Prompt-based mental health screening from social media text

Wesley Ramos dos Santos[1] **,**
Ivandré Paraboni[1]

[1]University of São Paulo (EACH-USP)
Av Arlindo Bettio 1000, São Paulo, Brazil

{wesley.ramos.santos,ivandre}@usp.br

Abstract. *This article presents a method for prompt-based mental health screening from a large and noisy dataset of social media text. Our method uses GPT 3.5. prompting to distinguish publications that may be more relevant to the task, and then uses a straightforward bag-of-words text classifier to predict actual user labels. Results are found to be on pair with a BERT mixture of experts classifier, and incurring only a fraction of its training costs.*

1. Introduction

Textual representation models used in NLP applications have changed dramatically in recent years, moving on from simple token counts (e.g., bag-of-words) [Pereira and Paraboni 2007] to transformer-based and large language models (LLMs) such as GPT[1] and Bloom[2], among many others. In particular, LLMs have successfully circumvented the need for labelled training data entirely, in so-called prompt-based methods that are now mainstream in the field, and have been shown to obtain higher accuracy in a wide range of tasks.

Despite its potential to reduce training costs, however, LLM prompting still requires careful consideration in applications involving large or noisy data, as it is often the case of social media. In particular, we notice that social media data may be available as a long history of publications, often stretching over thousands of posts, and yet many or most may be unrelated to the underlying task. This may be the case, for instance, when screening for user-level information such as mental health statuses (e.g., related to depression), a task that may challenge supervised and prompt-based methods alike.

Based on these observations, this work investigates the use of LLM prompting as an aid to mental health screening from social media text. Taking the issue of depression detection from Brazilian Twitter timelines as a case study, we propose a method that combines GPT prompting and standard bag-of-words classification for fast, computationally inexpensive results, and which is evaluated against SOTA results obtained by BERT mixture of experts [dos Santos et al. 2023b].

2. Related work

Depression detection may be seen as an instance of author profiling [Rangel et al. 2020, Flores et al. 2022, Pavan et al. 2023, Pavan et al. 2020] from social media text. Recent studies in the field are summarised in Table 1, categorised by text genre (Reddit, Twitter), text model (b=bag of words, BERT, e=embeddings, s=sentiment), and methods.

[1]https://platform.openai.com/docs/models
[2]https://huggingface.co/bigscience/bloom

Table 1. Depression prediction from text data.

Study	Genre	Text model	Methods
[Cohan et al. 2018]	reddit	b,e	FastText
[Aragón et al. 2019]	reddit	s	SVM
[Burdisso et al. 2020]	reddit	b	SS3
[Lin et al. 2020]	twitter	e	CNN
[Souza et al. 2020]	reddit	e	LSTM
[Souza et al. 2021]	reddit	e	LSTM+CNN
[Ansari and Ji 2022]	reddit,twitter	e,s	LR+LSTM
[dos Santos et al. 2023a]	twitter	BERT	Bi-LSTM
[dos Santos et al. 2023b]	twitter	BERT	mixture of experts

We notice a slight predominance of works based on Reddit. This may be explained by the greater ease of access and reuse of this type of data for research purposes, which is far more restricted in the case of Twitter/X. Reddit publications are taken as the basis for some of the best-known datasets available for the English language, including the SMHD [Cohan et al. 2018] and eRisk [Parapar et al. 2022] corpora.

As for the kinds of textual model under consideration, Table 1 reflects the natural evolution of the NLP and related fields, with an initial prevalence of bag-of-words models and feature engineering, and its gradual replacement by models based on word embeddings and, more recently, transformers. With regard to the computational methods used, a similar trajectory is generally observable, with the use of traditional (e.g., linear) classifiers being gradually replaced by sequence classification methods based on deep learning, including the more recent use of transformer-based architectures. In both cases - text representation and methods - we notice also that better results are usually accompanied by an increase in computational costs.

Finally, we notice that, with the exception of the work based on the *SetembroBR* [dos Santos et al. 2023a, dos Santos et al. 2020] corpus to be used in the present work, all of the above studies are devoted to the English language and, to the best of our knowledge, none make use of prompt-based methods to query a LLM for depression directly.

3. Method

As in [Cohan et al. 2018, Parapar et al. 2022] and others, our approach to depression detection relies on a dataset of social media (in our case, Twitter/X) texts that have been published by either individuals who self-reported a depression diagnosis, or by a random (control) group. Thus, the task at hand represents a binary classification problem intended to distinguish individuals who will most likely receive a depression diagnosis (called Diagnosed class) in the future from the general population (called Control class), and not to distinguish depressed from non-depressed individuals per se[3].

3.1. Data

We use of the depression portion of the SetembroBR corpus [dos Santos et al. 2023a], a collection of Twitter timelines published by Diagnosed and Control individuals in which

[3]In fact, the data does not convey any guaranteed 'non-depressed' individuals [dos Santos et al. 2023a].

only the data prior to the diagnosis date is kept. As in other language resources of this kind, the Control (i.e., random) subset is designed so as to be seven times larger than the Diagnosed group, making a heavily imbalanced classification task. Table 2 presents data descriptive statistics.

Table 2. SetembroBR depression corpus descriptive statistics.

Statistics	Diagnosed	Control	Overall
Users (timelines)	1,684	11,788	13,472
Words (million)	29.32	201.94	231,26
Publications (million)	2.43	16.99	19,42

In the present work, we follow the standard train/test split provided by the corpus, as described in [dos Santos et al. 2023a].

3.2. Approach

Screening for depression from social media gives rise to the questions of how to handle a large number of publications, most of which unlikely to be relevant to the task. To this end, we envisaged a prompt-based approach called *Prompt.Bow* that relies on GPT 3.5 to enrich an otherwise standard text classifier. These two steps - prompting and classification - are described individually as follows.

First as a means to identify messages that are potentially related to mental health, we use GPT 3.5. prompting to assess a random sample of 30,000 tweets. The (English-translated) prompt in question, adapted from the clinical description of depression in [American Psychiatric Association 2013], is shown in Figure 1.

Considering tweet X below, which of the following three options would be most likely?

Option 1: Tweet X has strong indications that the individual who wrote it may be suffering from some type of depression or anxiety disorder:

1.a - This may occur because the tweet explicitly mentions intense feelings of depression, despair, intense anxiety, or related symptoms.

1.b - The tweet strongly and explicitly suggests that the user is experiencing high levels of depression or anxiety, even if it is not explicitly stated.

Option 2: Tweet X has moderate indications that the individual who wrote it may be suffering from some type of depression or anxiety disorder:

2.a - This may be because the tweet mentions feelings of depression, anxiety, stress, or related symptoms, but the indicators are not as strong as in the `high' category.

2.b - The tweet indirectly suggests that the user is experiencing low levels of depression or anxiety based on the content, even if it is not explicitly stated.

Option 3: Tweet X has little or no indication that the individual who wrote it may be suffering from any type of depression or anxiety disorder:

3.a - Messages that have no relation to the topic.

3.b - Messages that use language in a colloquial manner without indicating that it has a medical basis.

Return only the option value (1, 2 or 3).

Tweet X is {*tweet text here*}

Figure 1. Prompt instruction to GPT.

By submitting this prompt to each of the 30,000 sample tweets, the data was categorised as having high (1), medium (2) or low (3) relevance to mental health. These GPT-labelled data was then taken as an input to train a T5 classifier [Raffel et al. 2020] to label the entire 19.42-million tweets in the corpus. Table shows the label distribution across the training portion of the data.

Table 3. Training data distribution according to relevance for depression. Tweets and tokens are shown in thousand units.

	Diagnosed class				Control class			
Relevance	Tweets	%	Tokens	%	Tweets	%	Tokens	%
high	39	2.0%	674	5.4%	153	1.1%	2,372	3.0%
medium	180	9.3%	3,663	29.4%	1,031	7.6%	20,566	25.7%
low	1,727	88.7%	8,122	65.2%	12,439	91.3%	56,971	71.3%

As expected, most publications (65.2% in the Diagnosed class, and 71.3% in the Control class) are deemed of low relevance for mental health prediction according to the prompted model. On the other hand, highly relevant publications are relatively rare (5.4% in the Diagnosed class, and 3.0% in the Control class).

More importantly, however, when prompting an LLM for mental health we are largely focusing on semantics, that is, on symptoms and other well-known clinical signs of depression. For instance, our current prompt allows the model to pinpoint a wide range of publications that may suggest, e.g., eating disorders or negative language use, both of which known to be related to depression, cf. [American Psychiatric Association 2013], but this is not to say that other factors can or should be overlooked.

In particular, we notice that LLM prompting may not explicitly account for more fine-grained linguistic indicators of depression such as the use of first person pronouns [Trifu et al. 2017], absolute terms [Al-Mosaiwi and Johnstone 2018] and other lexical factors (e.g., denoting emotion as in hate speech, cf. [da Silva et al. 2020]). These indicators, which may be present in any publication of low or high relevance to depression alike, are also important predictors of mental health statuses, and no data point should in principle be discarded solely based on the LLM output.

As a means to keep the full train data available to the classifier whilst distinguishing between more and less relevant messages (which clearly show different distributions across classes in Table 3), *Prompt.Bow* uses the category labels provided by the LLM to split the training data into high, medium, and low relevance subsets, from which we create three individual bag-of-words vectors (to be concatenated as discussed below). In other words, the training texts are split into three (low/medium/high relevance) categories according to the previous GPT prompt method, and we build an individual BoW model from each of these three subsets.

The resulting vectors are further reduced using univariate feature selection using F1 as a score function. The final 'high' and 'medium' vectors were reduced to $k = 6{,}000$ features each, and the final 'low' vector was reduced to $k = 3{,}000$ features, all of which concatenated as a single 15,000-word vector.

In addition to distinguishing between high-, medium- and low-relevance messages in this way, we use the information provided by the LLM also to help capture message

order, the underlying assumption being that certain patterns (e.g., a series of consecutive 'highly relevant' messages) may be indicative of depression. To this end, we created a bigram model of high/medium/low labels only, that is, representing sequences of 'high', 'medium' and 'low' labels only (and not text). This was further reduced to its $k = 40{,}000$ most relevant (bigram) features by performing F1 univariate feature selection, and then appended to the previous 15,000-word vector.

The resulting vector - a combination of three text models with different degrees of relevance to mental health, and a model that captures sequences of relevant messages - is taken as the input to a standard logistic regression classifier. This choice is motivated by the observation that much of the deep (e.g., semantics-driven) language processing had already been performed by the LLM, and that a simple text classifier should suffice for user label prediction.

4. Evaluation

Table 4 presents the results of the present *Prompt.Bow* model alongside the results reported in [dos Santos et al. 2023b] for BERT mixture of experts using BERTabaporu [da Costa et al. 2023], which currently stands as the SOTA for the present setting. In both cases, results are based on the standard test portion of the corpus.

Table 4. Classification results.

Model	Precision	Recall	F1
Prompt.BoW	0.64	0.72	0.66
BERT.MoE	0.64	0.67	0.65

Table 4 shows that results remain close, with a small advantage for the *Prompt.Bow* approach over BERT mixture of experts. However, by relying on an input provided by the pre-trained LLM, we notice that these results were obtained by using a computationally inexpensive classifier model (that is, leaving aside the costs of pre-training the LLM in the first place), which represents a stark contrasts to computationally-intensive BERT mixture of experts supervision.

5. Final remarks

This work presented an experiment in prompt-based mental health screening from a large, noisy corpus of social media publications that relies on LLM prompting to distinguish publications that may be more relevant to the task, and then uses a straightforward bag-of-words text classifier to predict actual user labels. This was shown to obtain competitive results if compared to a BERT-based classifier architecture that represents the SOTA in the present setting, but incurring only a fraction of its training costs.

As future work, we intend to refine the present method by using the formal definition of depression in [American Psychiatric Association 2013] as a prompt, and by fully integrating symptoms and linguistic indicators within a neural architecture for both depression and anxiety disorder prediction.

6. Acknowledgements

The present work has been financed by the São Paulo Research Foundation (FAPESP grant #2021/08213-0). The first author has been supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 - grant # 88887.475847/2020-00. This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation.

References

Al-Mosaiwi, M. and Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders 5th edition*. American Psychiatric Association, Arlington, VA.

Ansari, L. and Ji, S. (2022). Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on computational Social Systems*.

Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C., and y Gómez, M. M. (2019). Detecting depression in social media using fine-grained emotions. In *NAACL-2019 Proceedings*, pages 1481–1486, Minneapolis, USA. Assoc for Comp Ling.

Burdisso, S. G., Errecalde, M., and y Gómez, M. M. (2020). t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams. *Pattern Recognition Letters*, 138:130–137.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and v Goharian (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING-2018*, pages 1485–1497, Santa Fe, USA.

da Costa, P. B., Pavan, M. C., dos Santos, W. R., da Silva, S. C., and Paraboni, I. (2023). BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recents Advances in Natural Language Processing (RANLP-2023)*, pages 217–223.

da Silva, S. C., Ferreira, T. C., Ramos, R. M. S., and Paraboni, I. (2020). Data driven and psycholinguistics motivated approaches to hate speech detection. *Computación y Systemas*, 24(3):1179–1188.

dos Santos, W. R., de Oliveira, R. L., and Paraboni, I. (2023a). SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*.

dos Santos, W. R., Funabashi, A. M. M., and Paraboni, I. (2020). Searching Brazilian Twitter for signs of mental health issues. In *12th International Conference on Language Resources and Evaluation (LREC-2020)*, pages 6113–6119, Marseille, France.

dos Santos, W. R., Yoon, S., and Paraboni, I. (2023b). Mental health prediction from social media text using mixture of experts. *IEEE Latin America Tr.*, 21(6):723–729.

Flores, A. M., Pavan, M. C., and Paraboni, I. (2022). User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58(1):67–89.

Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., and Leung, H. (2020). *SenseMood: Depression Detection on Social Media*, pages 407–411. Association for Computing Machinery, New York, USA.

Parapar, J., Martin-Rodilla, P., Losada, D. E., and Crestani, F. (2022). Overview of eRisk 2022: Early Risk Prediction on the Internet. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, pages 821–850, Bologna, Italy.

Pavan, M. C., dos Santos, V. G., Lan, A. G. J., ao Trevisan Martins, J., dos Santos, W. R., Deutsch, C., da Costa, P. B., Hsieh, F. C., and Paraboni, I. (2023). Morality classification in natural language text. *IEEE transactions on Affective Computing*, 14(1):857–863.

Pavan, M. C., dos Santos, W. R., and Paraboni, I. (2020). Twitter Moral Stance Classification using Long Short-Term Memory Networks. In *9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319*, pages 636–647. Springer.

Pereira, D. B. and Paraboni, I. (2007). A language modelling tool for statistical NLP. In *5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC*, pages 1679–1688, Rio de Janeiro. Sociedade Brasileira de Computação.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rangel, F., Rosso, P., Zaghouani, W., and Charfi, A. (2020). Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, page 1–21.

Souza, V., Nobre, J., and Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *SBBD-2020*, pages 121–132, Porto Alegre, Brazil. SBC.

Souza, V., Nobre, J., and Becker, K. (2021). A deep learning ensemble to classify anxiety, depression, and their comorbidity from texts of social networks. *Journal of Information and Data Management*, 12(3):306–325.

Trifu, R., Nemes, B., Bodea-Hategan, C., and Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies*, 17:105–128.