Predição de transtorno depressivo em redes sociais: BERT supervisionado ou ChatGPT *zero-shot*?

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 first study on the use of the ChatGPT dialogue system in a complex and sensitive application, namely, the computational prediction of mental health disorders from social media text. To this end, we conducted an experiment to compare a traditional supervised approach based on BERT with a zero-shot strategy based on natural language prompts submitted directly to the dialogue system. Results of this evaluation, taking into account the accuracy of the classification task in view of the need for corpus annotation in the supervised approach, highlight different advantages of each alternative.*

Resumo. *Este artigo apresenta um primeiro estudo sobre o uso do sistema de diálogo ChatGPT em uma aplicação complexa e sensível: a predição computacional de transtornos de saúde mental a partir de textos provenientes de redes sociais. Para esse fim, foi conduzido um experimento comparando uma abordagem supervisionada tradicional baseada em BERT com uma estratégia zero-shot baseada em prompts em língua natural submetidos diretamente ao sistema de diálogo. Resultados desta avaliação, levando em conta a acurácia da tarefa de classificação face à necessidade de anotação prévia de córpus da abordagem supervisionada, destacam diferentes vantagens de cada alternativa.*

1. Introdução

As formas de representação textual utilizadas em aplicações de PLN têm evoluído de forma acelerada em anos recentes. Partindo-se de modelos simples baseados em contagens de *tokens*, como o *bag-of-words* e suas variações, no espaço de poucos anos passou-se ao uso de *word embeddings* estáticos, como os produzidos com word2vec [Mikolov et al. 2013], e posteriormente dependentes de contexto, como nos modelos de língua pré-treinados do tipo BERT [Devlin et al. 2019]. Este último, até recentemente considerado o modelo mais expressivo de sua categoria, com 340 milhões de parâmetros, começa a enfrentar a concorrência de modelos ainda maiores e mais sofisticados, denominados LLMs (*large language models*) como por exemplo GPT-3 [Brown et al. 2020] e Bloom [BigScience Workshop 2022], com 175 bilhões de parâmetros cada.

A crescente complexidade dos atuais LLMs pode entretanto restringir seu uso a aplicações para o idioma inglês, ou àquelas em que é possível fazer uso de uma versão multilíngue do modelo. Neste sentido, uma alternativa de interesse para a experimentação

em PLN foi a disponibilização da interface de diálogo ChatGPT[1], que combina um LLM da família GPT com métodos de aprendizado supervisionado e por reforço utilizando *feedback* humano para modelar a tarefa de conversação com usuários humanos com alto grau de realismo.

Embora sejam modelos do tipo generativo (i.e., que essencialmente completam um trecho de texto com sua continuação mais provável), LLMs e sistemas deles derivados (como ChatGPT) possuem, em virtude do grande volume de dados de treinamento a que tiveram acesso durante sua construção, ampla capacidade de associar pares de textos, e podem assim ser facilmente adaptados a diversas tarefas de interpretação de língua natural [Zhang et al. 2023]. Em especial, observa-se que com poucos meses de lançamento o sistema ChatGPT já começou a ser cogitado como uma possível alternativa a métodos supervisionados tradicionais (i.e., baseados em córpus de exemplos rotulados) por permitir a consulta direta ao LLM sem exemplos prévios. Em métodos deste tipo, ao invés de rotular um córpus de avaliações de produtos com informação de sentimento (e.g., positivo, neutro ou negativo), podemos em tese simplesmente submeter ao LLM uma instrução em língua natural perguntando qual o sentimento expresso em um determinado texto.

A ausência de exemplos de treinamento em métodos baseados em LLMs é uma forma de classificação de texto do tipo *zero-shot*. Embora haja considerações relativas à segurança e contaminação de dados (i.e., o fato de que estes modelos são constantemente atualizados e podem ter sido expostos aos dados de teste da tarefa em tempo de treinamento, cf. [Zhang et al. 2023]), o uso de ferramentas como ChatGPT em tarefas de PLN como sumarização, sistemas de respostas a perguntas, análise de sentimentos e outras [Qin et al. 2023] tem se difundido com rapidez, e recentemente foi levantada até mesmo a hipótese do 'começo do fim da tarefa de anotação de córpus' [Kuzman et al. 2023].

Com base nestas considerações, neste trabalho apresentamos um primeiro estudo sobre o uso de ChatGPT em uma aplicação complexa e enfocando um tema intencionalmente sensível, carregado de questões éticas do tipo que sistemas como ChatGPT notoriamente tentam evitar: a predição computacional de transtornos de saúde mental a partir de textos provenientes de redes sociais. Aplicações deste tipo, já amplamente desenvolvidas com uso de métodos supervisionados convencionais [Chancellor and Choudhury 2020, Su et al. 2020], são aqui tratadas pela primeira vez com uso de métodos *zero-shot*, não havendo (até onde temos conhecimento) risco de contaminação de dados.

De forma mais específica, o presente trabalho objetiva avaliar a detecção de indivíduos com maior risco de desenvolver transtorno depressivo a partir de suas postagens no Twitter brasileiro, utilizando para este fim uma abordagem supervisionada tradicional baseada em BERT e, como alternativa, uma estratégia do tipo *zero-shot* baseada em *prompts* submetidos diretamente ao sistema de diálogo ChatGPT, levando-se em conta a acurácia da tarefa de classificação face à necessidade de anotação prévia de córpus da abordagem supervisionada. As principais contribuições previstas são as seguintes:

- Método inédito do tipo *zero-shot* para detecção de transtorno depressivo a partir de postagens no Twitter em Português.
- Comparação com modelo supervisionado do tipo estado-da-arte para essa tarefa, baseado em BERT e Bi-LSTMs.

[1] `https://chat.openai.com/chat`

O restante deste artigo está organizado da seguinte forma. A seção 2 apresenta um breve levantamento de estudos existentes da área de predição de transtorno de depressão a partir de textos. A seção 3 introduz os modelos baseados em BERT e ChatGPT desenvolvidos. A seção 4 descreve a avaliação conduzida, e a seção 5 apresenta seus resultados. Finalmente, a seção 6 sumariza a presente discussão e apresenta futuras direções de pesquisa sobre o assunto.

Considerações éticas

Os modelos computacionais discutidos foram desenvolvidos com base em dados publicamente disponibilizados na plataforma Twitter, aqui tratados de forma anonimizada e confidencial. A presente abordagem linguístico-computacional não deve ser vista como substituto a outras formas de aquisição de conhecimento (em especial, derivadas da área médica) e não objetiva diagnosticar *indivíduos* com transtornos de saúde mental, mas apenas contribuir para a área de análise computacional de redes sociais enfocando o *estudo da linguagem* empregada nestas circunstâncias.

2. Trabalhos relacionados

A detecção transtorno depressivo com base em dados textuais (e.g., provenientes de redes sociais ou outras fontes) é tipicamente modelada na forma de um problema de aprendizado de máquina supervisionado, ou seja, fazendo uso de córpus de textos rotuladas com informações relativas ao estado de saúde mental de seus autores (e.g., usuários de redes sociais) para treino e teste de classificadores. Sob esta perspectiva, a tarefa pode ser vista como uma instância do problema de caracterização autoral [dos Santos et al. 2020b, Pavan et al. 2023, Flores et al. 2022] combinado à detecção de linguagem afetiva [da Silva et al. 2020]. Um levantamento de estudos recentes deste tipo é apresentado na Tabela 1, com indicação do gênero de texto considerado (Reddit, Twitter), a forma de representação dos dados textuais (b=*bag of words*, BERT [Devlin et al. 2019], d=características de domínio, e=*embeddings*, h=horário da publicação, i=imagens, l=atributos LIWC [Pennebaker et al. 2001], m=metadados, n=informações de rede, p=*part-of-speech*, s=atributos afetivos, t=tópicos, u=informações demográficas), e métodos computacionais (e.g., CNN=redes neurais convolucionais, LSTM=*long short-term neural networks*, LR=regressão logística, RF=*Random Forest*, etc.).

Dentre os estudos selecionados, observa-se uma ligeira predominância de trabalhos baseados na rede social Reddit. Esta preferência pode ser explicada pela maior facilidade de acesso e reúso de dados desse tipo para fins de pesquisa, o que é mais restrito no caso da plataforma Twitter. Assim, postagens Reddit são usadas em alguns dos conjuntos de dados mais conhecidos para o idioma inglês, como os córpus SMHD [Cohan et al. 2018] e eRisk [Losada and Crestani 2016], sendo este último também a base de uma série de desafios computacionais (ou '*shared tasks*') *Early Risk Prediction on the Internet* [Parapar et al. 2022].

Quanto aos tipos de modelos textuais utilizados, a Tabela 1 reflete a evolução natural da pesquisa em áreas correlatas do PLN, com predominância inicial de modelos do tipo *bag-of-words* e engenharias de características, e sua substituição gradual por modelos baseados em *word embeddings* e, mais recentemente, BERT [Devlin et al. 2019].

Tabela 1. Detecção de transtorno depressivo a partir de texto

Estudo	Gênero	Repres. textual	Método
[Cohan et al. 2018]	reddit	b,e	FastText
[Trotzek et al. 2018]	reddit	e,p,m,d	CNN
[Kumar et al. 2019]	twitter	d,s,h	ensemble
[Aragón et al. 2019]	reddit	s	SVM
[Cacheda et al. 2019]	reddit	h,m,n	RF
[Burdisso et al. 2020]	reddit	b	SS3
[Lin et al. 2020]	twitter	e,i	CNN
[Yazdavar et al. 2020]	twitter	b,s,t,i,n,l,u	RF
[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. 2023]	twitter	BERT	Bi-LSTM

No que diz espeito aos métodos computacionais empregados, de modo geral observa-se a mesma trajetória, com o uso de classificadores tradicionais baseados em contagens de *tokens* sendo gradualmente substituído por métodos de classificação de sequências baseados em aprendizado profundo, incluindo o uso mais recente de arquiteturas baseadas em *transformers*.

Com exceção dos estudos para o português em [dos Santos et al. 2020a, dos Santos et al. 2023], que introduziram o córpus denominado SetembroBR para detecção de transtorno de depressão e ansiedade no Twitter brasileiro, todos os trabalhos identificados são dedicados ao idioma inglês. Assim, este córpus será tomado como base no presente trabalho, conforme discutido nas próximas seções.

Finalmente, observa-se que nenhum dos estudos identificados faz uso de métodos *zero-shot*, baseados em *prompt* ou em modelos de língua de grande escala e similares. Destacamos, entretanto, que o sistema ChatGPT tem sido utilizado da forma que pretendemos aplicar ao presente problema de detecção de depressão em diversas outras tarefas de PLN, como a classificação de gênero linguístico [Kuzman et al. 2023], detecção de atitudes ou posicionamentos [Zhang et al. 2023] e análise de sentimentos [Qin et al. 2023].

3. Modelos de predição de transtorno depressivo

O estudo conduzido avalia dois modelos de predição de transtorno depressivo a partir de textos publicados na rede social Twitter. O primeiro modelo, baseado no sistema de diálogos ChatGPT, é uma abordagem do tipo *zero-shot* que não faz uso de nenhum tipo de anotação manual de dados. O segundo modelo segue uma abordagem supervisionada tradicional baseada em dados de um córpus rotulado e uso do modelo de língua BERT [Devlin et al. 2019]. O objetivo da avaliação é o de verificar qual destas duas abordagens apresenta melhores resultados para um subconjunto de *timelines* (conjuntos de postagens ou *tweets*) de teste do córpus SetembroBR [dos Santos et al. 2023].

O modelo de predição baseado no sistema ChatGPT foi construído da seguinte forma. Quando questionado diretamente sobre saúde mental ou outras questões de natureza sensível, o sistema ChatGPT geralmente evita fornecer uma resposta, usando justifi-

cativas como '*o diagnóstico de depressão só pode ser feito por um médico*', '*sou apenas um modelo de língua*' e similares. Assim, como forma de contornar estas medidas de proteção, foi utilizado um *prompt* que solicita informações sobre possíveis indicadores linguísticos de depressão que podem ou não estar presentes em um dado *tweet*. O *prompt* submetido ao sistema ChatGPT *14 March Version* foi o seguinte:

> *Desconsidere o diálogo anterior e avalie se o texto a seguir apresenta indicadores linguísticos de depressão em uma escala de 0 a 10, onde 0 significa nenhum indício e 10 significa indícios muito fortes. Ao responder, informe apenas um número de 0 a 10, sem justificar ou explicar sua resposta. O texto a ser analisado é* <**texto do tweet aqui**>.

A escala de 0 a 10 foi utilizada porque, em testes preliminares, este tipo de *prompt* se revelou mais preciso do que solicitações mais diretas, como no caso de resposta binária (e.g., '*indique se o texto possui ou não indicadores de depressão*'). Além disso, sendo um sistema de conversação dotado de memória de diálogo, foi solicitado explicitamente que as interações anteriores (que podiam ser referentes ao mesmo autor do *tweet* atual ou outro qualquer) fossem desconsideradas, forçando o sistema a analisar cada mensagem de forma independente das demais. Finalmente, para maior rapidez na resposta do sistema, foi solicitado que não fosse apresentada nenhuma explicação adicional motivando a resposta.

Com base neste método, uma coleção de *tweets* de teste foi rotulada pelo sistema ChatGPT com escores de 0 a 10. Para a combinação desses escores individuais em um rótulo de classe global (i.e., considerando todos os *tweets* de uma *timeline* de um determinado indivíduo), a média destes escores foi comparada a um valor de *threshold* fixo previamente computado a partir de dados de treino não utilizados na presente avaliação.

De forma mais específica, o sistema ChatGPT foi utilizado para rotular um conjunto de 30 *timelines* de treino da classe de *Controle* (ou seja, um conjunto de indivíduos aleatórios selecionados a partir da população geral) contendo 80 tweets cada. A seguir, foi computada a média de escores de todos os $30 * 80 = 2400$ *tweets* de treino, e este valor foi utilizado como *threshold* para definir os rótulos (*Diagnosticados* ou *Controle*) das *timelines* de teste.

Como alternativa ao modelo baseado em ChatGPT, foi considerada também uma abordagem tradicional de aprendizado supervisionado baseada em modelos de língua do tipo BERT. Para este fim, utilizou-se o modelo BERTabaporu [da Costa et al. 2023], um modelo BERT treinado com base em 2.9 bilhões de *tokens* obtidos a partir de 237 milhões de tweets em português. Nesta abordagem, a representação das *timelines* a serem rotuladas como *Diagnosticado* ou *Controle* é feita em sequências consecutivas de 10 *tweets* iniciadas em uma posição aleatória da *timeline* a cada época. Esta representação textual alimenta uma camada Bi-LSTM com 100 neurônios seguida de 3 camadas do tipo MLP, cada uma com dropout de 0,1 e função de ativação *softmax*.

4. Avaliação

Como forma de comparar o desempenho dos modelos baseados em ChatGPT e BERT discutidos na seção anterior, foi conduzido um experimento de aplicação destes modelos preditivos a uma porção do córpus SetembroBR [dos Santos et al. 2023] de *timelines* de usuários do Twitter brasileiro com diagnóstico de depressão, e de usuários

Tabela 2. Subconjunto Depressão do córpus SetembroBR

Métrica	Diagnosticados	Controle
Usuários	1684	11788
Tweets (milhões)	2,43	16,99
Tokens (milhões)	29,32	201,94

aleatórios formando um grupo de controle de proporção 7 vezes superior. Nesta definição do problema, também seguida em [Coppersmith et al. 2015, Losada et al. 2017, Lynn et al. 2018, Cohan et al. 2018, Losada et al. 2019, Parapar et al. 2022] e outros, o objetivo é distinguir indivíduos depressivos da população em geral, e não distinguir indivíduos depressivos de não-depressivos.

A Tabela 2 sumariza estatísticas descritivas da porção de dados referentes ao transtorno depressivo presentes no córpus.

O córpus possui uma divisão aleatória pré-definida entre *timelines* de treinamento (80%) e teste (20%) que foi respeitada no experimento realizado. Entretanto, os dados de treino são usados de forma diferente pelos dois modelos desenvolvidos. Para o modelo baseado em ChatGPT, a porção de treino foi utilizada apenas para cálculo do valor de *threshold* de separação das classes *Diagnosticados* e *Controle*, conforme descrito na seção anterior. O modelo BERT, por outro lado, utiliza a porção de treino completa do córpus para sua construção.

No que diz respeito aos dados de teste utilizados por ambos os modelos, observa-se que não seria praticável submeter manualmente para avaliação do sistema ChatGPT cada um dos cerca de 3,9 milhões de *tweets* de teste do córpus SetembroBR. Assim, na presente avaliação foi utilizado apenas um subconjunto reduzido de *timelines*, contendo cada uma um número também reduzido de *tweets*.

O experimento realizado baseou-se em um subconjunto de 50 *timelines* de cada classe, selecionadas aleatoriamente a partir do conjunto de teste do córpus, e cobrindo cada uma um intervalo fixo de 80 *tweets* consecutivos com a maior frequência possível dos termos 'depressão' e 'ansiedade'. Esta estratégia de seleção objetivou maximizar as chances de que, mesmo analisando-se uma porção reduzida dos dados, algum indício de discussão sobre questões de saúde mental pudesse ser encontrado no trecho avaliado, ressaltando-se que esta simplificação não representa uma vantagem para nenhum dos dois modelos sob avaliação (já que ambos utilizam os mesmos dados de teste), e não torna a tarefa computacional menos complexa (dado que tanto indivíduos das classe *Diagnosticados* como *Controle* podem mencionar ou não estes termos).

5. Resultados

Os dados de teste selecionados foram submetidos para avaliação dos modelos ChatGPT e BERT conforme descrito nas seções anteriores. Para este fim, foram computadas as medidas de acurácia por classe e a acurácia média de cada modelo, observando-se que o conjunto de teste é perfeitamente balanceado. A Tabela 3 apresenta os resultados obtidos.

Os presentes resultados motivam uma série de considerações. Em primeiro lugar, observa-se que, na média global, os dois modelos são essencialmente similares, um resultado que é em certo sentido inesperado tendo-se em vista a grande diferença metodológica

Tabela 3. Resultados de predição de transtorno de depressão.

Modelo	Diagnosticados	Controle	Média
ChatGPT	0,70	0,60	0,65
BERT	0,48	0,84	0,66

entre as duas abordagens. Entretanto, observando-se o comportamento individual de cada modelo nas classes *Diagnosticados* e *Controle*, esta diferença se reflete de forma mais evidente. O modelo baseado em ChatGPT possui capacidade relativamente elevada de classificar corretamente as *timelines* de indivíduos *Diagnosticados*, mas apresenta menor sucesso ao tratar o grupo *Controle*. No caso do modelo baseado em BERT, o efeito é o contrário, ou seja, uma menor acurácia na classe *Diagnosticados* é compensada por uma melhoria na classe *Controle*.

Uma possível explicação para estes resultados seria a de que o modelo baseado em ChatGPT é realmente superior ao modelo baseado em BERT quando aplicado à tarefa de identificar indicadores linguísticos de depressão mas, como a tarefa modelada pelo córpus é a de distinção entre *Diagnosticados* e um grupo de *Controle* aleatório (e que não representa uma classe negativa do tipo 'não diagnosticados', mas apenas uma população média), o conjunto *Controle* também apresenta alguns indicadores deste tipo (ainda que certamente em menor proporção do que no conjunto *Diagnosticados*). Assim, embora o modelo ChatGPT tenha ampla vantagem em relação ao modelo BERT na identificação destes indicadores, a separação entre estes casos e os exemplos aleatórios carece da noção de 'população média' que, crucialmente, está presente no conjunto de treino empregado pelo modelo BERT, e que possivelmente explica a superioridade do modelo BERT na classe *Controle*. Em outras palavras, a habilidade de detecção do sistema ChatGPT é significativa, mas o conceito de 'população média' ainda é algo que não foi adequadamente modelado pela presente engenharia de *prompts* dado que o modelo ChatGPT *zero-shot* não conta com exemplos do que seria essa população.

6. Conclusões

Este artigo apresentou um primeiro estudo sobre o possível uso da ferramenta ChatGPT em uma tarefa de PLN de natureza notadamente sensível e complexa - a predição de transtorno depressivo em redes sociais - e sua comparação com um método tradicional baseado em BERT. Nossos resultados indicam que, embora ambos modelos tenham obtido acurácia média semelhante, o modelo baseado em ChatGPT pode ser considerado superior no sentido de não fazer uso de dados rotulados manualmente, enquanto o modelo BERT supervisionado exige córpus de treinamento anotado.

Voltando à questão do título deste artigo - BERT supervisionado ou ChatGPT *zero-shot* - propomos uma resposta indireta. Apesar dos resultados médios similares, o método *zero-shot* é melhor na classe positiva (i.e., na detecção de usuários diagnosticados), enquanto o método supervisionado é melhor na classe negativa (ou grupo de *Controle*). Como essa diferença decorre da forma como a presente tarefa computacional é definida (ou seja, como uma tarefa de distinção entre indivíduos diagnosticados e indivíduos aleatórios que representam uma população média), não é possível verificar essa questão com base no córpus empregado no presente estudo. É possível entretanto que a vantagem do modelo ChatGPT seja ainda mais expressiva em um cenário de classificação

dito tradicional, como o da distinção entre indivíduos depressivos e não depressivos. Um estudo desta natureza é deixado como sugestão de trabalho futuro.

Mesmo considerando-se as peculiaridades da presente definição do problema, observamos também que a fragilidade do modelo baseado em ChatGPT parece não estar tanto no modelo de língua em si, mas sim na forma como o rótulo de classe é decidido. No presente estudo, optou-se por utilizar a média simples dos escores do sistema com uso de um valor de *threshold* previamente computado para decidir se a resposta do modelo seria *Diagnosticado* ou *Controle*, mas é possível que um método mais sofisticado possa aproximar esses resultados dos obtidos pelo modelo BERT supervisionado.

Finalmente, cabe observar que o uso de métodos baseados em ChatGPT e afins pode ser menos adequado a tarefas de caracterização autoral como no presente caso, e mais adequado a tarefas de interpretação de língua natural para extração de significado textual, como análise de sentimentos ou detecção de posicionamentos [Pavan et al. 2020, Pavan and Paraboni 2022] aos moldes apresentados em [Zhang et al. 2023]. Uma iniciativa de investigação desta natureza também é deixada como sugestão de trabalho futuro.

7. Agradecimentos

Esse trabalho conta com apoio FAPESP # 2021/08213-0. Os autores agradecem ao Centro de Inteligência Artificial (C4AI-USP) e ao apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation. O primeiro autor recebe apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 (# 88887.475847/2020-00).

Referências

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.

BigScience Workshop (2022). BLOOM (revision 4ab0472).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

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.

Cacheda, F., Fernandez, D., Novoa, F. J., and Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *J Med Internet Res*, 21(6):e12554.

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

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.

Coppersmith, G., Dredze, M., Harman, C., Kristy, H., and Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 31–39, Denver, 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)*, Varna, Bulgaria.

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.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 Proceedings*, pages 4171–4186, Minneapolis, USA.

dos Santos, W. R., de Oliveira, R. L., and Paraboni, I. (2023). 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. (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.

dos Santos, W. R., Ramos, R. M. S., and Paraboni, I. (2020b). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

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.

Kumar, A., Sharma, A., and Arora, A. (2019). Anxious depression prediction in real-time social data. In *Intl. Conf. on Advances in Engineering Science Management & Technology*, Dehradun, India.

Kuzman, T., Mozetič, I., and Ljubešić, N. (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. *arXiv preprint arXiv:2303.03953*.

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.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *LNCS 10456*, pages 346–360, Cham. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. In *LNCS 11696*.

Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., and Schwartz, H. A. (2018). CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, USA.

Mikolov, T., Wen-tau, S., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proc. of NAACL-HLT-2013*, pages 746–751, Atlanta, USA. Assoc for Comp Ling.

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.

Pavan, M. C. and Paraboni, I. (2022). Cross-target stance classification as domain adaptation. In Pichardo Lagunas, O., Martínez-Miranda, J., and Martínez Seis, B., editors, *Advances in Computational Intelligence - MICAI 2022 - Lecture Notes in Artificial Intelligence vol 13612*, pages 15–25, Cham. Springer Nature Switzerland.

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

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv preprint arXiv:2302.06476*.

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.

Su, C., Xu, Z., Pathak, J., and Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(116).

Trotzek, M., Koitka, S., and Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., Thirunarayan, K., Meddar, J. M., Myers, A., Pathak, J., and Hitzler, P. (2020). Multimodal mental health analysis in social media. *PLOS ONE*, 15(4):1–27.

Zhang, B., Ding, D., and Jing, L. (2023). How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.