UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-000/2021

SetembroBR v3: detecção de transtornos de depressão e ansiedade em redes sociais

Wesley Ramos dos Santos, Ivandré Paraboni

Agosto - 2021

Série de Relatórios Técnicos

PPgSI-EACH-USP
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo
03828-000 – São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

SetembroBR v3: detecção de transtornos de depressão e ansiedade em redes sociais

Wesley Ramos dos Santos, Ivandré Paraboni[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brazil

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

Resumo. *A observação de que indivíduos com transtornos de saúde mental como depressão e ansiedade são muitas vezes usuários regulares de redes sociais motiva uma ampla gama de pesquisas na área de Processamento de Línguas Naturais (PLN) dedicadas ao reconhecimento automático de situações de risco deste tipo a partir de textos. Este relatório descreve a construção de um recurso linguístico-computacional para este fim, o córpus SetembroBR versão 3, e algumas oportunidades de pesquisa que este recurso proporciona.*

1. Introdução

Transtornos de saúde mental como depressão e ansiedade são desafios bem conhecidos e uma crescente fonte de preocupação na sociedade moderna. De acordo com dados da Organização Mundial da Saúde [WHO 2017], a população brasileira apresentava já em 2017 a maior proporção de casos de depressão (5,8%, ou 12,2 milhões de indivíduos) e ansiedade (9,3%, ou 19,6 milhões de indivíduos) da América Latina. Ao mesmo tempo, diversos estudos demonstram que indivíduos com transtornos de saúde mental são usuários regulares de redes sociais em proporção similar à população em geral, que é estimada em cerca de 70% entre indivíduos de meia idade e até 97% entre indivíduos mais jovens [Naslund et al. 2020, Aschbrenner et al. 2018, Birnbaum et al. 2017, Brunette et al. 2019], e que indivíduos com estes transtornos frequentemente recorrem às redes sociais em busca do suporte de outros usuários com problemas semelhantes [Bucci et al. 2019].

Este cenário, e a observação de que transtornos de saúde mental frequentemente se refletem na linguagem empregada pelos indivíduos que sofrem dessas condições, levaram a um número expressivo de estudos no Processamento de Línguas Naturais (PLN) e áreas relacionadas, tendo como foco a detecção de transtornos como depressão, ansiedade, bipolaridade, anorexia, ideação suicida e automutilação em redes sociais [Yates et al. 2017, Yazdavar et al. 2017, Shen et al. 2017, Shen e Rudzicz 2017, Cohan et al. 2018, Trotzek et al. 2018, Lin et al. 2020, Mann et al. 2020]. Estudos deste tipo, que procuram identificar casos de maior gravidade e eventualmente sinalizar a necessidade de um indivíduo buscar ajuda, são também o foco desde relatório, que trata especificamente dos transtornos de depressão e ansiedade observáveis na plataforma Twitter de língua portuguesa.

A identificação computacional de indivíduos com transtornos de saúde mental a partir de redes sociais é uma tarefa complexa e não totalmente resolvida na pesquisa em PLN, sendo tipicamente modelada como um problema de aprendizado de máquina (AM) supervisionado. Em abordagens deste tipo, um córpus de publicações (e.g., puramente textuais ou multimodais) rotulado com informações relativas ao estado mental (e.g., depressivo ou não depressivo) dos usuários que as produziram é utilizado para treinamento e teste de modelos de classificação do fenômeno de interesse.

Na construção da base de dados, usuários com transtorno de saúde mental (ou pertencentes à classe positiva, na terminologia de AM) podem ser selecionados por meio de validação externa (e.g., por meio de avaliação médica especializada ou uso de questionários), obtendo-se, por exemplo, um córpus de indivíduo com e sem depressão a serem distinguidos com uso de modelos classificadores. Este método de aquisição de dados, apesar de confiável, é entretanto de alto custo, o que na prática limita o número de casos (ou instâncias) obtidos. Por este motivo, a seleção de indivíduos por validação externa não é a formulação mais comum deste problema computacional.

Como alternativa à validação externa, estudos da área de PLN têm explorado a ocorrência de autorrelatos em redes sociais, como em '*Fui ao psiquiatra hoje e ela me diagnosticou com depressão*' para selecionar rapidamente grandes massas de indivíduos da classe positiva com baixo custo. Mensagens deste tipo - que servem apenas para a seleção de usuários, mas que não fazem parte do córpus propriamente dito - são frequentes em redes sociais, e embora o método seja naturalmente mais sujeito a imprecisões ou falsidade (e que de resto são possíveis também no uso de questionários etc.), é geralmente aceito que estas dificuldades tendem a ser compensadas pelo maior volume de dados obtido [Coppersmith et al. 2015].

Dado entretanto que autorrelatos fornecem evidência apenas da possível *presença* de um transtorno, mas não da sua *ausência* - e que portanto não podem ser tomados por base para seleção de indivíduos que deveriam compor a classe negativa do aprendizado - utiliza-se neste caso uma formulação distinta para o problema. Mais especificamente, ao invés de utilizar um conjunto de usuários (por exemplo) sem depressão, utiliza-se um grupo de usuários de controle selecionados aleatoriamente (e que portanto inclui um certo número de indivíduos com depressão) de proporção várias vezes superior ao da classe positiva (normalmente proporcional à incidência do transtorno na população em geral).

Assim, o problema computacional a ser resolvido não consiste em distinguir indivíduos com e sem depressão, e sim em identificar indivíduos com probabilidade acima da média da população geral de ter depressão. Esta probabilidade, que entre indivíduos da classe positiva deveria ser próxima de 100%, deve ser muito mais baixa no grupo de controle.

O uso de autorrelatos para identificar indivíduos (provavelmente) diagnosticados com depressão/ansiedade configura assim um problema de aprendizado de máquina fortemente desbalanceado. Mesmo com estas dificuldades, entretanto, a seleção de indivíduos com transtorno de saúde mental por meio de seus próprios autorrelatos é o método de coleta de dados mais popular na pesquisa na área de PLN [Yates et al. 2017, Cohan et al. 2018, Losada et al. 2019] e será também o foco do presente relatório, envolvendo a questão da construção de uma base de dados (ou córpus) de treinamento e teste, e o desenvolvimento de modelos de classificação propriamente ditos. É importante destacar entretanto que a presente opção por uma perspectiva linguístico-computacional não deve ser vista como substituto a outras formas de aquisição de conhecimento (em especial, derivadas da área médica), e sim de forma complementar a estes métodos.

A detecção de transtornos de depressão e ansiedade a partir de textos faz o uso de córpus especialmente coletados e rotulados com este propósito. Recursos linguístico-computacionais deste tipo já foram desenvolvidos em quantidade expressiva para o idioma inglês, em alguns casos como subproduto de eventos do tipo *shared task* como o desafio

CLPsych-2015 [Coppersmith et al. 2015] e a série eRisk [Losada et al. 2019].

Sob a perspectiva da presente pesquisa, observa-se que os córpus existentes apresentam - além da diferença de idioma - diversas lacunas a serem exploradas. Por exemplo, observa-se que estes córpus são normalmente limitados a dados textuais, desconsiderando assim outras informações disponíveis a partir de redes sociais como as conexões de um indivíduo e suas interações com outros usuários, as quais têm se mostrado úteis na detecção de transtorno de ansiedade [Dutta e Choudhury 2020].

De especial interesse para este trabalho, observa-se que córpus existentes não levam em conta o *momento do diagnóstico* do transtorno de saúde mental, o que significa que o modelo de classificação tem acesso irrestrito a textos produzidos antes e depois da confirmação por um especialista. Embora na prática esta diferença possa ser pequena (i.e., um indivíduo verdadeiramente diagnosticado com depressão e outro igualmente deprimido, mas que ainda não foi ao psiquiatra, seriam próximos do ponto de vista da aplicação), um modelo baseado unicamente em dados *anteriores* ao diagnóstico pode ser de maior utilidade na detecção precoce destes transtornos, e assim servir ao propósito de prevenção em um sentido mais estrito do termo.

Independentemente do método de coleta de dados, entretanto, a detecção de transtornos de saúde mental apresenta os desafios comuns à boa parte das aplicações de PLN. Em especial, observa-se que a distinção entre indivíduos diagnosticados e aqueles que pertencem ao grupo de controle com base em características linguísticas é naturalmente dificultada pelo fato de que ambos os grupos podem discutir qualquer tipo de assunto, incluindo questões relacionadas à saúde mental (deles próprios, de terceiros etc.), em uma variedade de contextos reais e imaginários, fazendo uso de humor, sarcasmo, conjecturas e muitas outras formas de expressão.

Estas dificuldades sugerem assim não apenas a necessidade de modelos de representação e classificação textual sofisticados, como possivelmente o uso de conhecimento adicional, como os diversos indicadores linguísticos estabelecidos na literatura médica [Trifu et al. 2017]. Indicadores deste tipo incluem o uso frequente de pronomes de primeira pessoa[1], tempo passado em verbos de ação, inversão da ordem de palavras em tópicos, ênfases, presença de sentenças curtas, impessoais, truncadas ou ríspidas, elipses, tautologias, repetições e ausência de comparação.

O uso de termos absolutistas (e.g., todo, sempre etc.) também é mais fortemente associado a indivíduos com depressão [Al-Mosaiwi e Johnstone 2018], assim como o uso de expressões de conotação negativa em certas redes sociais [Seabrook et al. 2018]. Finalmente, alguns estudos exploram também o uso de informações não textuais como conexões da rede social, frequência e horário de publicações e outros [Choudhury et al. 2013, Shrestha e Spezzano 2019].

No contexto do presente projeto, consideramos ainda a possibilidade de que conhecimento proveniente de outras tarefas de PLN, como o reconhecimento de posicionamentos (ou *stance*) [Mohammad et al. 2016, 2017, dos Santos e Paraboni 2019, Siddiqua et al. 2019, Pavan et al. 2020] ou da caracterização autoral [dos Santos et al. 2017, Preotiuc-Pietro et al. 2017, Silva e Paraboni 2018a,b, Hsieh et al. 2018, Takahashi et al. 2018, Ramos et al. 2018, Pizarro 2019, dos Santos et al. 2020b, Rangel et al. 2020] de traços

[1]Para estudos de resolução pronominal, ver, e.g., Paraboni [1997], Paraboni e de Lima [1998].

de personalidade [2] como no modelo dos cinco grandes fatores (CGF ou Big Five) [Goldberg 1990] ou MBTI [Myers 1962], possa contribuir para solução computacional deste problema

Com base nestas observações, este documento descreve a criação de um recurso linguístico-computacional de larga escala dedicado ao português, sem similar conhecido, cobrindo tanto o problema de detecção de depressão (que é o mais frequentemente estudado na área) como também de transtorno de ansiedade, e incluindo dados textuais e não-textuais (e.g., relativos aos contatos e interações na rede social).

2. Trabalhos relacionados

O estudo de detecção precoce de transtornos de saúde mental sob uma perspectiva de PLN suscita questões de pesquisa referentes à natureza dos dados a serem empregados na construção de modelos computacionais, e referentes aos métodos computacionais propriamente ditos. Estas duas questões são discutidas individualmente nas seções a seguir.

2.1. Conjuntos de dados para detecção de depressão/ansiedade

O desenvolvimento de modelos computacionais de detecção de transtornos de saúde mental a partir de textos tipicamente faz uso de uma base de dados cuidadosamente projetada com exemplos de documentos escritos por indivíduos com estes transtornos, e por indivíduos de um grupo de controle. Assim, nesta seção apresentamos um breve levantamento dos recursos linguísticos-computacionais existentes na área, e discutimos como o córpus a ser proposto pretende se destacar do estado-da-arte. Para este fim, serão considerados aqui apenas os recursos construídos a partir de autorrelatos, e em proporção significativa para possível uso de métodos de AM, desconsiderando-se aqueles que realizam experimentos com bases de dados próprias de menor volume e/ou que não são disponibilizadas publicamente. A exceção é o estudo em Mann et al. [2020], que mesmo sendo baseado em inventários (e não autorrelatos) de depressão, foi incluído neste levantamento por ser um dos poucos exemplos de conjunto de dados existentes em português brasileiro encontrados na literatura.

Para cada um dos estudos selecionados, a tabela 1 apresenta o domínio, idioma e modalidade de dados, o número de instâncias de depressão/ansiedade, a relação entre usuários de controle (C) e diagnosticados (D) e o método de pareamento das classes. Detalhes adicionais são discutidos a seguir.

[2] Também estudado no contexto da seleção de conteúdo para geração de língua natural em, e.g., Paraboni e van Deemter [1999], Paraboni [2003].

Tabela 1. Córpus rotulados com informação de depressão/ansiedade.

Córpus	Domínio	Idioma	Modalidade	Dep.	Ans.	C/D	Pareamento
CLPsych-2015	Twitter	En	texto	477	-	1	idade/gênero
RSDD	Reddit	En	texto	9210	-	11,6	comportamento
Shen et. al.	Twitter	En	texto	1402	-	1	na
eRisk-2017	Reddit	En	texto	135	-	5,6	na
eRisk-2018	Reddit	En	texto	214	-	7	na
SMHD	Reddit	En	texto	14139	8783	9	comportamento
Mann et. al.	Instagram	Pt	texto,imagens	82	-	1,7	na

Com base neste levantamento, observa-se que os córpus existentes tendem a ser baseados principalmente nos domínios Reddit e Twitter, e que praticamente todos recursos identificados são do tipo textual e dedicados ao idioma inglês. A exceção é o já citado estudo em Mann et al. [2020], que se destaca por ser o único a considerar o domínio Instagram e baseado no idioma português, e também por incluir dados de imagens. Além disso, observa-se que o córpus SMHD é o único que contempla casos de transtorno de ansiedade. É importante observar, entretanto, que diversos métodos de detecção de transtornos de saúde mental discutidos na próxima seção fazem uso de dados multimodais (e.g., imagens e características da rede social etc.) mas que este tipo de informação, via de regra, não é parte integrante de um córpus publicamente disponibilizado para pesquisa.

No que diz respeito ao volume de dados, observa-se que os córpus considerados apresentam grande variação de tamanho (aqui medida em quantidade de usuários diagnosticados), e que recursos baseados em dados da rede social Reddit tendem a ser maiores. Isso é explicável pela organização dos dados em grupos de discussão (e.g., sobre depressão etc.) nessa plataforma e, consequentemente, pela maior facilidade de identificação de usuários diagnosticados. Dados no domínio Twitter, por outro lado, são consideravelmente menos estruturados, o que torna mais complexa a tarefa de seleção de indivíduos com transtorno de saúde mental.

No projeto dos córpus em discussão, observa-se de modo geral uma grande variação também na proporção entre usuários de controle e diagnosticados, indicada na coluna C/D da tabela, e que é um aspecto crucial da modelagem do problema de classificação a ser resolvido. Dentre os estudos selecionados, o córpus CLPsych-2015 é o único que se limita a um balanceamento artificialmente ideal, criado com o objetivo de simplificar o uso de métodos de AM no contexto da competição (ou *shared task*) de mesmo nome, enquanto que em todos os demais procurou-se considerar de alguma forma a necessidade de modelar um cenário mais realista. Dentre as estratégias adotadas, observa-se a modelagem de um grupo de controle de dimensão amplamente superior ao do grupo de diagnosticados [Yates et al. 2017, Cohan et al. 2018, Losada et al. 2017, 2018], o uso de múltiplas seleções de usuários aleatórios a partir de uma grande massa de dados [Shen et al. 2017], ou simplesmente o uso de inventários clínicos [Mann et al. 2020] como forma de selecionar usuários livres de depressão (a um custo possivelmente elevado, refletido no menor volume de usuários obtidos, porém dispensando o uso de um grupo de controle aleatorizado).

Outra questão fundamental no projeto de córpus deste tipo é o pareamento entre usuários diagnosticados e suas contrapartidas no grupo de controle, já que diferenças indesejáveis entre os dois conjuntos podem levar ao aprendizado de padrões espúrios. Assim, como forma de assegurar que os dois grupos sejam minimamente comparáveis, o córpus CLPsych-2015 se destaca como único a garantir pareamento de gênero e faixa etária entre usuários de ambas as classes, o que minimiza possíveis diferenças de vocabulário entre os dois grupos. O pareamento entre classes também é uma preocupação do córpus RSDD e sua extensão SMHD, em que usuários diagnosticados e de controle são selecionados com base no seu comportamento na rede social (e.g., grupos de discussão dos quais participam, frequência de postagens etc.). Este tipo de pareamento é um quesito importante no caso da plataforma Reddit, mas não possui contrapartida direta no domínio Twitter.

Finalmente, é importante destacar que nenhum dos córpus aqui discutidos leva em conta a noção de detecção *precoce* de transtornos, ou seja, anterior ao momento em que o indivíduo é oficialmente diagnosticado por um profissional da área de saúde. Em todos os casos discutidos, os dados constantes do córpus contemplam publicações feitas antes e depois do diagnóstico ou tratamento do indivíduo, e podem assim incluir um número potencialmente grande de indicadores a esse respeito. Por exemplo, uma vez que um indivíduo é diagnosticado ou começa a fazer tratamento para depressão, pode ser natural que ele passe a falar (ou falar mais) sobre este assunto, ou que relate as conversas que tem com o psicólogo, os efeitos da medicação etc. Todos estes indicadores, que em maior ou menor grau estão presentes em todos os córpus aqui discutidos, podem não só reduzir o grau de realismo do desafio computacional, como talvez reduzir a própria utilidade prática do método desenvolvido. Além disso, cabe observar que esta limitação está presente até mesmo nos córpus disponibilizados na série de desafios e-Risk [Losada et al. 2017, 2018, 2019], que mesmo possuindo o propósito explícito de detecção de primeiros sinais de transtornos de saúde mental como depressão, não fazem nenhuma distinção efetiva entre dados publicados antes ou depois do seu diagnóstico oficial, limitando-se apenas ao desafio computacional de classificar estes indivíduos com base no menor volume de dados possível seguindo a ordem cronológica das publicações.

Com base nestas observações, sugere-se assim a oportunidade de estudo e desenvolvimento de um novo recurso linguístico-computacional específico para o português e enfocando a detecção precoce de transtornos de depressão e ansiedade a partir de dados textuais e não-textuais no domínio Twitter[3]. Como um primeiro passo nessa direção, o grupo responsável pela presente pesquisa apresentou em dos Santos et al. [2020a] uma iniciativa de construção de um córpus do tipo pretendido por meio da seleção de usuários de interesse com base em autorrelatos em que há indicação explícita do momento do diagnóstico (e.g., 'Na semana passada o meu psicólogo me informou que eu tenho depressão'). Nesta abordagem, assim como em vários outros estudos da área [Coppersmith et al. 2015, Yates et al. 2017, Cohan et al. 2018], a seleção de usuários é feita de forma cuidadosa consultando-se a plataforma Twitter em busca de autorrelatos que correspondem a uma ampla gama de expressões regulares de interesse, e seguida de inspeção manual.

Diferentemente de estudos existentes, entretanto, na presente abordagem a coleta dos dados (tweets) propriamente dita é feita também com inspeção manual, explorando-se a

[3]A escolha do domínio Twitter é motivada pela sua maior popularidade no Brasil.

ordenação cronológica da plataforma Twitter de modo a restringi-los ao conjunto de dados anteriores ao evento relatado. Um exemplo de como esta porção de dados dita 'útil' (para fins de classificação) é delimitada é ilustrada pelo marcador [end] na tabela 2 com base em um autorrelato indicado pelo marcador [msg].

Tabela 2. Timeline de um usuário com marcador [end] indicando o término da porção de dados a ser considerada na predição de depressão, em que todos os tweets abaixo do ponto [end] são descartados.

Data	Marcador	Texto
Mon March 25		Deixei meu celular em casa e agora a bateria está morta.
Tue March 26		Eu assisti esse filme duas vezes no ano passado.
Thu March 28	[end]	tão feliz que finalmente comprei meus óculos novos LOL
Mon April 1		Vou dormir agora. Amanhã é um grande dia.
Wed April 3		@usuário você nunca me contou isso.
Mon April 8		Pensando em ligar para ela de novo hoje à noite...
Fri May 3		Oiiii! como você está?
Sun May 5	[msg]	Mês passado o psiquiatra me diagnosticou com depressão :(

Neste exemplo, observa-se que no dia 5 de maio (na parte inferior da timeline), o usuário relata um diagnóstico recebido em uma data não especificada do mês anterior (abril). Assim, todos os tweets anteriores ao mês de abril, até o ponto indicado como [end], são coletados para suporte à tarefa de detecção precoce de depressão. O restante dos dados (a partir de 1o. de abril e incluindo o próprio autorrelato em [msg]) são descartados.

Embora o passo adicional de revisão dos dados para delimitar sua porção 'útil' acarrete um aumento discreto nos custos da construção do córpus, o foco nas publicações cronologicamente anteriores a um evento clínico específico é uma abordagem possivelmente inédita na área, e que oferece suporte à construção de modelos computacionais potencialmente mais úteis do ponto de vista da aplicação prática. Em outras palavras, enquanto modelos existentes limitam-se a decidir se um indivíduo é ou não deprimido, um córpus construído da forma aqui sugerida permite o estudo e desenvolvimento de modelos para detectar a evolução do transtorno antes da intervenção clínica, que é o momento a partir do qual a aplicação computacional perde seu propósito.

2.2. Detecção computacional de depressão/ansiedade

A disponibilização de um córpus nos moldes descritos na seção anterior possibilita o emprego de diversos métodos para detecção computacional de transtornos de saúde mental. Nesta seção apresentamos um breve levantamento do estado-da-arte desta área e algumas de suas oportunidades.

Como em outras tarefas de PLN, alguns estudos da área modelam estes indicadores explicitamente na forma de características de aprendizado (em abordagens ditas de engenharia de características), enquanto que outros estudos concentram-se em reconhecer padrões de interesse a partir de dados em estado bruto (em abordagens ditas orientadas a dados). Em ambos os casos, uma ampla gama de métodos de aprendizado é considerada, com uma crescente predominância de modelos neurais.

A tabela 3 categoriza os estudos recentes da área de acordo com o tipo de problema considerado (d=depressão, a=ansiedade, s=sintomas de depressão, v=grau de severidade da depressão, *=outros), tarefa (det=detecção, e.d = detecção precoce, exp=explicação), domínio (Reddit, Twitter etc.)), idioma (En=inglês, Ch=chinês, Pt=português), características de aprendizado (e=embeddings, t=tópicos, n=informações de rede, u=informações referentes ao usuário, l=atributos psicolinguísticos LIWC, p=part-of-speech, d=dicionário de domínio, s=sentimentos/emoções, i=imagens, b=bag of words, m=metadados, t=informação temporal), e métodos de aprendizado (e.g., CNN=redes neurais convolucionais, LSTM=long short-term neural networks, MLP=multilayer perceptron, NN=outras arquiteturas neurais, LR = regressão logística, RF=Random Forest, DT=árvore de decisão etc.) Detalhes adicionais são discutidos a seguir.

Tabela 3. Modelos computacionais de detecção de depressão e ansiedade

Estudo	Alvo	Tarefa	Domínio	Líng.	Atributos	Método
[Yates et al. 2017]	d	det	Reddit	En	e	CNN
[Yazdavar et al. 2017]	s	det	Twitter	En	t	LDA
[Shen et al. 2017]	d	det	Twitter	En	n, u, i, t, d, l, e	LR
[Shen e Rudzicz 2017]	a	det	Reddit	En	e, t, l	MLP
[Loveys et al. 2017]	a,*	det	Twitter	En	s	VADER
[Cohan et al. 2018]	d,a,*	det	Reddit	En	b, e	FastText
[Song et al. 2018]	d	exp	Reddit	En	d, s, t, m, p	MLP+RNN
[Trotzek et al. 2018]	d	e.d	Reddit	En	e, p, m, t, d	CNN
[Nascimento et al. 2018]	d	det	Blogs	Pt	s	DT
[Shen et al. 2018]	d	det	Weibo	Ch	s, p, t, m, i, u, n	NN
[Kumar et al. 2019]	d,a	det	Twitter	En	d, s, t	ensemble
[Aragón et al. 2019]	d	e.d	Reddit	En	s	SVM
[Cacheda et al. 2019]	d	e.d	Reddit	En	t, m, n	RF
[Burdisso et al. 2020]	d	e.d	Reddit	En	b	SS3
[Lin et al. 2020]	d	det	Twitter	En	e, i	CNN
[Yazdavar et al. 2020]	d	det	Twitter	En	b, s, t, i, n, l, u	RF
[Souza et al. 2020b]	d,a	det	Reddit	En	e	LSTM
[Mann et al. 2020]	d.sev	det	Instagram	Pt	i, e	NN

No que diz respeito ao tipo de transtorno a ser detectado, observa-se que estudos focados na depressão (d) tendem a ser muito mais frequentes do que os focados no transtorno de ansiedade (a). Além disso, observa-se que o estudo em Kumar et al. [2019] trata da questão da ansiedade depressiva, um transtorno misto que combina sintomas de depressão e ansiedade, e que o estudo em Souza et al. [2020b] investiga a sua comorbidade. Alguns estudos enfocam ainda a detecção de sintomas [Yazdavar et al. 2017] ou grau de severidade [Mann et al. 2020] destes transtornos.

Considerando-se a tarefa computacional a ser implementada, é possível observar que a maioria dos estudos identificados trata da detecção (det) de depressão/ansiedade ou sua detecção com base no menor volume possível de evidência (e.d), que é uma formulação específica do problema adotada na série de *shared tasks* eRisk [Losada et al. 2019], e/ou

baseada nos córpus derivados destas competições. A preocupação com a necessidade de explicar os passos que levaram ao resultado da classificação é o foco do estudo em Song et al. [2018].

Assim como no caso dos córpus discutidos na seção anterior, os estudos identificados são, com poucas exceções, baseados em dados nos domínios Twitter ou Reddit. Vários destes estudos desenvolvem seus próprios conjuntos de dados (especialmente quando considerando características não textuais da rede social ou outras não disponibilizadas em córpus de uso público), enquanto outros reutilizam um córpus de referência na área, como o RSDD [Yates et al. 2017], SMHD [Cohan et al. 2018] ou eRisk [Losada et al. 2017] no domínio Reddit, ou ainda o córpus de tweets discutido em Shen et al. [2017].

Também conforme as características dos córpus discutidos na seção anterior, a maioria dos estudos identificados é dedicada ao idioma inglês (En). Há entretanto iniciativas isoladas para diversos outros idiomas, e que não são discutidos aqui por razão de brevidade. Estas iniciativas incluem os idiomas árabe [Almouzini et al. 2019], japonês [Tsugawa et al. 2015], coreano [Park et al. 2013], romeno [Briciu e Lupea 2018], russo [Semenov et al. 2015], e tailandês [Katchapakirin et al. 2018]. No caso da língua portuguesa (Pt), identificamos o estudo em Mann et al. [2020], que trata da questão da detecção de depressão em um córpus multimodal no domínio Instagram.

Finalmente, no que diz respeito aos modelos e métodos computacionais empregados, observa-se exemplos de estudos que exploram praticamente todo tipo de informação disponibilizada em redes sociais, incluindo dados textuais, temporais, estruturais da rede, e ainda características demográficas e comportamentais dos usuários, geralmente de forma combinada. A variedade de métodos de aprendizado também é significativa, com uma certa preferência por modelos neurais. Por outro lado, observa-se que nenhum dos estudos identificados faz uso de métodos mais recentes como os modelos de língua pretreinados do tipo BERT [Devlin et al. 2019], ELMo [Peters et al. 2017] e similares, o que representa uma possível oportunidade de melhoria a exemplo do observado em diversas outras áreas do PLN, especialmente considerando-se que modelos deste tipo começam a ser disponibilizados para o português [Souza et al. 2020a, Felix et al. 2020].

3. O córpus SetembroBR versão 3

O córpus SetembroBR é um recurso linguístico-computacional básico que vem se juntar a outras iniciativas do gênero já desenvolvidas pelo grupo responsável para o PLN em português, como os córpus Stars [Teixeira et al. 2014] e Stars2 [Paraboni et al. 2017], dentre outros[4].

O córpus consiste de uma coleção de 22,5 milhões de tweets e 265 milhões de palavras escritas em português por 14.961 usuários únicos. Entre eles, há 2.470 os chamados usuários diagnosticados que revelaram um diagnóstico ou tratamento para depressão, transtorno de ansiedade ou ambos. O restante são usuários de controle (aleatórios). Indivíduos do sexo feminino são igualmente prevalentes nos grupos de Diagnóstico (76,6 %) e Controle (77,07 %). Esse desequilíbrio de gênero pode ser explicado pelo fato de

[4]Estes recursos foram utilizados predominantemente para geração de língua natural, nas tarefas de seleção de conteúdo [Paraboni 2000, Paraboni e van Deemter 2002a,b, dos Santos Silva e Paraboni 2015] e realização superficial [de Lucena et al. 2010, Pereira e Paraboni 2008]

que a seleção do usuário não é verdadeiramente aleatória, ou seja, escolhemos deliberadamente descartar cronogramas de usuário excessivamente curtos, mensagens escritas em idiomas estrangeiros etc.

Além dos dados de texto, também coletamos as listas de seguidores e amigos de todos os usuários do corpus. Isso equivale a uma rede de 13 milhões de usuários únicos e listas de todas as menções de usuários (identificadas pelo caractere @ no texto do Twitter), o que corresponde a uma média de 735 mil menções de usuários em cada classe. Todas as informações relacionadas ao usuário foram totalmente anonimizadas.

O corpus de dados textuais é composto por tweets publicados ao longo de um período de 8 anos, de 2014 ao início de 2021. A maior concentração de dados no período 2017-2019 reflete o fato de a tarefa de coleta de dados ter sido realizada a partir de 2019.

A Tabela 4 apresenta estatísticas descritivas dos subconjuntos Depressão (esquerda) e Ansiedade (direita) do córpus, divididos em informações textuais (topo), temporais (meio) e relacionadas à rede social (parte inferior). As colunas C/D e C/A mostram as relações de números de usuários de Controle / Depressão e Controle / Ansiedade, e têm como objetivo ilustrar o balanceamento de cada conjunto de dados. As proporções devem ser próximas de 7 para o número de usuários, tweets e palavras, e próximas de 1 para as demais estatísticas. Mais detalhes são discutidos a seguir.

Tabela 4. Estatísticas descritivas do córpus SetembroBR v3.

Estatística	Depressão	Controle	Total	C/D	Ansiedade	Controle	Total	C/A
Todos usuários	1283	8981	10264	7	1767	12369	14136	7
Feminino	76.0%	76.0%	76.0%	1	77.4%	77.4%	77.4%	1
Tweets (milhões)	1.96	13.70	15.66	7	2.8	19.96	22.81	7
Palavras (milhões)	23.48	160.82	184.31	6.8	34.90	233.41	268.31	6.7
Tweets/usuários	1528	1526	1526	1	1616	1614	1614	1
Palavras/tweets	12	11.7	11.8	1	12.2	11.7	11.8	1
Dias (média)	522	661	644	1.3	559	610	603	1.1
Dias (máximo)	3712	4163	4163	1.1	4004	4209	4209	1.1
Amigos (média)	733	722	723	1	788	756	760	1
Seguidos (média)	1398	2452	2320	1.8	1062	2381	2216	2.2
Menções (média)	3165	3157	3158	1	3336	3331	3332	1

Nas estatísticas relacionadas ao texto na linha superior da Tabela 4, notamos que todas as razões C/D e C/A estão dentro das expectativas de acordo com o procedimento de balanceamento de corpus discutido na seção anterior. Da mesma forma, as estatísticas relacionadas ao tempo na linha do meio - que se referem ao número de dias decorridos entre a primeira e a última mensagem postada pelos usuários - também são razoavelmente equilibradas.

Em relação às classes Diagnosticado em ambos os subconjuntos, notamos que o intervalo entre a primeira (mais antiga) mensagem e a última (ou seja, a mensagem mais recente antes da data de autorrelato do diagnóstico ou tratamento) é de cerca de 1,5 anos (540 dias). Isso sugere que pode ser possível predizer depressão/ansiedade com ampla

antecedência[5]. Na prática, no entanto, a distribuição do tempo varia consideravelmente de um indivíduo para outro.

Finalmente, a Tabela 5 apresenta a distribuição de usuários diagnosticados em uma série de categorias (condição, tipo de evento, especialista e tratamento medicamentoso) conforme relatado pelos próprios usuários (ou seja, quando disponível). Notamos que o grande número de especialistas rotulados como 'não especificados' não significa falta de suporte clínico (que é um requisito para os dados coletados), uma vez que eventos de diagnóstico/tratamento genuínos podem ser determinados com base em outros tipos de evidências, incluindo hospitalização, prescrição de drogas controladas etc.

Tabela 5. Distribuição de indivíduos diagnosticados conforme condição (depressão/ansiedade), tipo de evento, especialista, e medicação.

Categoria		Depressão	Ansiedade
Evento	diagnóstico	67.2%	62.1%
	tratamento	32.8%	37.9%
Especialista	psiquiatra	22.1%	17.6%
	psicólogo	19.6%	15.5%
	médico	13.3%	25.8%
	neurologista	0.3%	1.1%
	terapeuta	1.2%	0.9%
	não especificado	43.5%	39.1%
Medicação	antidepressivo	22.4%	14.6%
	remédio	18.5%	30.7%
	ansiolítico	0.1%	1.8%
	tarja preta	1.7%	2.0%
	floral	0.0%	0.4%
	não especificado	57.3%	50.5%

Notamos que o uso do termo 'ansiedade' é um pouco menos formal do que o uso de 'depressão', cobrindo uma ampla gama de fenômenos que vão desde eventos individuais relacionados à ansiedade (muitas vezes referidos como 'ataques') até o transtorno de ansiedade generalizada (TAG) propriamente dito. Essa diferença entre as maneiras pelas quais os dois termos são empregados no corpus é observada nos tipos de especialistas envolvidos em eventos relacionados à ansiedade, mais frequentemente representados por médicos de clínica geral (atualmente rotulados como 'médico' e frequentemente referidos como 'médico' ou 'doutor') do que em eventos relacionados à depressão (com um maior envolvimento de profissionais de saúde mental), e também nos tipos de tratamento medicamentoso em consideração (com menos menções ao uso de antidepressivos e mais menções à 'remédio' de forma não específica).

4. Resultados preliminares

Objetivando reportar resultados iniciais de referência para futuros estudos sobre o córpus SetembroBR v3, consideramos a seguir quatro classificadores de texto que fazem

[5]Outros tipos de avaliação de risco de depressão de longo prazo com base em dados desde a infância até a vida adulta são discutidas em, por exemplo, [Lynn et al. 2018].

uso de contagens TF-IDF e word embeddings estáticos e dependentes de contexto. Estes modelos estão resumidos na Tabela 6 e são discutidos a seguir.

Tabela 6. Modelos de referência

Modelo	Método	Características
LogReg	Regressão logística	K-best TF-IDF
LSTM	Long short-term memory network	word embeddings
CNN	Convolutional neural network	word embeddings
BERT	Long short-term memory network	BERT

Todos os modelos recebem como entrada uma amostra de 200 tweets selecionados aleatoriamente de cada usuário, sem substituição[6]. Os parâmetros do modelo são definidos executando-se validação cruzada de 5 partições sobre os dados de treinamento e, subsequentemente, são testados classificando-se uma parte não vista dos dados e computando-se os resultados da média ponderada por classe.

LogReg representa um sistema de baseline do tipo classificador de documento simples, que recebe como entrada um vetor de contagens TF-IDF construído a partir de um conjunto de tweets do usuário concatenados como um único documento. O vetor é reduzido às suas $k = 15.000$ melhores características com o auxílio da seleção de características univariada usando ANOVA e F1 como uma função de escore. O modelo é treinado usando-se regressão logística com pesos de classe balanceados, penalidade L2 e solver lbfgs.

LSTM representa um modelo de rotulagem de sequências padrão com base em uma rede do tipo LSTM que toma como entrada uma representação de documento construída concatenando até 4.000 tokens de cada usuário de entrada, e usando o 5.000 palavras mais frequentes para o vocabulário. A arquitetura consiste em uma camada de embeddings de palavras autotreinadas de tamanho 50 seguida pela função de ativação ReLu. Esta camada LSTM alimenta uma sequência de camadas totalmente conectadas com regularização de dropout e softmax como função de ativação na saída. Uma função de entropia cruzada binária com pesos de classe balanceados é definida como um meio de lidar com o desequilíbrio de classe.

Na mesma linha, CNN também representa um modelo de rotulagem de sequências, e mais uma vez tomando tweets concatenados como entrada. O modelo é baseado na abordagem original de rede convolucional (CNN) em Yates et al. [2017], que foi adaptada do domínio Reddit em inglês para o Twitter em português. As principais mudanças envolveram ajustar os tamanhos dos documentos de entrada e outros parâmetros relacionados ao tamanho do texto, e substituir os embeddings de palavras em inglês por um modelo de CBOW de tamanho 50 disponibilizado em Hartmann et al. [2017]. Em outros aspectos, CNN usa a mesma implementação fornecida em Yates et al. [2017], que consiste de duas camadas convolucionais com pooling médio e 25 filtros cada, e com funções de ativação linear e ReLU na primeira e segunda camadas convolucionais, respectivamente. A parte

[6] A seleção aleatória de mensagens tem o objetivo de evitar suposições sobre a distribuição de mensagens ao longo do tempo.

totalmente conectada é composta por uma camada de 50 neurônios com ativação ReLU e uma camada de saída com softmax.

Finalmente, BERT usa um modelo do tipo BERT [Devlin et al. 2019] que foi pretreinado em texto em Português em Souza et al. [2020a], e ajustado para as presentes tarefas usando a arquitetura LSTM. Devido à limitação de entrada de 512 tokens do modelo BERT, este modelo difere dos anteriores - todos projetados para classificar conjuntos de tweets concatenados - no sentido de que os tweets são classificados individualmente, e apenas as primeiras n = 20 previsões de nível de tweets (de acordo com sua probabilidade) são combinadas para decidir o rótulo da classe final (em nível de usuário). Quanto à arquitetura, os embeddings de BERT são tomados como uma entrada para uma camada LSTM seguida por uma camada de saída totalmente conectada com ativação softmax e regularização de dropout, e usando entropia cruzada binária com pesos de classe balanceados como uma função de perda. O modelo é treinado em no máximo três épocas e, como 80% de todas as mensagens contêm até 30 tokens cada, a entrada para o modelo BERT também é preenchida com zeros totalizando 30 tokens.

A Tabela 7 sumariza os resultados da classificação de depressão e ansiedade obtidos pelos modelos descritos acima.

Tabela 7. . Classificação de depressão e ansiedade medida em F1 ponderada. Os melhores resultados de medida F1 para cada classe são destacados.

	Depressão			Ansiedade		
Modelo	P	R	F1	P	R	F1
LogReg	0.88	0.71	0.76	0.86	0.67	0.73
LSTM	0.79	0.74	0.76	0.79	0.78	**0.78**
CNN	0.83	0.81	0.82	0.82	0.73	0.77
BERT	0.85	0.82	**0.83**	0.83	0.72	0.76

No caso da classificação de depressão (no lado esquerdo da Tabela 7), notamos que BERT e CNN são as alternativas de melhor desempenho geral. A diferença entre BERT e CNN não é estatisticamente significativa, mas a diferença entre BERT e LSTM é ($\chi = 203$, $\alpha = 0,05$, $p = 0$), bem como a diferença entre BERT e LogReg ($\chi = 242$, $\alpha = 0,05$, $p = 0$). No caso de classificação de transtorno de ansiedade (lado direito da Tabela 7), LSTM é a alternativa de melhor desempenho. A diferença entre LSTM e o segundo melhor modelo (CNN) é estatisticamente significativa ($\chi = 390$, $\alpha = 0,05$, $p = 0$).

5. Considerações

Este documento descreveu a construção do corpus SetembroBR versão 3 e suas características. O córpus encontra-se atualmente em uso em estudos de detecção de depressão/ansiedade a partir de dados textuais e não textuais, incluindo questões de interpretabilidade e outras aplicações.

6. Agradecimentos

O primeiro autor é beneficiário de bolsa CAPES nro. 88887.475847/2020-00. Os autores deste trabalho agradecem também ao Centro de Inteligência Artificial (C4AI-USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation.

Referências

Al-Mosaiwi, M. e 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.

Almouzini, S., khemakhem, M., e Alageel, A. (2019). Detecting arabic depressed users from twitter data. *Procedia Computer Science*, 163:257–265.

Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C., e y Gómez, M. M. (2019). Detecting depression in social media using fine-grained emotions. Em *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 1481–1486, Minneapolis, USA. Association for Computational Linguistics.

Aschbrenner, K. A., Naslund, J. A., Grinley, T., Bienvenida, J. C. M., Bartels, S. J., e Brunette, M. (2018). A survey of online and mobile technology use at peer support agencies. *Psychiatric Quarterly*, páginas 1–10.

Birnbaum, M. L., Rizvi, A. F., Correll, C. U., Kane, J. M., e Confino, J. (2017). Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and nonpsychotic mood disorders. *Early Intervention in Psychiatry*, 11(4):290–295.

Briciu, A. e Lupea, M. (2018). Studying the language of mental illness in romanian social media. Em *IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, páginas 21–28.

Brunette, M., Achtyes, E., Pratt, S., Stilwell, K., Opperman, M., Guarino, S., e Kay-Lambkin, F. (2019). Use of smartphones, computers and social media among people with smi: opportunity for intervention. *Community Mental Health Journal*, páginas 1–6.

Bucci, S., Schwannauer, M., e Berry, N. (2019). The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2):277–297.

Burdisso, S. G., Errecalde, M., e 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., e Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *J Med Internet Res*, 21(6):e12554.

Choudhury, M. D., Gamon, M., Counts, S., e Horvitz, E. (2013). Predicting depression via social media. Em *International AAAI Conference on Web and Social Media (ICWSM)*. AAAI.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., e v Goharian (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. Em *27th International Conference on Computational Linguistics*, páginas 1485–1497, Santa Fe, USA. Association for Computational Linguistics.

Coppersmith, G., Dredze, M., Harman, C., Kristy, H., e Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. Em *Second Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 31–39, Denver, USA. Association for Computational Linguistics.

de Lucena, D. J., Paraboni, I., e Pereira, D. B. (2010). From semantic properties to

surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

Devlin, J., Chang, M., Lee, K., e Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Em Burstein, J., Doran, C., e Solorio, T., editores , *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)*, páginas 4171–4186. Association for Computational Linguistics.

dos Santos, V. G., Paraboni, I., e Silva, B. B. C. (2017). Big five personality recognition from multiple text genres. Em *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, páginas 29–37, Prague, Czech Republic. Springer-Verlag.

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

dos Santos, W. R. e Paraboni, I. (2019). Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. Em *Recents Advances in Natural Language Processing (RANLP-2019)*, páginas 1069–1075, Varna, Bulgaria.

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

dos Santos Silva, D. e Paraboni, I. (2015). Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition & Computation*, 15(03):186–225.

Dutta, S. e Choudhury, M. D. (2020). Characterizing anxiety disorders with online social and interactional networks. Em *HCI International 2020 – Late Breaking Papers: Interaction, Knowledge and Social Media*, páginas 249–264, Cham. Springer International Publishing.

Felix, N., Soares, A., e Castro, P. (2020). *Deep Learning for Named Entity Recognition in Legal Domain*. Tese de Doutorado , Universidade Federal de Goiás.

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59 6:1216–29.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., e Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *11th Brazilian Symposium in Information and Human Language Technology - STIL*, páginas 122–131, Uberlândia, Brazil.

Hsieh, F. C., Dias, R. F. S., e Paraboni, I. (2018). Author profiling from facebook corpora. Em *11th International Conference on Language Resources and Evaluation (LREC-2018)*, páginas 2566–2570, Miyazaki, Japan. ELRA.

Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., e Kaewpitakkun, Y. (2018). Facebook social media for depression detection in the thai community. Em *15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, páginas 1–6.

Kumar, A., Sharma, A., e Arora, A. (2019). Anxious depression prediction in real-time social data. Em *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)*, Dehradun, India.

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

Losada, D. E., Crestani, F., e Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. Em *Lecture Notes in Computer Science vol 10456*, páginas 346–360, Cham. Springer.

Losada, D. E., Crestani, F., e Parapar, J. (2018). Overview of eRisk: Early Risk Prediction on the Internet. Em *Lecture Notes in Computer Science vol 11018*, páginas 343–361, Cham. Springer.

Losada, D. E., Crestani, F., e Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. Em *Lecture Notes in Computer Science vol 11696*.

Loveys, K., Crutchley, P., Wyatt, E., e Coppersmith, G. (2017). Small but mighty: Affective micropatterns for quantifying mental health from social media language. Em *Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 85–95, Vancouver, Canada. Association for Computational Linguistics.

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

Mann, P., Paes, A., e Matsushima, E. H. (2020). See and read: Detecting depression symptoms in higher education students using multimodal social media data. Em *Proceedings of the International AAAI Conference on Web and Social Media*, páginas 440–451.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., e Cherry, C. (2016). A dataset for detecting stance in tweets. Em *10th Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia.

Mohammad, S. M., Sobhani, P., e Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Myers, I. B. (1962). *The Myers-Briggs type indicator*. Consulting Psychologists Press.

Nascimento, R., Parreira, P., dos Santos, G., e Guedes, G. P. (2018). Identificando sinais de comportamento depressivo em redes sociais. Em *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, Brazil. SBC.

Naslund, J. A., Bondre, A., Torous, J., e Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, 5:245–257.

Paraboni, I. (1997). Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, PUCRS, Porto Alegre.

Paraboni, I. (2000). An algorithm for generating document-deictic references. Em *Procs. of workshop Coherence in Generated Multimedia, associated with First Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon*, páginas 27–31.

Paraboni, I. (2003). *Generating references in hierarchical domains: the case of Document Deixis*. Tese de Doutorado , University of Brighton.

Paraboni, I. e de Lima, V. L. S. (1998). Possessive pronominal anaphor resolution in Portuguese written texts. Em *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, páginas 1010–1014. Association for Computational Linguistics.

Paraboni, I., Galindo, M., e Iacovelli, D. (2017). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 51(2):439–462.

Paraboni, I. e van Deemter, K. (1999). Issues for the generation of document deixis. Em *Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (esslli99)*, páginas 44–48.

Paraboni, I. e van Deemter, K. (2002a). Generating easy references: the case of document deixis. Em *INLG-2002, New York*, páginas 113–119.

Paraboni, I. e van Deemter, K. (2002b). Towards the generation of document-deictic references. Em *Information sharing: reference and presupposition in language generation and interpretation*, páginas 329–352. CSLI Publications.

Park, S., Lee, S. W., Kwak, J., Cha, M., e Jeong, B. (2013). Activities on facebook reveal the depressive state of users. *J Med Internet Res*, 15(10):e217.

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

Pereira, D. B. e Paraboni, I. (2008). Statistical surface realisation of Portuguese referring expressions. Em *Gotal-2008, Lecture Notes in Artificial Intelligence 5221*, páginas 383–392, Gothenburg, Sweden. Springer-Verlag.

Peters, M. E., Ammar, W., Bhagavatula, C., e Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. Em *Proc. of ACL-2017*, páginas 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Pizarro, J. (2019). Using N-grams to detect Bots on Twitter. Em Cappellato, L., Ferro, N., Losada, D., e Müller, H., editores , *CLEF 2019 Labs and Workshops, Notebook Papers*, página 10. CEUR-WS.org.

Preotiuc-Pietro, D., Liu, Y., Hopkins, D., e Ungar, L. (2017). Beyond binary labels: Political ideology prediction of twitter users. Em *55th Annual Meeting of the Association for Computational Linguistics*, páginas 729–740, Vancouver. Association for Computational Linguistics.

Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., e Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. Em *11th International Conference on Language Resources and Evaluation (LREC-2018)*, páginas 1138–1145, Miyazaki, Japan. ELRA.

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

Seabrook, E. M., Kern, M. L., Fulcher, B. D., e Rickard, N. S. (2018). Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates. *Journal of Medical Internet Research*, 20(5):e168.

Semenov, A., Natekin, A., Nikolenko, S., Upravitelev, P., Trofimov, M., e Kharchenko, M. (2015). Discerning depression propensity among participants of suicide and depression-related groups of vk.com. Em *Analysis of Images, Social Networks and Texts*, páginas 24–35, Cham. Springer International Publishing.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., e Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. Em *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, páginas 3838–3844.

Shen, J. H. e Rudzicz, F. (2017). Detecting anxiety on Reddit. Em *Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 58–65, Vancouver, Canada. Association for Computational Linguistics.

Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., Tang, J., Tiropanis, T., Chua, T.-S., e Hall, W. (2018). Cross-domain depression detection via harvesting social media. Em *Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, páginas 1611–1617. International Joint Conferences on Artificial Intelligence Organization.

Shrestha, A. e Spezzano, F. (2019). Detecting depressed users in online forums. Em *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, páginas 945–951.

Siddiqua, U. A., Chy, A. N., e Aono, M. (2019). Tweet stance detection using an attention based neural ensemble model. Em *NAACL-HLT 2019*, páginas 1868–1873, Minneapolis, USA.

Silva, B. B. C. e Paraboni, I. (2018a). Learning personality traits from Facebook text. *IEEE Latin America Transactions*, 16(4):1256–1262.

Silva, B. B. C. e Paraboni, I. (2018b). Personality recognition from Facebook text. Em *13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122*, páginas 107–114, Canela. Springer-Verlag.

Song, H., You, J., Chung, J.-W., e Park, J. C. (2018). Feature attention network: Interpretable depression detection from social media. Em *32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Souza, F., Nogueira, R., e Lotufo, R. (2020a). BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems (BRACIS) - LNCS 12319*, Cham. Springer.

Souza, V., Nobre, J., e Becker, K. (2020b). Characterization of anxiety, depression, and their comorbidity from texts of social networks. Em *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, páginas 121–132, Porto Alegre, Brazil. SBC.

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., e Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. Em *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, página 12, Avignon, France.

Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., e Yamasaki, A. K. (2014). Generating relational descriptions involving mutual disambiguation. Em *Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403*, páginas 492–502, Kathmandu, Nepal. Springer.

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

Trotzek, M., Koitka, S., e 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*.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., e Ohsaki, H. (2015). Recognizing depression from twitter activity. Em *33rd Annual ACM Conference on Human*

Factors in Computing Systems, páginas 3187–3196, New York, USA. Association for Computing Machinery.

WHO (2017). Depression and Other Common Mental Disorders: Global Health Estimates - Licence: CC BY-NC-SA 3.0 IGO. Technical report, World Health Organization.

Yates, A., Cohan, A., e Goharian, N. (2017). Depression and self-harm risk assessment in online forums. Em *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, páginas 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

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

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