A Bag-of-Users approach to mental health prediction from social media data

Rafael Lage de Oliveira and **Ivandré Paraboni**
University of São Paulo (EACH-USP)
Av Arlindo Bettio 1000, São Paulo, Brazil
rlagedo@gmail.com,ivandre@usp.br

Abstract

Computational models of mental health prediction from social media data are typically built from the textual contents produced by the individuals to be assessed, but the use of non-textual information available from the network structure may also have relevant predictive power. Based on these observations, this work presents initial experiments on mental health prediction from textual and non-textual Twitter data in Portuguese, comparing a traditional content-based approach using BERT with models based on social media connections (friends, followers and mentions), and which is inspired from the well-known Bag-of-Words text representation. Results highlight an advantage for the model based on textual contents, but also suggest that the use of non-textual information may provide a significant contribution to these tasks.

1 Introduction

The detection of mental health disorders such as depression and anxiety from social media data is a current application of great social interest, and has been the focus of a wide range of recent studies in NLP and related fields (Lin et al., 2020; Chancellor and Choudhury, 2020; Su et al., 2020; Parapar et al., 2023). Computational models of this kind are typically built from the textual content produced by the individuals to be assessed (e.g., social media users) and, although user-generated text is possibly the richest source of information for tasks of this kind, the use of non-textual information available from the network structure (e.g., connections between users) may also have relevant predictive power (Cheng and Chen, 2022; Zogan et al., 2022b) according to the principle of *homophily* (McPherson et al., 2001), i.e., the tendency of users with similar interests establish connections.

Using non-textual social media information as learning features for mental heath prediction may however pose a number of challenges. In particular, although the number of social media connections may be large (e.g., a typical Twitter user may have thousands of friends and followers), hence suggesting a potentially rich source of information, it remains unclear how often we will find a connection between, e.g., two depressed individuals. For instance, in the SetembroBR depression and anxiety disorder corpus (dos Santos et al., 2023), 3903 individuals were randomly sampled from a large pool of Portuguese-speaking Twitter users based on their diagnoses and, crucially, these individuals are largely unrelated, that is, they do not make a connected community.

When social media users are unacquainted to each other, building meaningful graph representations may not be possible, and the use of established social network measures (e.g., of distance between nodes) for prediction purposes may become unhelpful. However, this is not to say that social media connections *to other individuals* (i.e., users not represented in the corpus) are unhelpful as well. On the contrary, homophily suggests that some of these individuals may be prone to interacting with particular users or accounts (e.g., a discussion forum on mental health issues, a celebrity known for having disclosed their mental health struggle, etc.), and this information may be predictive of mental health statuses alongside more traditional (i.e., user-generated) information.

One possible way of using non-textual information for mental health prediction when a fully connected network is unavailable is by regarding social media connections not as relations between the individuals represented in the corpus, but rather as *atomic properties* of these individual. More specifically, the information that a user u follows, e.g., a Twitter account that promotes information on mental health, may be regarded as a learning feature to help classify u as being depressed or not, and this may be implemented, for instance, by modelling social media relations as sets of connections.

Based on these observations, this work presents an initial study of depression and anxiety disorder prediction in the Portuguese language from textual and non-textual data alike. Using the aforementioned SetembroBR corpus as a basis, we compare a traditional content-based approach built from pre-trained BERT (Devlin et al., 2019) with models solely based on social media connections in a so-called *Bag-of-Users* approach. In doing so, the objective of the study is to compare the two types of strategy, which may be seen as a first step towards the development of multimodal predictive models for these tasks.

The rest of this paper is structured as follows. Section 2 reviews existing work in mental health prediction from mutimodal social media data. Section 3 introduces our present models for depression and anxiety disorder prediction in Portuguese. Section 4 describes our main experiments. Section 5 draws our conclusions and discusses future work.

2 Related work

Table 1 summarises recent studies in mental health prediction based on multimodal social media data. These studies are categorised by task (A=anxiety, D=depression), genre (In=Instagram, Fb=Facebook, Fl=Flicker, Sw=Sina Weibo, Tw=Twitter), language (Ch=Chinese, En=English), textual features (bow=Bag-of-Words, we=word embeddings, lex=lexicon, LIWC (Pennebaker et al., 2001), st=sentiment), non-textual features (ti=time, pc=posts, mc=mentions, rt=reposts, rc=replies, lc=likes, ac=friends, fc=followers, cc=comments, vc=views, nm=other).

Among the selected studies, we notice that one of our target applications – depression prediction – is common in the field, but the second – anxiety disorder prediction – was only addressed from a multimodal perspective in Mendu et al. (2020). Regarding the type of social media under consideration, we notice that the use of microblog data from Twitter and Sina Weibo prevails. Moreover, all identified studies are dedicated to either English or Chinese languages.

Although the use of word embeddings as a textual representation is common, we notice that simpler strategies based on Bag-of-Words or LIWC lexical category counts are also popular. This may be explained by the observation that many of the existing studies are more focused on the use of network-related features, and that in many of these studies the text model tends to take second place. Furthermore, representations of this kind may simplify the combination of textual and non-textual features (e.g., by vector concatenation) than would otherwise be the case if, for instance, using word embeddings sequences.

Regarding the kinds of non-textual features under consideration, we notice that these are largely based on user counts (e.g., number of friends, etc.). Structural information, however, does appear in two studies (Sinha et al., 2019; Ruch, 2020) dedicated to the related issues of detecting symptoms of depression and suicidal thoughts, which were not part of the present survey.

3 Models

We envisaged an experiment to compare the use of textual and non-textual features in mental health prediction using Portuguese social media data. To this end, textual features were computed using a pre-trained BERT (Devlin et al., 2019) language model, and non-textual features correspond to social media connections represented by relationships with Twitter friends and followers, and @ mentions of other users.

All models were built from the SetembroBR corpus (dos Santos et al., 2020, 2023) of Twitter timelines (i.e., lists of timestamped text publications), divided into two classes: those produced by individuals who have been diagnosed with depression or anxiety disorder (hereby referred to as the 'Diagnosed' class), and a seven times larger group of random individuals (hereby 'Control' group)[1]. In this setting, every Diagnosed user is paired with its seven Control counterparts according to gender[2], timeline length and publication dates.

The corpus conveys 46.8 million tweets written in Portuguese by 18,819 unique users, and their sets of friends, followers, and mentions. Table 2 presents descriptive statistics of the textual and non-textual portions of the data, showing the mean number of connections (friends, followers and mentions) on the top, and mean text statistics (number of timelines, tweets and tokens) at the bottom.

For the textual models, in which case the task may be seen as an instance of Portuguese text au-

[1] A similarly heavy class imbalance, intended to help distinguish diagnosed from random individuals, is adopted in Yates et al. (2017); Losada et al. (2017); Cohan et al. (2018).

[2] Estimated by the linguistic gender expressed in text, as in, e.g., Paraboni and de Lima (1998).

Model	Task	Genre	Lang.	Textual	Non-textual
(Yang et al., 2020)	D	Fb	En	LIWC	ti,pc,ac
(Wu et al., 2020)	D	Fb	Ch	we	ti,pc
(Mendu et al., 2020)	A	Fb	En	bow,LIWC	ti,nm
(Xu et al., 2020)	D	Fl	En	bow,LIWC	ti,vc
(Alsagri and Ykhlef, 2020)	D	Tw	En	bow	ti,pc,mc,rc
(Ghosh and Anwar, 2021)	D	Tw	En	LIWC,st	ti,pc,cc,rt
(Zogan et al., 2021)	D	Tw	En	we,lex	ti,pc,rt,ac,fc
(Bi et al., 2021)	D	Sw	Ch	bow,LIWC,lex	fc,ac,lc,cc,rt
(Cheng and Chen, 2022)	D	In	Ch	we	ti
(Zogan et al., 2022a)	D	Tw	En	we,LDA	ti,pc,rt,ac,fc

Table 1: Existing work using non-textual features for mental health prediction.

Statistics	Depres.	Ctrl	Anxiety	Ctrl
Friends	659	710	678	729
Followers	777	945	810	975
Mentions	125	122	115	114
Timelines	1,684	11,788	2,219	15,533
Tweets (mi)	2.43	16.99	3.43	23.98
Tokens (mi)	29.32	201.94	42.24	281.51

Table 2: SetembroBR descriptive statistics, taken from dos Santos et al. (2023).

thor profiling (da Silva et al., 2020; Flores et al., 2022; Pavan et al., 2023), we used the BERT approach introduced in dos Santos et al. (2023). This consists of the Portuguese Twitter BERT model in da Costa et al. (2023), which has been presently fine-tuned for the tasks at hand. In this approach, user timelines are classified in batches of 10 consecutive tweets each, and the class label (to be associated with the timeline under analysis) is decided by majority vote. The model architecture consists of a BiLSTM network with ReLU activation function followed by a fully connected layer with softmax activation and using a cross entropy type loss function with balanced class weights. The model is trained in up to three epochs and the input messages are truncated to 30 tokens.

For the non-textual models, connections between users of the corpus and their friends, followers and mentions of other network users were represented as binary 'Bag-of-Users' models indicating whether each individual in the corpus had a relationship with others, mostly not represented in the corpus. A fragment of this representation is illustrated as follows, showing three corpus users (who may belong to the Diagnosed or Control class), and some of their friendship relations.

	Friend 1	Friend 2	...	Friend N
User 1	1	0	...	0
User 2	0	0	...	0
User 3	0	0	...	1

As in a conventional (i.e., textual) Bag-of-Words approach, this representation is highly sparse, with approximately one million possible connections (or dimensions), but a very low number of actual connections per user. Thus, we initially attempted to select only the 15 thousand users with the highest number of connections for each (Diagnosed and Control) class, but even with this pruning the representation of friends, followers and mentions was still highly sparse. For that reason, a second feature selection method was used, once again inspired by techniques normally used in text pre-processing.

We performed univariate feature selection over a development portion of the training data using F1 as a score function to select the K most relevant characteristics (i.e., connections) for each of the three non-textual models. More specifically, candidate K values were attempted based on the maximum number of connections available in each of the three (friends, followers and mentions) networks, with 500-unit decreases until identifying the K value that maximised the F1 measure. The optimal K values obtained for the depression (D) and anxiety (A) prediction tasks are summarised in Table 3.

Model	Depression	Anxiety
Friends	14,500	17,000
Followers	13,000	21,000
Mentions	19,500	10,500

Table 3: Non-textual model K values.

	Depression			Control			Anxiety			Control		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	0.34	0.49	**0.40**	0.92	0.87	0.89	0.36	0.36	**0.36**	0.91	0.91	0.91
Friends	0.25	0.44	0.32	0.92	0.82	0.86	0.23	0.43	0.30	0.91	0.80	0.85
Followers	0.22	0.60	0.32	0.92	0.69	0.79	0.20	0.50	0.29	0.91	0.72	0.80
Mentions	0.36	0.32	0.34	0.90	0.92	0.91	0.30	0.31	0.30	0.90	0.90	0.90

Table 4: Main results. Best F1 scores for the positive class in each task are highlighted.

4 Evaluation

From the fixed training and test portions of the SetembroBR corpus data described in (dos Santos et al., 2023), we built and evaluated both BERT and Bag-of-User models. Results are summarised in Table 4.

For both tasks, results suggest that the BERT textual model is still superior to the non-textual alternatives. However, we notice that the difference may in some cases be considered small, particularly if one takes into account the computational cost involved in building these models, which is vastly superior in the case of BERT. Moreover, the observation that learning features that do not rely upon user-generated contents have considerable predictive power is a useful insight in its own right.

5 Final remarks

This study presented initial experiments on mental health prediction from social media data in Portuguese using on textual and non-textual data alike, and focusing on settings in which the available social media users are in principle unacquainted to each other, in which case standard network-related metrics or models may be unhelpful.

As an alternative to these methods, a so-called Bag-of-Users approach, analogous to a simple count-based text model, was presented. Although results obtained from this method still point to the advantage of the model based on textual content using BERT, the use of non-textual information in this way also presents a potentially useful contribution, and suggests that the combination of the two strategies (for example, with the use of ensemble methods) may improve current results.

Thus, in addition to an investigation on how to combine textual and non-textual data into a single model, as future work we also envisage improving the representation of non-textual models using more expressive network embeddings representations, such as those computed by using node2vec (Grover and Leskovec, 2016) and related methods.

6 Acknowledgements

The present research has been supported by the São Paulo Research Foundation (FAPESP grant #2021/08213-0).

References

Hatoon S. Alsagri and Mourad Ykhlef. 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, E103D(8):1825 – 1832.

Yanting Bi, Bing Li, and Hongzhe Wang. 2021. Detecting depression on sina microblog using depressing domain lexicon. In *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 965–970.

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

Ju Chun Cheng and Arbee L. P. Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, 59(2):319–339.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, 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. Assoc for Comp Ling.

Pablo Botton da Costa, Matheus Camasmie Pavan, Wesley Ramos dos Santos, Samuel Caetano da Silva, and Ivandré Paraboni. 2023. BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recents Advances in Natural Language Processing (RANLP-2023)*, pages 217–223, Varna, Bulgaria.

Samuel Caetano da Silva, Thiago Castro Ferreira, Ricelli Moreira Silva Ramos, and Ivandré Paraboni. 2020. Data driven and psycholinguistics motivated approaches to hate speech detection. *Computación y Systemas*, 24(3):1179–1188.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 Proceedings*, pages 4171–4186, Minneapolis, USA.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*.

Wesley Ramos dos Santos, Amanda Maria Martins Funabashi, and Ivandré Paraboni. 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. ELRA.

Arthur Marçal Flores, Matheus Camasmie Pavan, and Ivandré Paraboni. 2022. User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58(1):67–89.

Shreya Ghosh and Tarique Anwar. 2021. Depression intensity estimation via social media: A deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6):1465 – 1474.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *KDD16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, San Francisco, USA. Association for Computing Machinery.

Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. *SenseMood: Depression Detection on Social Media*, pages 407–411. Association for Computing Machinery, New York, USA.

David E. Losada, Fabio Crestani, and Javier Parapar. 2017. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *LNCS 10456*, pages 346–360, Cham. Springer.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Sanjana Mendu, Anna Baglione, Sonia Baee, Congyu Wu, Brandon Ng, Adi Shaked, Gerald Clore, Mehdi Boukhechba, and Laura Barnes. 2020. A framework for understanding the relationship between social media discourse and mental health. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Ivandré Paraboni and Vera Lucia Strube de Lima. 1998. Possessive pronominal anaphor resolution in Portuguese written texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1010–1014. Assoc for Comp Ling.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. eRisk 2023: Depression, Pathological Gambling, and Eating Disorder Challenges. In *Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science, vol 13982*, Cham. Springer.

Matheus Camasmie Pavan, Vitor Garcia dos Santos, Alex Gwo Jen Lan, Jo ao Trevisan Martins, Wesley Ramos dos Santos, Caio Deutsch, Pablo Botton da Costa, Fernando Chiu Hsieh, and Ivandré Paraboni. 2023. Morality classification in natural language text. *IEEE transactions on Affective Computing*, 14(1):857–863.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Alexander Ruch. 2020. Can x2vec save lives? integrating graph and language embeddings for automatic mental health classification. *Journal of Physics: Complexity*, 1(3).

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. #suicidal - a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 941–950, New York, NY, USA. Association for Computing Machinery.

Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(116).

Min Wu, Chih-Ya Shen, En Tzu Wang, and Arbee Chen. 2020. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54.

Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Inferring social media users' mental health status from multimodal information. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6292–6299. European Language Resources Association.

Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. 2020. A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54:102141.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of EMNLP-2017*, pages 2968–2978, Copenhagen, Denmark. Assoc for Comp Ling.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: Learning multimodalities with user post summarization for depression detection on social media. In *44th International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, pages 133–142, New York, USA. Association for Computing Machinery.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022a. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281 – 304.

Hamad Zogan, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022b. Depression detection with multi-modalities using a hybrid deep learning model on social media. *World wide web*, 25(1):281–304.