BERTabaporu: Assessing a Genre-specific Language Model for Portuguese NLP

Pablo da Costa **Matheus Pavan** **Wesley dos Santos** **Samuel da Silva** **Ivandré Paraboni**
School of Arts, Sciences and Humanities
University of São Paulo
Av Arlindo Bettio 1000, São Paulo, Brazil
{pablo.costa,matheus.pavan,wesley.ramos.santos,samuel.caetano.silva,ivandre}@usp.br

Abstract

Transformer-based language models such as Bidirectional Encoder Representations from Transformers (BERT) are now mainstream in the NLP field, but extensions to languages other than English, to new domains and/or to more specific text genres are still in demand. In this paper we introduced BERTabaporu, a BERT language model that has been pre-trained on Twitter data in the Brazilian Portuguese language. The model is shown to outperform the best-known general-purpose model for this language in three Twitter-related NLP tasks, making a potentially useful resource for Portuguese NLP in general.

1 Introduction

Transformer-based language models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) are now mainstream in the NLP field, but extensions are still much in demand. New BERT models have been fine-tuned or built from scratch for many languages other than English (e.g., Maltese in (Micallef et al., 2022)), for specific domains (e.g., mental health in (Ji et al., 2022)), and for particular text genres (e.g., Twitter data in (Nguyen et al., 2020)).

In the case of our target language - Brazilian Portuguese - the first and best-known representative of this trend is BERTimbau, a general purpose BERT model built from a large collection of web documents (Souza et al., 2020). Despite its popularity among the Portuguese NLP community, however, we notice that the particular kind of language employed on contemporary social media may be distinct from the training data considered in previous work, making existing models potentially less suitable to handle recent social media text. In particular, we notice that tweets are not only shorter and less structured than pieces of news, but words such as 'Covid' may not be recognised by older language models.

Based on these observations, we may ask whether Twitter-related applications may benefit from a more genre-specific model built from social media data - as opposed to more standard or general text - perhaps along the lines of BERTweet, a Twitter-specific model for the English language described in (Nguyen et al., 2020), or the multilingual TwHIN-BERT (Zhang et al., 2022), also built from Twitter data. To shed light on this issue, this work introduces BERTabaporu, a BERT model built from a collection of 238 million tweets written by over 100 thousand unique Twitter users, and conveying over 2.9 billion tokens in total. The model has been evaluated in three Twitter-related text classification tasks, and its results are compared to those obtained by the general purpose, web-based BERTimbau in (Souza et al., 2020). In doing so, our goal is to introduce a novel resource for Portuguese NLP, and foster further applications based on Twitter text data in this language.

The main contributions made in this work are (i) a novel BERT model trained on large Twitter corpus in the Portuguese language; and (ii) comparison with the best-known existing Portuguese BERT in three Twitter text classification tasks, namely, stance, mental health and political alignment prediction.

The rest of this article is structured as follows. Section 2 reviews existing work that have introduced BERT-like models for Portuguese, for other languages and domains. Section 3 describes how our current work, the BERTabaporu model, has been built. Section 4 introduces the downstream tasks in which BERTabaporu is to be assessed. Section 5 reports results obtained for each of the evaluation tasks, and compares these to the results obtained by existing work. Finally, section 6 draws our conclusions and suggestions of future work.

Proceedings of Recent Advances in Natural Language Processing, pages 217–223
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_024

2 Background

Since the original English and multilingual (or mBERT) models described in (Devlin et al., 2019), similar resources devoted to these and to dozens of other languages have been created. Models of this kind are either trained from scratch or fine-tuned (often from mBERT) using either general purpose or domain- or genre-specific text data. Noteworthy examples include BERTweet (Nguyen et al., 2020), a 16-billion token model built from Twitter data in the English language, domain-specific models fine-tuned for mental health (Ji et al., 2022), abusive language use (Caselli et al., 2021), biomedical texts (Schneider et al., 2020) and others, alongside general purpose models for a wide range of languages including, e.g., Estonian (Tanvir et al., 2021), Maltese (Micallef et al., 2022), Romanian (Masala et al., 2020) and Czech (Sido et al., 2021).

In the case of the Portuguese language, although model repositories such as Hugging Face provide a considerable number of BERT models - particularly for the Legal domain, and even including a few models trained on Twitter data - we have identified only three models of this kind that are more fully documented in the NLP literature. These are summarised in Table 1 and further discussed below.

Two of the existing Portuguese models - BioBERTpt and PetroBERT - are fine-tuned to perform domain-specific tasks in the clinical/biomedical and oil and gas industry domains, respectively. This makes web-based BERTimbau (Souza et al., 2020) the more closely related alternative to our own work (BERTabaporu, on the top row of the table). BERTimbau is the first, and arguably the best-known Portuguese BERT model to date, and it has been trained from scratch using a general purpose web corpus in which great care has been taken to minimise the effects of duplicate data, making it a suitable baseline to our current work.

Based on these observations, our present work BERTabaporu may be seen as a general purpose, Twitter-specific alternative to BERTimbau built from a slightly larger dataset (2.9 billion tokens against 2.7 billion in BERTimbau), and which should be able to outperform the existing web-based model in Twitter-oriented tasks.

3 The BERTabaporu model

In order to gather unlabelled text data to built BERTabaporu, we selected a number of existing tweet repositories and collected additional data online. As a means to minimise the effect of duplicated training data, BERTabaporu has been built from tweets originally posted by over 100 thousand unique Twitter users excluding their retweets. In this user-centred method, although some data duplication may still occur (namely, if an individual rewrites the same text that another user has authored), we assume that the effect of duplicates is likely to be small.

Users were selected from a number of pseudo-random tweet sources based on a number of seed topics, such as Covid-19, politics, mental health and others, and then their entire public timelines were collected regardless of the topics under discussion. Thus, it should be clear that the data is by no means limited to these topics, and that we did not search for *individual tweets* about any particular topic, but rather used the seed topics as a guideline to identify *users* timelines, and then collect all their publications (which will inevitably discuss a very broad range of subjects besides the seed topic.) In other words, our data consists of a collection of pseudo-random user timelines, and not a collection of tweets about the seed topics, and should not be seen as being significantly biased towards any particular topic.

Selected user timelines comprised three main categories: (i) timelines of random users (about 33%); (ii) timelines of users who discussed Covid-19, politics, mental health issues, vaccines or other Covid-19 measures at least once (about 46%); and timelines of friends with whom these users most frequently interact (about 21%).

From the selected timelines, all non-Portuguese tweets were removed. From the remainder, emoticons, non-alphabetic characters, URLs and usernames were removed, and numbers were replaced by '1'. Finally, timelines conveying fewer than 80 tweets were discarded. Table 2 summarised descriptive statistics of our training dataset.

From the above unlabelled data, we pre-trained a monolingual BERT model from scratch using both BERT-BASE and BERT-LARGE architectures. The base version uses 12 transformer layers, a hidden size of 768, and 8 attention heads. The large version uses 24 transformer layers, a hidden size of 1024, and 16 attention heads. In both cases, the vocabulary is initialised with 64K tokens. Pre-training is performed across 1M steps, with a sequence length of 128 for the first 90% of the steps

Model	Domain	Text genre	Tokens	Training
BERTabaporu (ours)	general	Twitter	2.9 bi	from scratch
BERTimbau (Souza et al., 2020)	general	web	2.7 bi	from scratch
BioBERTpt (Schneider et al., 2020)	clinical/biomed.	notes, abstracts	44.1 mi	fine-tuned
PetroBERT (Rodrigues et al., 2022)	oil and gas	notes, reports, theses	na	fine-tuned

Table 1: Documented pre-trained BERT models devoted to the Portuguese language.

User source	Timelines	%	Tweets (th)	Sentences (th)	Tokens (th)
Random	32,879	32.6 %	102,489	113,183	1,111,397
Covid-19	9,021	9.0 %	18,384	21,945	233,968
Politics	5,767	5.7 %	12,416	16,614	155,380
Mental health	3,790	3.8 %	8,653	9,541	100,734
Vaccine	27,861	27.7 %	57,913	83,048	898,287
Friends' timelines	21,369	21.2 %	38,044	44,154	436,929
Overall	100,687	100.0 %	237,899	288,485	2,936,697

Table 2: BERTabaporu training data descriptive statistics.

and a sequence length of 512 for the remaining 10% steps. The models use a batch size of 512, and a warm-up of 1% of the total number of steps. Training was performed on v2-8 TPUs, taking approximately 120 hours for both configurations. The resulting language model is publicly available for reuse[1].

4 Evaluation

We envisaged a number of Twitter text classification experiments to compare our current Twitter BERTabaporu model with the general-purpose alternative in (Souza et al., 2020). In doing so, we would like to show that genre-specific BERTabaporu obtains superior results in these evaluation scenarios.

4.1 Downstream evaluation tasks

Evaluation will focus on three downstream tasks - stance, mental health statuses and political alignment prediction - all of which modelled as binary classification tasks, and based on Twitter text data in the Portuguese language. Two of these tasks - stance and political alignment prediction - consist of classifying individual tweets, whereas mental health prediction consists of classifying Twitter users (or rather, the sets of tweets published on their Twitter timelines.) The choice of these tasks is intended to provide variation in input definition (i.e., individual tweets versus entire timelines), in the degree of explicitness of class labels (e.g., learning the stance explicitly expressed in text versus the implicit political leaning of its author), and in corpus labelling methods (tweet- and user-level annotation, or label propagation) as discussed in the next section.

Stance prediction is the computational task of inferring an attitude in favour or against a set target topic (Mohammad et al., 2016; dos Santos and Paraboni, 2019). For instance, '*A universal basic income would alleviate poverty*' conveys a stance in favour of the target 'universal basic income'. The task is analogous to sentiment analysis, but stance and sentiment are not necessarily correlated (Aldayel and Magdy, 2021; Pavan et al., 2020). In our current setting, we focus on six stance prediction tasks based on targets that have been popular discussion topics on Brazil social media (Brazilian presidents, Covid-related measures, and local institutions.) In these tasks, given an input tweet known to convey a stance towards a particular target, the goal is to decide whether this represents a stance in favour or against it.

Mental health statuses prediction consists of determining whether an individual is prone to a mental health disorder based on their publications, e.g., on social media. Computational models of this kind have been popular in the NLP field (Shen et al., 2017; Losada et al., 2017; Cohan et al., 2018) under multiple task definitions. These include, for instance, deciding whether an individual is depressed or not (Yazdavar et al., 2020), measuring the degree of severity of the underlying disorder (Mann

[1] https://huggingface.co/pablocosta/bertabaporu-large-uncased

et al., 2020), symptoms detection (Yazdavar et al., 2017), and others. In our current setting, we focus on two independent subtasks of this kind, namely, depression and anxiety disorder prediction. Given a set of tweets published by a particular individual (i.e., a Twitter timeline), and which may or may not disclose mental health information, the goal is to predict whether the individual is likely to receive a diagnosis for depression/anxiety in the future.

Finally, political alignment prediction is the task of inferring whether an individual is a supporter of the former (right-leaning) government of Brazil or not, based on tweets that they have authored. The task may be seen as an instance of author profiling (Rangel et al., 2016, 2020; dos Santos et al., 2020b; Pavan et al., 2023), in which the goal is to infer, e.g., the political leaning (or other demographics) of the individual who published a given tweet that may or may not convey politics-related information. We notice that the task is distinct from previous stance prediction in that the target (i.e., the issue of being for or against the government) is generally not under discussion. Thus, for instance, '*Churches are not supposed to pay taxes*' would more likely be written by a supporter of a conservative government.

4.2 Task datasets

For the stance prediction task, we used a corpus of for/against stances towards six polarised target topics (presidents Lula versus Bolsonaro, the Covid-19 Sinovac vaccine versus Hydroxychloroquine, and a TV network versus the church) described in (Pavan and Paraboni, 2022). The dataset comprises 46.8K manually labelled tweets, and the for/against classes are roughly balanced across targets.

For the mental health prediction task, we used two datasets comprising Twitter timelines of individuals with a diagnosis for depression and anxiety disorder described in (dos Santos et al., 2020a, 2023). The depression dataset contains 13.5K timelines, and the anxiety dataset contains 17.8K timelines in total. As in (Yates et al., 2017) and others, the positive class (i.e., the diagnosed-related data) consists of timelines of individuals who self-disclosed a depression/anxiety diagnosis, as in e.g., '*Last week the doctor told me I have anxiety disorder*'[2]. The negative class, on the other hand, consists of timelines of random users, and it is designed so as to be seven times larger than in the positive class, making this a heavily imbalance classification task. Positive instances are manually labelled at the user (or timeline) level, and matched to their seven random counterparts according to gender, publication dates and number of tweets.

Finally, for the political alignment prediction task, we used a corpus of tweets written by individuals who were identified as being supporters of the current Brazilian president, or against him. The distinction was made based on the use of certain hashtags as described in (da Silva and Paraboni, 2023). For instance, individuals who use the hashtag '#EleNão' ('not him', a popular anti-government slogan during the presidential elections) are labelled as being anti-government, and so every tweet that this individual wrote is labelled in the same way (by label propagation) regardless of its actual contents. The present dataset consists of a random selection of 4010 politically-related tweets from this corpus, and it is class-balanced.

Table 3 summarises descriptive statistics about the evaluation corpora under consideration by reporting the number of positive and negative instances and overall number of tokens of each subset.

4.3 Models

The six stance classifier models were built by making use of a common architecture that was further optimised for each task through grid search. This common architecture consists of a token embedding layer, a recurrent layer of Bidirectional Long Short-Term Memory cells, a multi-head self-attention mechanism and a dense layer with sigmoid activation to produce the output predictions. All layers use dropout regularisation in their inputs. The parameters to be optimised through grid search were the number of BERT layers (last only, or last four), the number of LSTM layers (1 or 2), the number of LSTM hidden dimensions (16 or 1280), attention density (32 or 64) and the number of attention heads (1 or 16). The token embedding layer consists of the pre-trained BERT language models (either from the general-purpose BERTimbau in (Souza et al., 2020), or from our present genre-specific BERTabaporu), and the output is taken to be the hidden state of selected last layers as determined through grid search. In the cases in which there are multiple layers for a single token, these are concatenated. In the 'last four' layers setting,

[2]The self-report itself not included in the corpus data, which conveys only publications prior to the moment of the diagnosis.

Task	(-) instances	(+) instances	Tokens
Stance-Lula	4,514	3,806	422,064
Stance-Bolsonaro	5,565	3,849	259,521
Stance-Hydroxychloroquine	3,978	4,017	277,824
Stance-Sinovac	4,058	3,915	252,663
Stance-Church	3,539	3,598	322,289
Stance-Globo TV	3,341	2,672	214,876
Depression	1,684	11,788	231.26 mi
Anxiety	2,219	15,533	323.75 mi
Political alignment	1,995	2,015	64,275

Table 3: Evaluation corpora descriptive statistics.

we use a 3072-dimensional vector as the embedding representation for each token.

The two mental health classifier models (for depression and anxiety disorder prediction, respectively) were built by using a BERT model (once again, either BERTimbau or BERTabaporu) that has been fine-tuned to each of these two individual tasks. Due to the 512-token input limitation in BERT, these models are trained and tested at 10-tweet batches, which are subsequently combined to decide the final (user-level) class label according to a majority vote. This procedure is repeated for 50 epochs using a random starting point within the user's timeline to select 10 consecutive tweets as the input to the pre-trained BERT model, whose final layer represents the actual text to be classified. This representation is fed into a Bidirectional Long Short-Term Memory layer using RELu activation followed by a fully connected output layer using softmax activation and dropout regularisation, and using binary cross-entropy with balanced class weights as a loss function. The model is trained in a maximum of three epochs and, given 80% of all tweets in the corpus are up to 30-tokens long, the input to BERT is zero-padded to 30 tokens.

Finally, for the political alignment prediction task, we simply used a vanilla BERT architecture consisting of either of the two BERT models under evaluation with softmax activation to produce the output predictions.

5 Results

Table 4 summarises the results obtained by using the general web-based BERTimbau model (Souza et al., 2020) and our current, domain-specific Twitter BERTabaporu model across tasks. In all cases, we follow the existing train-test split available from each corpus and report results over the test set.

From these results we notice that domain-specific BERTabaporu (right side of Table 4 outperforms the more general BERTimbau model in all tasks. The perceived gain is statistically significant at $p < 0,001$ according to a McNemar test (McNemar, 1947) in all tasks except for the smaller political alignment task, in which case results from both models were found to be equivalent. This outcome suggests that using a more genre-specific pre-trained language model may indeed improve results if compared to more general alternatives.

6 Final remarks

This paper introduced BERTabaporu, a BERT language model pre-trained on Twitter data in the Brazilian Portuguese language. Compared to previous work, the present models has been found to outperform the best-known general-purpose model for this language in three Twitter-related text classification tasks, namely, stance, mental health statues, and political alignment prediction, and may be potentially useful to many others Twitter-related applications in the Portuguese language.

As future work, we intended to extend the present analysis by assessing the use of BERTabaporu in other Portuguese NLP tasks. Moreover, since the present model has been trained on a considerably large dataset, there is the question of whether BERTabaporu may be helpful even in non-Twitter evaluation settings. An investigation along these lines is also left as future work.

Acknowledgements

The present research has been supported by the São Paulo Research Foundation (FAPESP grant #2021/08213-0). 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. W. dos Santos has been supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil

Task	BERTimbau (web)			BERTabaporu (Twitter)		
	P	R	F1	P	R	F1
Stance-Lula	0.80	0.80	0.80	0.85	0.84	**0.85**
Stance-Bolsonaro	0.80	0.79	0.80	0.88	0.87	**0.87**
Stance-Hydroxychloroquine	0.79	0.79	0.79	0.85	0.85	**0.85**
Stance-Sinovac	0.81	0.81	0.81	0.86	0.86	**0.86**
Stance-Church	0.83	0.83	0.83	0.87	0.87	**0.87**
Stance-Globo TV	0.85	0.85	0.85	0.90	0.90	**0.90**
Depression	0.61	0.70	0.63	0.68	0.66	**0.67**
Anxiety	0.59	0.64	0.60	0.64	0.64	**0.64**
Political alignment	0.63	0.63	0.63	0.67	0.66	**0.66**

Table 4: Classification results for different Twitter-related tasks using BERT models trained on web (left) and Twitter data (right). The highest F1 score for each task is highlighted.

(CAPES) - Finance Code 001 - grant # 88887.475847/2020-00. S. da Silva has been supported by the Brazilian Foundation CAPES - Coordination for the Improvement of Higher Education Personnel, under grant 88882.378103/2019-01.

References

Abeer Aldayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *5th Workshop on Online Abuse and Harms (WOAH-2021)*, pages 17–25, Online. Association for Computational Linguistics.

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 *27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. 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 *13th Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

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

Paulo Mann, Aline Paes, and Elton H. Matsushima. 2020. See and read: Detecting depression symptoms in higher education students using multimodal social media data. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 440–451.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *3rd Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Assoc. for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *EMNLP-2020 proceedings*, pages 9–14, Online. Association for Computational Linguistics.

Matheus Camasmie Pavan and Ivandré Paraboni. 2022. Cross-target stance classification as domain adaptation. In *Advances in Computational Intelligence - MICAI 2022 - Lecture Notes in Artificial Intelligence vol 13612*, pages 15–25, Cham. Springer Nature Switzerland.

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.

Matheus Camasmie Pavan, Wesley Ramos dos Santos, and Ivandré Paraboni. 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.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In *CLEF 2016 Evaluation Labs and Workshop, Notebook papers*, pages 750–784, Évora, Portugal. CEUR-WS.org.

Francisco Rangel, Paolo Rosso, Wajdi Zaghouani, and Anis Charfi. 2020. Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, page 1–21.

Rafael B. M. Rodrigues, Pedro I. M. Privatto, Gustavo José de Sousa, Rafael P. Murari, Luis C. S. Afonso, João P. Papa, Daniel C. G. Pedronette, Ivan R. Guilherme, Stephan R. Perrout, and Aliel F. Riente. 2022. PetroBERT: A domain adaptation language model for oil and gas applications in portuguese. In *Computational Processing of the Portuguese Language*, pages 101–109, Cham. Springer International Publishing.

Wesley Ramos dos Santos, Amanda Maria Martins Funabashi, and Ivandré Paraboni. 2020a. 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.

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 and Ivandré Paraboni. 2019. Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In *Recents Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria.

Wesley Ramos dos Santos, Ricelli Moreira Silva Ramos, and Ivandré Paraboni. 2020b. Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

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.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Misoslav Konopík. 2021. Czert – Czech BERT-like model for language representation. In *Proceedings of RANLP-2021*, pages 1326–1338, Online. INCOMA Ltd.

Samuel Caetano da Silva and Ivandré Paraboni. 2023. Politically-oriented information inference from text. *Journal of Universal Computer Science*, 29(6):570–595.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems (BRACIS) - LNCS 12319*, Cham. Springer.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 1191–1198.

Aamir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M. Meddar, Annie Myers, Jyotishman Pathak, and Pascal Hitzler. 2020. Multimodal mental health analysis in social media. *PLOS ONE*, 15(4):1–27.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv 2209.07562*.