comorbidity. The use of the dictionary contributed to better results, however, more investigations on this field need to be done to distinguish between classes. Another study by Alghamdi et al. (2020) focuses on the Arabic forum 'Nafsany,' dedicated to mental health discussions where users share their stories, emotions, and challenges while engaging in interactive conversations via comments. They collected forum posts related to depression and manually labelled them as indicative of depression or not. They developed ArabDep, an Arabic lexicon comprising diagnostic terms for depression. Classification was performed using two distinct approaches: a lexicon-based method leveraging ArabDep, and machine learning approaches involving the training of multiple models. Their study showed that the lexicon enhanced the classification results, which inspired our work.

Research into the detection of changes in depression levels via social media usually addresses shifts from depression to a suicidal condition, rather than addressing changes in language when the onset of depression occurs. Sawhney et al. (2020) conducted a study aimed at assessing suicidal risks on Twitter. They employed a time-aware transformer model and utilized a dataset containing instances of suicide ideation. They applied their model to analyze a dataset consisting of 34,306 tweets authored by 32,558 users. The primary objective was to classify individuals' suicide risk levels based on the patterns and content of their tweet sequences. It however became clear from the ratio of tweets to users that the quantity of tweets per user was not sufficient to discern a comprehensive longitudinal pattern. These challenges are taken into consideration while designing our study.

3. Dataset

3.1. Data Collection

For data collection, we targeted platforms that have a high influx of textual content in English, have a large number of active users, allow sequential posting, and have a high rate of posting frequency. The most well-known social media platform that meets these conditions is X.[3] Our goal was to collect data for individuals reportedly diagnosed with depression — that is, individuals posting text stating that they were diagnosed with depression, and where the individual is active on the platform on a weekly basis. We performed multiple steps in order to obtain high-quality data. The official Twitter API was used to search and we collected tweets based on keywords. First, we collected tweets with the search words 'I was diagnosed with depression on [specific month and year]' in the date range November 2018 to February 2019. The dates were selected to allow the collection of a sufficient number of tweets before and after diagnosis. A total of 2034 tweets were pulled and manually inspected to select only original tweets — that is, not replicated tweets or tweets telling a story about someone else. Only users who mentioned a specific month and year of diagnosis were selected. This resulted in 120 users. We then retrieved posts for three years before the diagnosis date and three years after the diagnosis date, resulting in 1.9 million tweets. To the best of our knowledge, this is the first dataset that contains posts before and after depression diagnosis **for the same users**. This allows comparison and monitoring of linguistic changes before

[3] https://developer.twitter.com/en/docs/twitter-api

and after diagnosis.

3.2. Lexicon Building and Data Filtering

The number of retrieved tweets is substantial, and the data may contain irrelevant posts on a variety of topics that may act as noise for the learning process. This calls for a filtering stage where we aim to keep only tweets that are related to, and representative of, a current mental health state.

To do that, we created a lexicon that can be used to screen posts for depression symptoms. If the post contains one of the lexicon words (in any PoS form) it will be included, otherwise, it will be discarded. The lexicon contains 598 words related to depression symptoms according to the Center for Epidemiologic Studies Depression (CES-D) validated scale. We started by building up a lexicon from existing words related to the depression scale questions which can be categorized into: poor appetite and eating disturbance, feeling down and depressed, concentration problems, feeling tired or having little energy, sleep disturbance, loss of interest, self-blame and shame, loneliness, and suicidal thoughts. For each category, we created a list of relevant words. The lexicon words are collected from multiple resources for each of these categories including the NLTK interface for WordNet `synset('food.n.02')`,[4] the NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013), the list of emotion words from Berkeley well-being institute,[5] the interest words list from university of Washington, [6] the suicidal words list in (Alhamed et al., 2022), and some words were extracted using Thesaurus [7] and Relatedwords.[8] Example words from sleep category `sleep, nap, awake, insomnia...etc.` , and from suicide category `die, jump, dead...etc.`

3.3. Data Pre-processing

Different pre-processing techniques were applied to the tweets using regular expression operations:[9] **tweet cleaning** from irrelevant information and X special characters and words such as RT, Fav, and Mentions; **stemming** using NLTK stemmer (NLTK, 2021) to remove redundant suffixes in a word; and **tokenisation** using the NLTK tokenizer. [10]

	Number	Avg. length (in words)
Users	120	-
All Tweets	1,969,645	14
Filtered Tweets	1,213,061	14
Chunks	28,697	433

Table 1: Dataset statistics (a chunk contains posts for one week).

3.4. Data Chunks Creation

Filtered posts are combined into sets of chunks for each user. Each chunk contains posts for a one-week period. This is to reflect the clinically validated scales of depression that are designed to measure depression such as the CES-D. According to these, patients are asked about symptoms of depression they faced *over the last week*.

The final dataset contains 28k chunks that are balanced between classes, namely “Before" and “After" depression diagnosis. The dataset statistics are illustrated in Table 1.

4. Data Analysis

In this section, we analyzed and compared linguistic and measurable metrics of posts before and after reported depression diagnosis in order to understand the differences between the two classes. We use three popular methods, namely LIWC, PoS, and Posting frequency. The data used in this section is not filtered by lexicon.

4.1. LIWC

Users' tweets were analyzed using Linguistic Inquiry and Word Count (LIWC) to understand the linguistic characteristics of users before and after being reported diagnosed with depression. This tool provides text analysis which includes occurrences of emotions in posts, mentioning friends and family, the occurrence of some topics such as health and food in posts, and the use of question and exclamation marks. A subset of the results with main topics is shown in Figure 1 (the rest of the topics in the results show no difference between classes). There is a decrease in both positive emotions and the use of exclamation marks, as well as a slight increase in focusing on the past and cognition after the diagnosis. This is expected as negative thoughts and focusing on the past are symptoms of depression. However, overall most of the characteristics are similar before and after diagnosis, with minor differences that are insufficient for distinguishing between the two classes.

[4] https://www.nltk.org/howto/wordnet.html

[5] https://www.berkeleywellbeing.com/list-of-emotions.html

[6] https://www.washington.edu

[7] https://thesaurus.yourdictionary.com/appetite

[8] https://relatedwords.org/relatedto/sleep

[9] https://docs.python.org/3/library/re.html

[10] https://www.nltk.org/api/nltk.tokenize.html

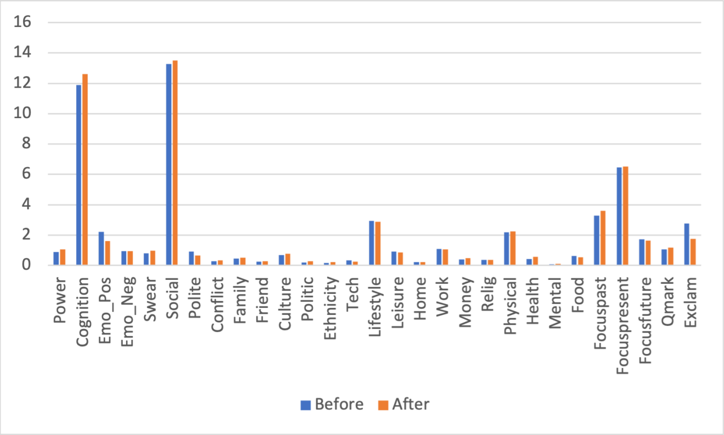

Figure 1: LIWC analysis: the percentage of used emotions, specific topics, and some punctuation marks in users' tweets before and after diagnosis of depression.

POS	Before Diagnosis	After Diagnosis
Noun, common (NN)	19.24%	***15.8%***
Noun, plural (NNS)	2.59%	3.04%
Noun, proper (NNP)	1.32%	1.77%
Verb, base form (VB)	3.14%	3.85%
Verb, present tense (VBZ)	2.11%	2.41%
Verb, past tense (VBD)	1.89%	2.25%
Adverb (RB)	3.91%	4.61%
Determiner (DT)	4.86%	***5.83***%
Pronoun (PRP)	3.73%	***4.49%***
Adjective (JJ)	8.24%	8.04%
Preposition (IN)	5.62%	***6.90%***

Table 2: Part of speech tags for posts before and after diagnosis.

4.2. POS analysis

Analyzing the distribution of Part of speech (POS) is a well-known methodology in NLP. We tracked the differences in POS for tweets before and after diagnosis using NLTK. [11] Results can be seen in Table 2, which shows the proportion of different POS in tweets before and after diagnosis. There is a clear decrease in the proportion of common nouns after depression. This is in line with previous studies on POS inspection for people suffering from depression (Bucur et al., 2021). Also, there are slight increases in the proportion of determiners, pronouns, and prepositions.

4.3. Posting Frequency

X is a micro-blogging social media platform that allows users to post up to 2400 tweets per day. Here we analyze posting frequency for people potentially suffering from depression before and after diagnosis. X has four categories of posts:

Original: Posts written by the user in his/her timeline.

Quoted: Posts are originally written by someone else and the user quotes this tweet and adds a comment then posts it in his/her timeline.

Replied: Posts are posted by someone and the user replies to this specific tweet.

Retweeted: Posts are originally written by someone else and the user reposts it in his/her timeline.

A hypothesis is that users will tweet less after a depression diagnosis considering their low mood and energy. in Figure 2 we show a comparison in posting frequency for each of the four categories before and after diagnosis. It can be seen that there is an obvious decrease in the number of original tweets while the reply tweets increased sharply. We investigate possible reasons for this increase by manual inspection of posts. There was a significant presence of conversations about self-disclosure in the replies. A possible explanation for this could be that users find it comfortable to talk about their feelings in a reply to someone and more acceptable than generating new original tweets. Quoted and retweeted tweets show smaller changes. Therefore, overall posting frequency has not changed before and after depression diagnosis, instead, the types of posts have changed.

5. Classification Models

Here we turn to using machine learning algorithms on this data, building classification models to learn the differences between posts pre- and post-depression diagnosis. The classification unit is a chunk of tweets and the same user will have some chunks labelled as positive (after depression)

[11] https://www.nltk.org/api/nltk.tag.pos_tag.html

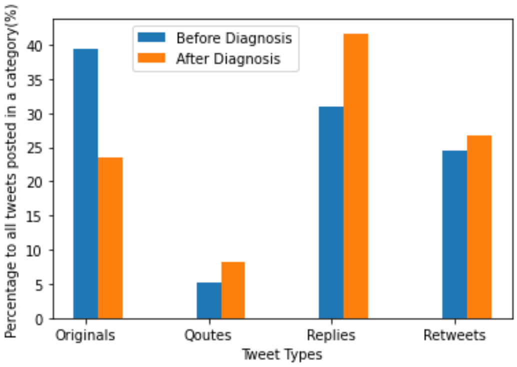

Figure 2: Posting frequency for people potentially suffering from depression before and after diagnosis.

and negative (before depression). Algorithms and settings used are shown in the following sections.

5.1. Classical ML Models

As baselines, we experiment with two traditionally well performing classical ML algorithms, SVM and Random Forests.

Support Vector Machines (SVM) We used SVM, we performed grid search hyper-parameter tunning and the best parameters were (kernel= 'linear', C=1, random_state=42). As SVM model only deal with numerical values, words are vectorized into numerical form (embedding) using Gensim Word2Vec[12] with 5k words and the word2vec specifications are num_features = 300, num_workers = 3. We took the average of the Word2Vec embeddings of the words in posts.

Random Forests (RF) We used Scikit learn RF We performed grid search hyper-parameter tunning for RF parameters ('max_depth': [5, 8, 15, 25, 30, 50,100], 'max_features': [2, 3,'sqrt'], 'min_samples_leaf': [1, 2, 5, 10], 'min_samples_split': [2, 5, 10, 15, 100], 'n_estimators': [100, 200, 500, 800, 1200]) and the best parameters found are (max_depth=50, max features=sqrt, min samples leaf=2, min samples split=2, n_estimators=200) which is used in this study. Words are vectorized into numerical form (embedding) using Word2Vec with the same vectors used in SVM.

5.2. Transformer-based Models

BERT Bidirectional Encoder Representations from Transformers (BERT) has proven to give high-accuracy results in classification tasks. Multiple comparisons between BERT and other text classification models have shown that BERT outperformed other models (Al-Garadi et al., 2021; Matero et al., 2019; Acheampong et al., 2021). Since BERT is bi-directional, it captures wider and contextual features (Devlin et al., 2019). The Hugging face library (Wolf et al., 2019) is used for BERT tokenization and fine-tuning, namely the'bert-base-uncased' model card.

RoBERTa Robustly Optimized BERT Pretraining Approach (RoBERTa) Liu et al. (2019a) has also been effectively used in text classification tasks. It often outperforms traditional classifiers and deep learning algorithms (Bozanta et al., 2021). The reason behind this performance might be its adaptability and strong performance on a variety of NLP benchmarks (Liu et al., 2019b). The Hugging face library is used for RoBERTa tokenization and fine-tuning, namely the 'roberta-base' model card.

MentalBERT MentalBERT is a model initialized with BERT-Base and subsequently trained using mental health-related posts collected from Reddit (Ji et al., 2022). The Hugging face library is used for MentalBERT tokenization and fine-tuning, namely the 'mental/mental-bert-base-uncased' model card.

5.3. Large Language Models (LLMs)

LLMs shown to have superior performance in comprehending and generating human-like text. We experiment with different types of LLMs to classify our texts. This is done in a zero-shot fashion with different prompts.

Prompts Prompts are questions or statements that are provided to the model to initiate and guide a conversation or specific task or to generate desired text. We experimented with different prompts, as illustrated in Table 3.

GPT We used the GPT-3 “text-curie-001" and GPT-3.5 “text-davinci-003" versions, as these have shown a strong ability to understand human-like emotional context (Tavast et al., 2022), and in sentiment analysis (Kheiri and Karimi, 2023). Since GPT-3 “text-curie-001" can handle a maximum of 2049 tokens divided between prompt and data, chunks were trimmed to fit this limit. We used the Official OpenAI Python library[13] to collect responses with the settings (temperature = 0.2, max tokens = 30).

[12] https://radimrehurek.com/gensim/models/word2vec.html

[13] https://platform.openai.com/docs/libraries/python-library

Do you think the person who wrote this text is depressed?
Do you think the person who wrote this text is depressed? answer yes or no
Do you think the person who wrote this text is depressed? return a probability percentage
To what extent do you think the person who wrote this text is depressed?
To what extent do you think the person who wrote this text is depressed? answer with one word only
Classify if the person who wrote this text is depressed
Classify if the person who wrote this text is depressed, reply with one word only

Table 3: Prompts used for LLMs answers

Google Bard Bard is Google's experimental commercial AI chat service designed to operate in a conversational manner, much like ChatGPT. The primary distinction lies in the fact that Google's service retrieves its information directly from the internet (Google-AI, 2023).

Alpaca Alpaca is LLM developed by Stanford University, based on Meta's LLaMA model. Alpaca is a promising new tool for NLP since it is open source. It is small and affordable, yet it is still capable of performing complex tasks (Taori et al., 2023).

6. Results

For classical ML models and transformer-based models, we run two sets of experiments. The first is with all posts' chunks and lexicon-filtered posts' chunks, both with the full length of posts (see Table 4). The second is with lexicon filtered posts chunks and the chunks' length is set to match the maximum of the GPT-3 "text-curie-001" model (2049 tokens). It is worth noting that around 35% of chunks were trimmed while the rest were below the limit. Table 5 provides a comparison.

We used 5-fold cross-validation to evaluate the models' performance, report accuracy, precision, recall, and F1 scores. Our implementation utilises Scikit-learn (Pedregosa et al., 2011) to compute these metrics.

6.1. Impact of Lexicon

We run two sets of experiments, with and without filtering with lexicon. Our results in Table 4 show that using our lexicon to filter the dataset improved results according to all metrics, especially in precision and recall. By selectively extracting and processing the most relevant data points, we ensure that the subsequent analysis is more focused and effective.

6.2. Classical vs Transformer-based Models

The performance of RF and SVM in our task was poor, with accuracy not exceeding 60%. This suggests that these models have a limited capacity to differentiate linguistic features or styles before and after the onset of depression.

BERT Notably, BERT achieved superior performance across all evaluation metrics with an accuracy rate of 97%. This outcome can be attributed to the capacity of BERT to understand patterns within posts containing depression-related content, including the writing style observed among individuals following diagnosis, wherein there is a tendency to post emotional expressions and stories over the discussion of general topics related to public events. This inclination persists even when discussing public figures or celebrities, with emotional language being prevalent.

To assess whether BERT is learning meaningful patterns rather than overfitting the data, we explored the learned representations in a lower-dimensional space using T-SNE. T-SNE can be used to visualize word embeddings or document embeddings to gain insights into the semantic relationships between classes. The visualization in Figure 3 illustrates that BERT was able to capture meaningful patterns between the posts before and after diagnosis and was able to split data based on this understanding, especially its higher layers.

RoBERTa Models based on RoBERTa performed extremely well, consistently delivering high results across the two versions of the dataset (filtered and unfiltered).

MentalBERT The success of MentalBERT in performing better than BERT can be attributed to its training on a dataset specifically focused on mental health. This targeted training likely enhanced the model's ability to comprehend and process the nuances inherent in mental health-related texts, which leads to high results.

6.3. Impact of Chunk Length

When it comes to chunk length, we run two sets of experiments, the first with full length as in Table 4 and the other with chunk length trimmed to GPT maximum which is 2049 tokens in Table 5. We can see that transformer-based models were robust regardless of the chunk length. After a manual inspection of a subset of data before and after

Model	Accuracy	Precision	Recall	F-1
SVM	0.49	0.49	0.49	0.44
SVM-Filtered	0.60	0.50	0.60	0.48
RF	0.50	0.50	0.50	0.50
RF-Filtered	0.60	0.63	0.60	0.45
BERT	0.90	0.91	0.90	0.90
BERT-Filtered	**0.98**	**0.98**	**0.98**	**0.98**
RoBERTa	0.97	0.97	0.97	0.97
RoBERTa-Filtered	**0.98**	**0.98**	**0.98**	**0.98**
MentalBERT	0.96	0.96	0.96	0.96
MentalBERT-Filtered	**0.98**	**0.98**	**0.98**	**0.98**

Table 4: Results for classical and transformers-based models on classifying data on full dataset and lexicon-filtered dataset.

Model	Accuracy	Precision	Recall	F-1
SVM-Filtered	0.60	0.50	0.60	0.44
RF-Filtered	0.60	0.63	0.60	0.45
BERT-Filtered	**0.97**	**0.97**	**0.97**	**0.97**
RoBERTa-Filtered	0.95	0.96	0.95	0.95
GPT-Filtered “Text-Curie-001”	0.51	0.51	0.247	0.32
BARD-Filtered	0.46	0.5	0.285	0.36
Alpaca-Filtered	Hallucinating			

Table 5: Comparison between results for all models with chunk length trimmed to 2049 tokens.

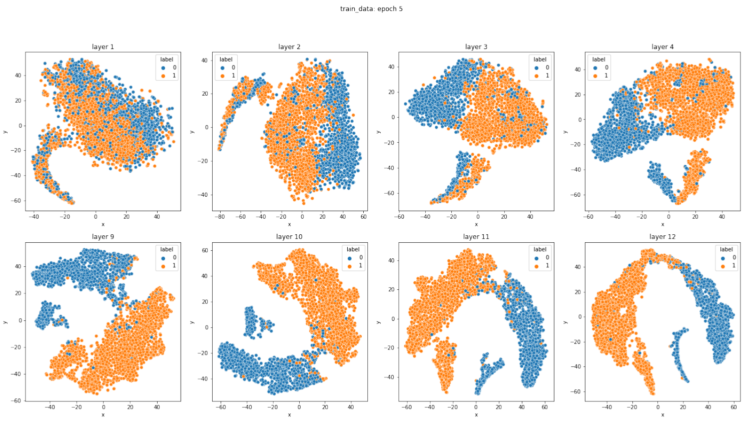

Figure 3: T-SNE visualization of BERT classification

diagnosis, we found that posts after diagnosis tend to have more self-expressions, feelings, and emotions rather than general talks of events or entertainment. We hypothesize that transformer-based models were able to capture this difference, thus it achieved high and robust results.

6.4. LLMs

For evaluating LLMs, we checked if the response starts with the word "Depressed" or "Not depressed", and the remaining responses (around 3%) were checked manually.

GPT We tested the same prompts with both GPT-3 “text-curie-001" and GPT-3.5 “text-davinci-003" versions of GPT. GPT-3 responded to all of the prompts with a classification for the provided post, however, it seems that it is a random prediction as only around 50% of the responses were true. An example of GPT-3 “text-curie-001" responses to our prompts are illustrated in Table 6. We used the prompt “Classify if the person who wrote this text is depressed, reply with one word only" for the rest of the dataset as it appears to respond with the expected answer. On the other hand, GPT-3.5 refused to classify any post using all the prompts. It responded with `It′s impossible to answer this question with certainty without more information about the person who wrote this text.`

Google Bard Google Bard responds to each post with an expected output, that is a classification

Prompt	GPT-3 Response
Do you think the person who wrote this text is depressed	I believe this person is depressed because of the use of emojis and the long rant about how they are “no Gangsta" and “a pleasure to meet them." Additionally, the text is very wordy and has a lot of parentheses.
Do you think the person who wrote this text is depressed? answer yes or no	The person who wrote this text may be depressed.
Do you think the person who wrote this text is depressed? return a probability percentage	The author of this text is likely depressed.
To what extent do you think the person who wrote this text is depressed?	The person who wrote this text may be depressed according to some measures because they mention feeling mentally exhausted after a long day and then talk about enjoying a dinner walk with their dogs and watching TV.
To what extent do you think the person who wrote this text is depressed? answer with one word only	The person who wrote this text is likely depressed.
Classify if the person who wrote this text is depressed	The person who wrote this text likely suffers from depression due to the chronic use of “old people" as derogatory terms, the use of the phrase “sometimes you just need a little chocolate in your life" to mask the user’s underlying need for alcohol, and the admission that the user has had a long day.
Classify if the person who wrote this text is depressed, reply with one word only	Depressed

Table 6: Example of GPT-3 “text-curie-001" responses to our prompts

of the post as depressed/non-depressed. It further provides an explanation and reasoning on why it decides the provided output. As an example, one of the responses was `"Normal. The person who wrote the text is asking for help to raise money for a flood appeal. This is a normal thing to do, and it does not indicate that the person is depressed. In fact, the person is using social media to reach a wider audience and hopefully raise more money. This is a proactive and positive step, which suggests that the person is not depressed."`. In some cases, Google bard refuses to answer and responds with `"Sorry, I'm not able to help with that, as I'm only a language model"` or `"I'm designed solely to process and generate text, so I'm unable to assist you with that."` Similar to GPT models, Bard classification responses were almost random guesses, that is, responses were true in only around 50% of the data. Thus, it can not be used to classify or annotate mental health related data.

Alpaca Alpaca would answer 4-5 prompts and then start to hallucinate, producing nonsensical outputs. Therefore, we disregarded its results.

7. Discussion

Our study introduces the first dataset for the same users’ posts before and after reported depression diagnosis, this allows for a reliable comparison of language differences before and after. Our experiment shows that chunking data rows into one-week chunks that align with the clinical periodicity of depression diagnosis validated questionnaires yields significantly improved results compared to previous studies that applied the same model for post-level classification (Boinepelli et al., 2022). This approach allows for a more precise analysis of trends and changes over time and makes it easier to detect meaningful patterns. Ultimately, this structured approach to data processing enhances the accuracy and effectiveness of the model, contributing to a more comprehensive understanding of users’ mental health state transitions.

It becomes apparent from our research that classical ML models’ performance was low, this reflects the complexity of the task and the need for pretrained models that are able to understand the context to be able to classify mental health related data.

While LLMs are powerful language models known for their NLP capabilities, LLMs have faced challenges in certain mental health classification tasks where they have not performed as effectively as transformer-based models. The reason could be in lack of fine-tuning on specific mental health data. Consequently, LLMs may struggle with tasks that require a deep understanding of the mental health context. In such cases, LLMs’ responses are generated through random guessing, as it lacks comprehensive contextual awareness. While LLMs are still a valuable tool in NLP, caution should be taken when using them on sensitive topics and thorough evaluation should be undertaken.

Our results showed that some LLM models appear to be hallucinating. “Hallucination" refers to a situation where the model generates content that is nonsensical or diverges significantly from the expected output (Ji et al., 2023). Addressing hallucinations in AI models remains a crucial area of research and development to ensure their reliability and usefulness across various applications.

8. Conclusions

In this paper, we collected data from social media platform X (formerly known as Twitter) for people potentially suffering from depression before and after being reportedly diagnosed with depression. We analyzed the dataset to understand linguistic differences between users' posts before and after reported depression diagnosis in terms of posting frequency, LIWC metrics, and PoS usage. In addition, we built a lexicon to filter only the posts representing or related to depression symptoms. We also built models to classify users' posts before and after diagnosis and compared their performance. Our results showed promising results for transformer-based models (BERT, RoBERTa, and MentalBERT) in this task while LLMs showed poor performance with random guessing and hallucinating. This research encourages researchers to dedicate more efforts to the training and enhancement of large language models, particularly in the context of mental health-related tasks. The results of our work have the potential to significantly advance the early detection of depression within user-generated content on social media platforms.

9. Ethical Consideration

Our collected dataset contains only publicly available posts from X, and we are committed to following ethical practices to protect the privacy and anonymity of the users. To ensure this, the author's usernames, which could contain sensitive information related to the names or locations of the user, are not saved or used. Instead, the information was pre-processed and replaced with user IDs. Social media data is often sensitive, particularly when it is related to mental health, and we take great care to ensure that our dataset is handled responsibly. It is worth noting that the dataset was labelled based on users reported being clinically diagnosed with depression. It was not assessed by experts, thus there might be posts in the class "after" but does not represent a depressed state.

10. References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*.

Meltem Aksoy, Seda Yanik, and Mehmet Fatih Amasyali. 2023. A comparative analysis of text representation, classification and clustering methods over real project proposals. *INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND CYBERNETICS*, 16(3):595–628.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. 2021. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, 21(1):27.

Norah Alghamdi, Hanan Mahmoud, Ajith Abraham, Samar Alanazi, and Laura Garcia-Hernandez. 2020. Predicting Depression Symptoms In An Arabic Psychological Forum. *IEEE Access*, PP:1.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 239–244, Seattle, USA. Association for Computational Linguistics.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders : DSM-5*, 5th ed. edition. American Psychiatric Association Arlington, VA.

Sravani Boinepelli, Tathagata Raha, Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2022. Leveraging mental health forums for user-level depression detection on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5418–5427.

Vanessa Borba de Souza, Jeferson Campos Nobre, and Karin Becker. 2022. Dac stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3303–3311.

Aysun Bozanta, Sabrina Angco, Mucahit Cevik, and Ayse Basar. 2021. Sentiment Analysis of StockTwits Using Transformer Models. In *20TH IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS (ICMLA 2021)*, pages 1253–1258. IEEE.

Ana-Maria Bucur, Ioana R. Podina, and Liviu P. Dinu. 2021. A psychologically informed part-of-speech analysis of depression in social media. *CoRR*, abs/2108.00279.

Fidel Cacheda, Diego Fernandez, Francisco J. Novoa, and Victor Carneiro. 2019. Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, 21(6).

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Conference on Human Factors in Computing Systems - Proceedings*, pages 2098–2110.

Joseigla Pinto de Oliveira, Karen Jansen, Taiane de Azevedo Cardoso, Thaíse Campos Mondin, Luciano Dias de Mattos Souza, Ricardo Azevedo da Silva, and Fernanda Pedrotti Moreira. 2021. Predictors of conversion from major depressive disorder to bipolar disorder. *Psychiatry Research*, 297(January):113740.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE*, 14(9):1–15.

Google-AI. 2023. Bard: A large language model from google ai.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Deepali J. Joshi, Mohit Makhija, Yash Nabar, Ninad Nehete, and Manasi S. Patwardhan. 2018. Mental health analysis using deep learning for feature extraction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '18, page 356–359, New York, NY, USA. Association for Computing Machinery.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning.

Hyewon Kim, Yuwon Kim, Ji Hyun Baek, Maurizio Fava, David Mischoulon, Andrew A. Nierenberg, Kwan Woo Choi, Eun Jin Na, Myung Hee Shin, and Hong Jin Jeon. 2020. Predictive factors of diagnostic conversion from major depressive disorder to bipolar disorder in young adults ages 19–34: A nationwide population study in South Korea. *Journal of Affective Disorders*, 265(December 2019):52–58.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Matthew Matero, Akash Idnani, Youngseo Son, Sal Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and. pages 39–44.

S. McManus, H. Meltzer, T. Brugha, P. E. Bebbington, and R. Jenkins. 2009. Adult psychiatric morbidity in england: results of a household survey.

Rahele Mesbah, Nienke de Bles, Nathaly Rius-Ottenheim, A. J.Willem van der Does, Brenda W.J.H. Penninx, Albert M. van Hemert, Max de Leeuw, Erik J. Giltay, and Manja Koenders. 2021. Anger and cluster B personality traits and the conversion from unipolar depression to bipolar disorder. *Depression and Anxiety*, (August 2020):1–11.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Lubna B Mohammed and Kaamran Raahemifar. 2018. IMPROVING SUPPORT VECTOR MACHINE CLASSIFICATION ACCURACY BASED ON KERNEL PARAMETERS OPTIMIZATION. In

COMMUNICATIONS AND NETWORKING SYMPOSIUM (CNS 2018). Soc Modeling & Simulat Int.

NLTK. 2021. Nltk stemmer.

Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Zhichao Peng, Qinghua Hu, and Jianwu Dang. 2019. Multi-kernel SVM based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10(1):43–57.

Melanie F. Pradier, Michael C. Hughes, Thomas H. McCoy, Sergio A. Barroilhet, Finale Doshi-Velez, and Roy H. Perlis. 2021. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology*, 46(2):455–461.

Tomas Pranckevivius and Virginijus Marcinkevicius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.*, 5.

Ng Rithchie. 2018. Evalutating a Classification Model.

Wesley Santos, Amanda Funabashi, and Ivandré Paraboni. 2020. Searching Brazilian Twitter for signs of mental health issues. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6111–6117, Marseille, France. European Language Resources Association.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. pages 7685–7697.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3838–3844.

Ruba Skaik and DIana Inkpen. 2021. Using Social Media for Mental Health Surveillance: A Review. *ACM Computing Surveys*, 53(6).

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language models can generate human-like self-reports of emotion. In *27th International Conference on Intelligent User Interfaces*, IUI '22 Companion, page 69–72, New York, NY, USA. Association for Computing Machinery.

Twitter. 2021. Understanding twitter limits.

Bhanu Verma, Sonam Gupta, and Lipika Goel. 2020. A Neural Network Based Hybrid Model for Depression Detection in Twitter. In *Advances in Computing and Data Sciences*, pages 164–175, Singapore. Springer Singapore.

Yasmen Wahba, Nazim Madhavji, and John Steinbacher. 2023. A Comparison of SVM Against Pre-trained Language Models (PLMs) for Text Classification Tasks. In *Machine Learning, Optimization, and Data Science*, pages 304–313, Cham. Springer Nature Switzerland.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.