

Desenvolvimento de Ferramenta de Análise de Sentimentos para Identificação de Possíveis Sinais de Comportamento Depressivo na Rede Social Twitter

Luan Mendes Gonçalves Freitas¹, Marcelo Ladeira¹, Marcos Fagundes Caetano¹

¹Departamento de Ciência da Computação (CIC)
Universidade de Brasília (UnB), 70910-900, Brasília, DF, Brazil.

luan.mendes@aluno.unb.br, {mladeira,mfcaetano}@unb.br

Resumo. *O projeto é baseado nos trabalhos dos pesquisadores De Choudhury et al. e Coppersmith et al. para detectar sinais de comportamento depressivo dos usuários a partir da análise das suas postagens no Twitter. São construídas duas bases de tweets, em português, uma da pré-pandemia (01/01/2018 a 31/12/2019), com N=71.232 usuários e uma na pandemia (01/01/2020 a 31/12/2021), com N=70.370 usuários. Essas bases contêm usuários declarados depressivos e usuários não declarados depressivos (controle). São consideradas as seguintes questões de pesquisa: análise de cinco novos atributos na performance dos modelos e a mobilidade de usuários entre as classes depressivo e não-depressivo após a pandemia. As bases de dados são compostas por dez atributos propostos por De Choudhury et al. e os cinco novos atributos. São induzidos modelos de aprendizagem de máquina (classificadores). Os modelos induzidos alcançam desempenhos superiores a performance de modelos propostos por De Choudhury para tweets em língua inglesa.*

1. Introdução

As doenças mentais, entre elas a depressão, tornou-se uma das principais causas de desequilíbrio emocional em todo o mundo [Corbanezi 2018]. Nos dias de hoje, as pessoas estão usando as mídias sociais, como *Twitter*, *Facebook*, *Google+*, entre outras redes, para compartilhar pensamentos e opiniões com seus amigos, familiares ou conhecidos [Park et al. 2012]. A análise das mensagens divulgadas em seus perfis sociais pode servir de base para obter evidências sobre os seus comportamentos.

Usando-se do conhecimento e técnicas de Mineração de Dados (*DM*) juntamente com a literatura de pesquisa em psicologia, psiquiatria, neurociência e sociolinguística sobre saúde humana, tornou-se possível prever possíveis sinais de padrões de comportamento de depressão [De Choudhury et al. 2013] e transtorno de estresse pós-traumático [Coppersmith et al. 2014]. Nesse contexto de transtorno mental de depressão De Choudhury et al. [De Choudhury et al. 2013] propôs um método para detecção de sinais de comportamento depressivo, baseado em atributos construídos a partir dos conteúdos de postagens e das atividades dos usuários do *Twitter*, tornando possível classificá-los como tendo evidência de serem depressivos ou não depressivos.

Esse projeto visa detectar sinais padrão de comportamento depressivo ou não-depressivo em um usuário no *Twitter*, conforme os trabalhos dos pesquisadores De Choudhury et al. [De Choudhury et al. 2013] e Coppersmith et al. [Coppersmith et al.

2014], utilizando-se métodos de Mineração de Dados para construção de modelos de aprendizagem de máquina supervisionado [Dos Santos and Ladeira 2014, Géron 2019]. São objetivos desta pesquisa: (1) Aplicar métodos e técnicas de Mineração de Dados para a construção de 2 bases de dados de *tweets* em português para 2 períodos: pré-pandemia COVID-19 (2018-2019) e pandemia COVID-19 (2020-2021); (2) Utilizar métodos de aprendizagem de máquina supervisionados para gerar um modelo que será incorporado a uma ferramenta de análise de sentimentos capaz de identificar evidências de possíveis sinais de padrões de comportamento depressivo e não depressivo; (3) Realizar uma análise comparativa dos resultados obtidos com os resultados da pesquisadora De Choudhury et al. [De Choudhury et al. 2013], se melhora ou não as previsões dos algoritmos.

De Choudhury et al. [De Choudhury et al. 2013] e Coppersmith et al. [Coppersmith et al. 2014] utilizaram *tweets* em inglês. As contribuições dessa pesquisa são analisar *tweets* em português, avaliar a mobilidade entre os grupos depressivos e não depressivos antes da pandemia COVID-19 (2018 a 2019) e durante a pandemia (2020 a 2021), e avaliar cinco novos possíveis atributos. Os possíveis atributos considerados são a quantidade nos *tweets* de: caracteres orientais (japonês, chinês e coreano), *emojis*, links, mídia (fotos, vídeos e *gifs*), e curtidas.

Esse artigo está organizado da seguinte forma. A Seção 2 descreve trabalhos relacionados ao contexto de detecção de depressão na rede social Twitter. A Seção 3 fornece uma visão geral dos conceitos teóricos e a pesquisa no *Twitter*, mineração de dados e aprendizagem de máquina supervisionada. A Seção 4 apresenta a metodologia utilizada. A Seção 5 apresenta os resultados da exploração de dados nos *tweets* coletados nos 2 períodos pré-pandemia (2018-2019) e pandemia (2020-2021) e a performance dos modelos supervisionados nos dados de validação e de teste das 2 bases de dados. A Seção 6 apresenta as conclusões e sugestões para futuros trabalhos.

2. Trabalhos Relacionados

von Sperling [von Sperling and Ladeira 2019, von Sperling 2019] extrai atributos comportamentais relacionados ao engajamento, emoção, e estilos linguísticos a partir de *tweets* em português para criar uma base de dados de usuários do Twitter que foram diagnosticados com depressão. De Choudhury et al. [De Choudhury et al. 2013] constrói classificadores (precisão = 0,742, recall = 0,629, F1 = 0,681) para identificar usuários com possíveis sinais de depressão, mensurando atributos comportamentais relacionados ao engajamento, emoção, estilos linguísticos, rede social, e menções a medicamentos antidepressivos. Coppersmith et al. [Coppersmith et al. 2014] propõe heurísticas para automatizar partes da construção do conjunto de dados para a depressão, resultando em um conjunto de dados muito maior do que o utilizado por De Choudhury et al., além de ampliar o escopo para outros transtornos.

3. Referencial Teórico

A Rede Social *Twitter* atualmente é uma das principais redes sociais, na qual seus usuários compartilham e trocam mensagens entre si. As mensagens de texto na rede social são denominadas *tweets* e são limitada em 280 caracteres, contendo texto puro, imagens, links para endereços externos e vídeos curtos. A *API* do *Twitter* pode ser usada para recuperar e analisar dados do *Twitter* de forma programática.

Mineração de Dados (*Data Mining (DM)*) é o processo de extrair informação válida, previamente desconhecida, a partir de grandes bases de dados [Géron 2019]. Os dois principais objetivos em Mineração de Dados são a predição e a descrição de dados. A predição envolve o uso de um conjunto de variáveis, para prever os valores desconhecidos ou futuros de uma outra variável objetiva. E a descrição dos dados consiste na descoberta de padrões que os descrevam.

Ao longo dos anos, várias pessoas criaram definições sobre a Aprendizagem de Máquina (em inglês: *Machine Learning (ML)*). Existem vários tipos de Aprendizagem de Máquina, mas o mais adequado para o projeto é a Aprendizagem de Máquina Supervisionada. No projeto usaremos os principais algoritmos supervisionados: Análise Discriminante Linear (*LDA*), Árvore de Decisão (*DT*), Floresta Randômica (*RF*), Gradient Boosting (*GB*), K-Vizinhos mais Próximos (*KNN*), *Naive Bayes (NB)*, Perceptron Multi-Camadas (*MLP*), Regressão Logística (*LR*), *Bagging*, *Boosting* e Votação.

As métricas de avaliação são usadas para avaliar o desempenho de modelos de classificação de aprendizado de máquina que irão comparar as previsões obtidas por um modelo, com os valores reais da base de dados [Géron 2019]. As principais métricas de avaliação são: Precisão, *Recall*, *F1-score* e Acurácia.

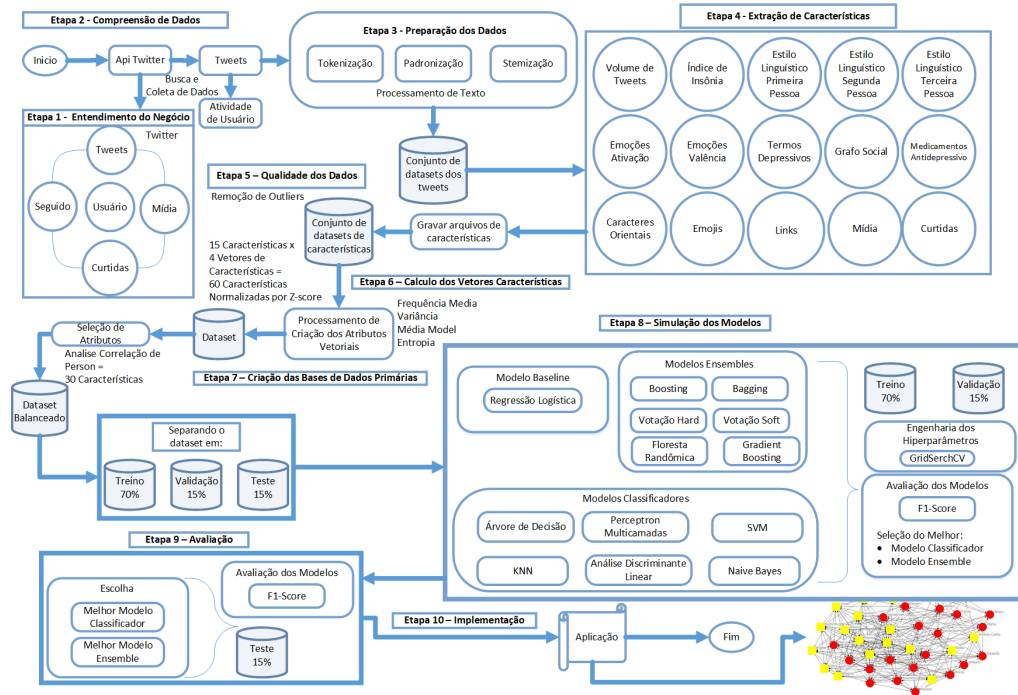
Existem dois pontos a que um classificador pode resultar: *Underfitting* e *Overfitting*. Um classificador pode resultar em *overfitting*, quando ele apresenta um alto grau de precisão para os exemplos de treinamento, e uma alta taxa de erro para o conjunto de teste. Um classificador pode resultar em *underfitting*, quando, provavelmente, o conjunto de treinamento é pouco representativo, logo o modelo não se ajusta ao próprio dado de treino.

4. Metodologia

A metodologia utilizada, ilustrada na Figura 1, é composta de 10 etapas e é baseada no modelo de referência CRISP-DM [Schröer et al. 2021].

1. **Entendimento do Negócio.** Tem como finalidade a compreensão da estrutura dos dados, a *API Twitter* e *framework Snsrape* [JustAnotherArchivist 2022] desenvolvido por terceiros, para coletar *tweets* em um determinado intervalo de tempo.
2. **Compreensão dos Dados.** Consiste em aplicar os serviços da *API* do *Twitter* e *framework* de terceiros para realizar busca e coleta dos *tweets*.
3. **Preparação dos Dados.** Nesta etapa são realizadas duas tarefas: Pré-Processamento (Preparação e Limpeza de dados) e Processamento de Texto (Tokenização, Padronização e Stemização).
4. **Extração das Características.** Consiste na extração dos dados quantitativos das características dos *tweets*, de forma individual. No projeto são definidas 15 séries de atributos, especificados na Seção 4.3.
5. **Qualidade dos Dados.** Essa etapa tem como propósito identificar usuários que estão dentro dos critérios de exclusão (*outliers*), ou seja, com menos de 30 *tweets* postados na conta de *Twitter* ou mais de 300 *tweets* postados em um dia.
6. **Sumarização dos Vetores Características.** Consiste na tarefa de sumarizar cada uma das 15 séries de atributos em 4 métricas vetoriais: frequência média, variância, média móvel e entropia de cada usuário resultando em 60 atributos, com seus respectivos rótulos de classe (depressão e controle).

Figura 1. Metodologia do Projeto

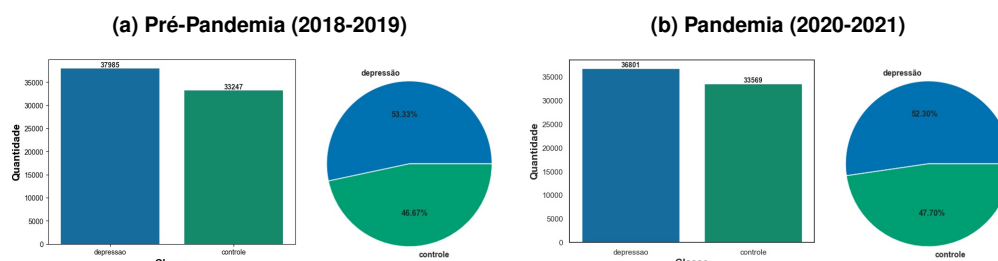


7. **Criação das Bases de Dados Primárias.** Consiste na criação das 2 bases de dados normalizados com a técnica de z-score: pré-pandemia e pandemia e a retirada da identificação dos usuários.
8. **Simulação dos Modelos.** Após a divisão dos *datasets* em treino, validação e teste (70%, 15% e 15%) os dados de treino são aplicados a algoritmos para construção de 13 modelos supervisionados. Os parâmetros desses modelos são refinados com os dados de validação, utilizando validação cruzada com *10-folds*. Esses parâmetros são calibrados pelo método *GridSearchCV*.
9. **Avaliação dos Modelos.** Nesta etapa, os dois modelos que apresentaram melhor desempenho na métrica acurácia, são submetidos ao conjunto de teste, para estimar a performance deles nos dados de teste.
10. **Implementação.** O modelo que obteve o melhor desempenho, tendo como critério de escolha o menor tempo de processamento e o maior nível de acurácia, é incorporado na ferramenta de análise de sentimentos para detectar sinais de comportamento depressivo.

4.1. Compreensão dos Dados

No projeto foi realizado duas coletas de dados de *tweets* públicos postados em dois períodos distintos: de 01 de janeiro de 2018 a 31 de dezembro de 2019 (antes do início da pandemia COVID-19) e de 01 janeiro de 2020 a 31 dezembro de 2021 (época da pandemia). Os dados coletados foram para ambas as classes: depressão e controle (não depressão). Para o período pré-pandemia foram coletados um total de 71.232 usuários (depressivos = 37.985, controle = 33.247) e para o período pandemia foram coletados um total de 70.370 usuários (depressivos = 36.801, controle = 33.569), como destacado na Figura 2.

Figura 2. Distribuição e Balanceamento entre as Classes



Para evitar inconsistências nos dados, os usuários pertencentes à ambas as classes foram eliminados, tornando-as disjuntas.

- **Classe Depressão:** Foram filtrados usuários no *Twitter* que se declaram publicamente depressivos. Seus dados coletados incluem pelo menos *tweets* que constam algumas das frases de autodeclarações de depressão (Tabela 1).
- **Classe Controle:** Para poder validar os dados e aplicá-los nos modelos de aprendizagem de máquina supervisionado que se diferenciam da classe depressão, filtrou-se no *Twitter*, usuários em que não constam nenhum *tweets* com frases de autodeclarações de depressão.

Para realizar a coleta de *tweets* de usuários declarados publicamente depressivos foi utilizado o framework *Snsrape* [JustAnotherArchivist 2022], ao invés da *API* do *Twitter Tweepy* [von Sperling and Ladeira 2019] para filtrar e coletar todos os *tweets* dos usuários nos períodos pré-pandemia e pandemia.

Tabela 1. Palavras-chave da Classe Depressão

"estava com depressão"	"fui diagnosticada com depressão"	"fui diagnosticado depressiva"	"sou deprê"
"estava com deprê"	"fui diagnosticada com deprê"	"fui diagnosticado depressivo"	"tenho depressão"
"estou com depressão"	"fui diagnosticada depressiva"	"fui diagnosticado deprê"	"tenho deprê"
"estou com deprê"	"fui diagnosticada depressivo"	"minha depressão"	"tinha depressão"
"fui depressiva"	"fui diagnosticada deprê"	"minha deprê"	"tinha deprê"
"fui depressivo"	"fui diagnosticado com depressão"	"sou depressiva"	"tô depressiva"
"fui deprê"	"fui diagnosticado com deprê"	"sou depressivo"	"tô depressivo"

Na filtragem de usuários para a classe controle utilizaram-se as mesmas palavras-chave, mas com operador de negação '-' na filtragem de busca, ou seja, *tweets* que não contêm essas palavras-chave.

4.2. Preparação de Dados

Os *tweets* extraídos são pré-processados para transformar o texto de entrada em um modelo padronizado, procedendo da seguinte forma: (1) Tokenização: identificação e listagem de *tokens*; (2) Padronização: normalizar os *tokens* convertendo maiúsculas em minúsculas, remoção de caracteres não latino, *URLs* e *hashtags*; (3) Stemização: Convenção de palavras para a forma raiz da palavra. Logo após o procedimento, é deixado uma cópia do texto original para extração dos atributos: caracteres orientais e *emojis*.

4.3. Extração das Características

Nesta seção apresentaremos as características que são extraídas dos textos e atividades dos usuários candidatos na rede social *Twitter* por dia, pois essas características apresentam o

estado emocional do usuário na interação com outros usuários na rede social *Twitter*. São 15 atributos, sendo do número 1 a 10 são introduzidos pela pesquisadora De Choudhury et al. [De Choudhury et al. 2013] e do número 11 a 15 os novos atributos:

1. **Volume de Tweets.** Quantidade de *tweets* postados;
2. **Índice de Insônia.** Relação de quantidade *tweets* postados no período da noite (21:00-6:00) pela quantidade *tweets* postados no período do dia (6:01- 20:59);
3. **Estilo Linguístico na 1ª pessoa.** Quantidade de palavras na 1ª pessoa do caso reto no singular e plural;
4. **Estilo Linguístico na 2ª pessoa** - Quantidade de palavras na 2ª pessoa do caso reto no singular;
5. **Estilo Linguístico na 3ª pessoa** - Quantidade de palavras na 3ª pessoa do caso reto do singular e plural;
6. **Valência de Emoções.** Cálculo da média de variância (estado emocional), usando-se a base de dados *ANEW-BR* [Kristensen et al. 2011];
7. **Ativação de Emoções.** Cálculo da média de ativação (reação emocional), usando-se a base de dados *ANEW-BR* [Kristensen et al. 2011];
8. **Termos Depressivos.** Cálculo da Média das palavras com valência menor que 4 (palavras negativas) usando a base de dados *ANEW-BR* [Kristensen et al. 2011];
9. **Grafo Social.** Quantidade de respostas de seguidores respondendo a um *tweet*;
10. **Medicamentos Antidepressivos.** Usando-se uma base de dados de medicamentos antidepressivos, calcula-se a frequência dos termos de medicamentos;
11. **Caracteres Orientais.** Quantidade de caracteres no intervalo *unicode* japonês, chinês e coreano.
12. **Emojis.** Quantidade de *emojis* no intervalo *unicode* de *emojis*;
13. **Frequência de Links.** Quantidade de links nos *tweets*;
14. **Mídia.** Quantidade de fotos, vídeos e *gifs* nos *tweets*;
15. **Número de Curtidas.** Quantidade de curtidas nos *tweets*.

4.4. Qualidade dos Dados

No final do processo de extração de características, são excluídos os usuários que possuem mais de 300 postagens de *tweets* em um dia, ou, que possuem menos de 30 *tweets* na conta do *Twitter*.

4.5. Cálculo dos Vetores de Características

Para cada atributo extraído obtém-se a série temporal diária do atributo para cada usuário. Assim, para cada série temporal do atributo são aplicadas 4 métricas estatísticas para geração de um único registro por usuário (Frequência Média, Variância, Média Móvel Ponderada e Entropia). Tendo como consequência a geração de 60 (15x4) atributos para cada usuário com seu rótulo de classe.

4.6. Criação das Bases Dados

As duas bases de dados com 60 atributos são normalizadas pela técnica *Z-Score*, para os atributos ficarem na mesma escala de valor. Logo depois, foi feita engenharia de seleção de atributos utilizando-se a técnica Matriz de Correlação de Pearson, para selecionar os 30 atributos mais significativos, ou seja, baixa correlação entre características, conforme mostrado na Tabela 2, os atributos selecionados.

Tabela 2. Características Seleccionadas

Volume de Tweets Frequência Média	Volume de Tweets Média Móvel Ponderada	Índice de Insônia Variância	Índice de Insônia Média Móvel Ponderada	Pronome 1ª Pessoa Frequência Média	Pronomes 1ª Pessoa Variância
Pronomes 2ª Pessoa Média Móvel Ponderada	Pronomes 2ª Pessoa Entropia	Pronomes 3ª Pessoa Frequência Média	Pronomes 3ª Pessoa Média Móvel Ponderada	Valência Média Móvel Ponderada	Valência Entropia
Ativação Média Móvel Ponderada	Ativação Entropia	Termos Depressivos Variância	Termos Depressivos Média Móvel Ponderada	Grafo Social Variância	Grafo Social Média Móvel Ponderada
Medicamentos Antidepressivos Frequência Média	Medicamentos Antidepressivos Média Móvel Ponderada	Caracteres Orientais Variância	Caracteres Orientais Média Móvel Ponderada	Emojis Variância	Emojis Entropia
Frequência de Links Média Móvel Ponderada	Frequência de Links Entropia	Mídia Variância	Mídia Média Móvel Ponderada	Número de Curtidas Frequência Média	Número de Curtidas Média Móvel Ponderada

4.7. Simulação dos Modelos - Aprendizagem de Máquina

Como estratégia na construção de um modelo de aprendizagem supervisionado eficiente, as bases de dados foram divididas de forma aleatória em: conjunto de treinamento (70%), conjunto de validação (15%) e conjunto de teste (15%) e aplicados nos modelos com validação cruzada de *10-folds*. A Tabela 3, mostra os 2 tipos de algoritmos (classificadores (C) e ensemble (E)) e seus hiperparâmetros adotados através da técnica *GridSearchCV* para se ter melhor desempenho. Assim, o modelo Regressão Logística com os seus hiperparâmetros básicos (*default*) foi escolhido como *baseline* para servir de modelo comparativo entre os modelos.

Os dados e códigos utilizados neste projeto encontram-se no Github¹.

Tabela 3. Modelos Supervisionados Seleccionados

Classificadores		
Algoritmo	Hiperparâmetros Utilizados	Classe Sklearn
Análise Discriminante Linear (LDA)	<code>shrinkage=0, solver='lsqr', tol=1e-6</code>	LinearDiscriminantAnalysis
Árvore de Decisão (DT)	<code>max_depth=9</code>	DecisionTreeClassifier
K-Vizinhos mais Proximos (KNN)	<code>n_neighbors=9</code>	KNeighborsClassifier
Naive Bayes (NB)	<code>alpha=1e-09</code>	BernoulliNB
Perceptron Multi-Camadas (MLP)	<code>hidden_layer_sizes=(10,30,10), max_iter=100</code>	MLPClassifier
Máquina de Vetores de Suporte (SVM)	<code>C=1, max_iter=1000</code>	LinearSVC
Regressão Logística (LR) (Baseline)		LogisticRegression
Ensemble		
Algoritmo	Hiperparâmetros Utilizados	Classe Sklearn
Floresta Randômica (RF)	<code>max_depth=10, n_estimators=90</code>	RandomForestClassifier
Gradient Boosting (GB)	<code>n_estimators=100</code>	GradientBoostingClassifier
Bagging	<code>base_estimator=MLP, n_estimators=10</code>	BaggingClassifier
Boosting	<code>base_estimator=RF, n_estimators=10</code>	AdaBoostClassifier
Votação Hard	<code>estimators=[LDA, DT, RF, GB, KNN, NB, MLP, SVM], voting='hard'</code>	VotingClassifier
Votação Soft	<code>estimators=[LDA, DT, RF, GB, KNN, NB, MLP], voting='soft'</code>	

5. Resultados

Esta seção apresenta os resultados obtidos nos experimentos das bases de dados nos períodos pré-pandemia (2018-2019) e pandemia (2020-2021) com os dados de treino, validação e teste.

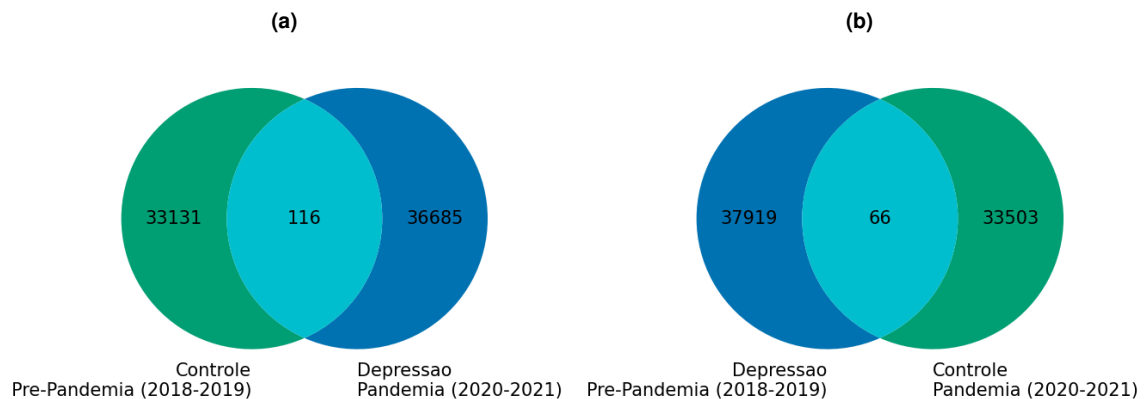
5.1. Exploração de Dados

Explorando-se as bases de dados geradas foi encontrada uma interseção entre as bases: Como mostrado na Figura 3a, no período pré-pandemia foram encontrados 116 usuários que não eram depressivos, mas que tornaram depressivos no período da pandemia. A *COVID-19* no Brasil pode ter sido o gatilho para despertar depressão interna nas pessoas. Não reunimos informações suficientes para comprovar se foi a *COVID-19* ou não. Por

¹<https://github.com/luanfreitas5/UnBSense>

outro lado, na Figura 3b, mostra que no período pandemia 66 usuários não se autodeclararam depressivos, mas que no período da pré-pandemia eram depressivos. De acordo com a literatura sobre psicologia humana, psiquiatria, neurociência e sociolinguística sobre saúde humana, vários fatores não relacionados a *COVID-19* podem causar depressão nas pessoas.

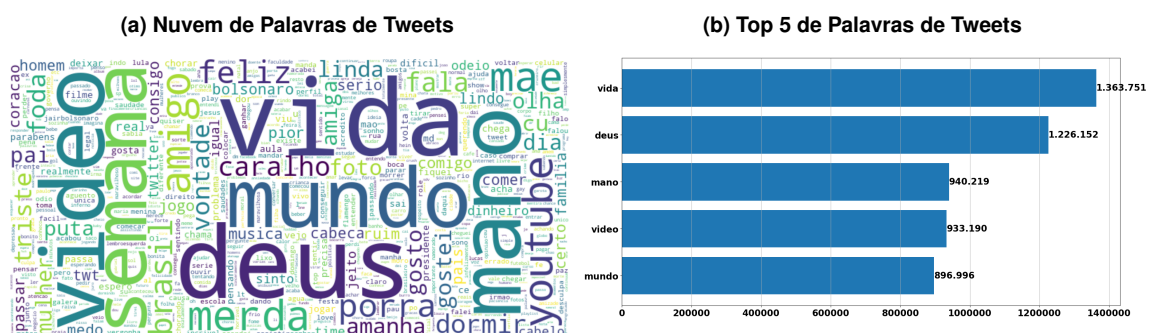
Figura 3. Interseções entre as Base de Dados



5.2. Nuvem Palavras

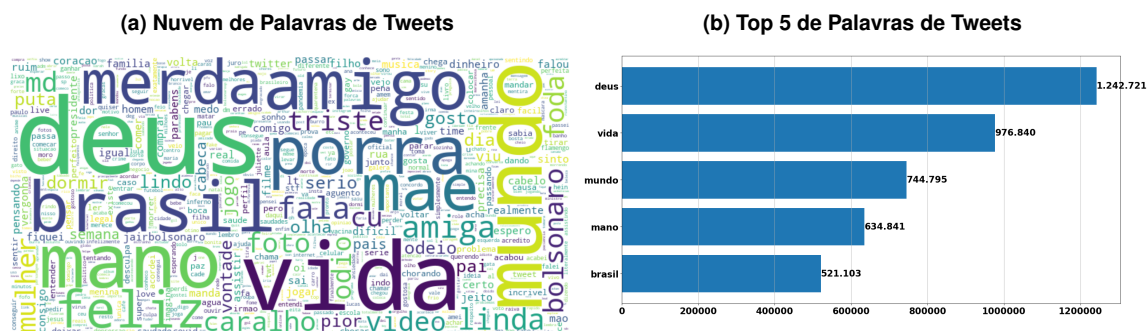
Na Figura 4a, apresenta-se a nuvem de palavras de todos *tweets* postados por usuários candidatos das classes de depressão e de controle no período pré-pandemia; através do Gráfico 4b, analisa-se que os termos mais predominantes usados pelos usuários em seus *tweets* no período pré-pandemia são “vida” com uma taxa de 1.363.751 ocorrências, “deus” com taxa de 1.226.152 ocorrências, “mano” com taxa de 940.219 ocorrências, “video” com taxa de 933.190 ocorrências e “mundo” com taxa de 896.996 ocorrências.

Figura 4. Análise de Tweets de Usuários de Depressão e Controle - Pré-Pandemia



Na 5a, apresenta-se a nuvem de palavras de todos *tweets*, os postados por usuários candidatos da classe de depressão e de controle no período pandemia; através do Gráfico 5b foi analisado que os termos mais predominantes usados pelos usuários em seus *tweets* no período pandemia são “deus” com taxa de 1.242.721 ocorrências, “vida” com uma taxa de 976.840 ocorrências, “mundo” com taxa de 744.795 ocorrências, “mano” com uma taxa de 634.841 ocorrências e “brasil” com taxa de 521.103 ocorrências.

Figura 5. Análise de Tweets de Usuários de Depressão e Controle - Pandemia



5.3. Simulação dos Modelos - Treino e Validação

Nesta seção serão apresentados os resultados dos experimentos realizados com os 2 melhores modelos, um de classificação e um de ensemble pré-selecionados na avaliação dos 13 modelos. A Tabela 4 apresenta o desempenho dos 13 modelos induzidos com validação cruzada 10-folds e seus hiperparâmetros configurados via GridSearchCV nas bases pré-pandemia (2018-2019) e pandemia (2018-2019), somente com os 10 atributos da De Choudhury et al. [De Choudhury et al. 2013], utilizando as métricas de Precisão (P), Recall (R), F1-Score (F1) e Acurácia (ACC) para avaliar e comparar os modelos, tendo como métrica alvo o f1-Score.

Tabela 4. Pré-Seleção dos Modelos

Pré-Pandemia					Pandemia				
Classificadores					Classificadores				
Modelo	Acurácia	Precisão	Recall	F1-Score	Modelo	Acurácia	Precisão	Recall	F1-Score
Perceptron Multicamadas (MLP)	76.1	75.5	87.4	81.0	Perceptron Multicamadas (MLP)	74.5	74.5	87.4	80.4
Árvore de Decisão (DT)	74.7	74.1	86.9	80.0	Árvore de Decisão (DT)	72.9	72.9	87.2	79.4
KNN	70.5	70.6	84.7	77.0	SVM	65.9	65.9	89.2	75.8
SVM	68.3	68.9	83.1	75.3	KNN	68.0	69.3	83.5	75.7
Análise Discriminante Linear (LDA)	66.9	67.3	84.2	74.8	Análise Discriminante Linear (LDA)	65.0	64.9	90.7	75.6
Regressão Logística (LR) (Baseline)	58.3	58.3	100.0	73.7	Regressão Logística (LR) (Baseline)	59.9	59.9	100.0	74.9
Naive Bayes (NB)	59.1	71.3	49.9	58.7	Naive Bayes (NB)	57.8	70.9	50.3	58.9
Ensemble					Ensemble				
Modelo	Acurácia	Precisão	Recall	F1-Score	Modelo	Acurácia	Precisão	Recall	F1-Score
Gradient Boosting (GB)	77.2	75.5	90.1	82.2	Gradient Boosting (GB)	75.5	74.4	90.0	81.5
Boosting	77.3	76.2	88.8	82.0	Boosting	75.6	75.0	88.8	81.3
Floresta Randômica (RF)	76.6	74.8	90.1	81.8	Floresta Randômica (RF)	74.8	73.5	90.6	81.2
Bagging	76.5	75.9	87.6	81.3	Bagging	75.1	75.2	87.3	80.8
Votação Soft	76.1	75.8	86.7	80.9	Votação Hard	74.5	74.4	87.7	80.5
Votação Hard	75.7	75.7	85.8	80.5	Votação Soft	74.1	74.4	86.4	80.0

Os modelos *Gradient Boosting (GB)* e Perceptron Multi-camadas (MLP) foram se destacam melhor entre os modelos e serão usados nos experimentos para analisar quais dos 5 novos atributos (caracteres orientais, *emojis*, mídias, frequência de links e número de curtidas) tem efeito significativos na interação com os 10 atributos da pesquisadora De Choudhury et al. [De Choudhury et al. 2013] (volume de *tweets*, índice de insônia, estilo linguístico de 1ª, 2ª e 3ª pessoa, emoções de valência e ativação, termos depressivos, grafo social e medicamentos antidepressivos).

A Tabela 5, apresenta o desempenho dos modelos *Gradient Boosting (GB)* e Perceptron Multicamadas (MLP) induzidos com os dados de validação das 2 bases de dados pré-pandemia (2018-2019) e pandemia (2020-2021). Os experimentos são realizados induzindo um modelo com os 10 atributos da pesquisadora De Choudhury et al. [De Choudhury et al. 2013] interagindo com cada um dos 5 novos atributos, se um atributo melhorar o desempenho do modelo no resultado do experimento anterior ele é

atribuído junto com os 10 atributos da pesquisadora no experimento seguinte e excluído na lista de interações, repetindo até resultado aceitável.

Tabela 5. Experimentos de Validação (Pré-Pandemia e Pandemia)

Modelo	Pré-Pandemia (2018-2019)																			
	Atributos De Choudhury				Atributos De Choudhury + Caracteres Orientais				Atributos De Choudhury + Emojis				Atributos De Choudhury + Links				Atributos De Choudhury + Mídia			
	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1
GB	77.2	75.5	90.1	82.2	77.2	75.5	90.2	82.2	77.3	75.9	89.5	82.1	77.0	75.4	89.9	82.0	77.3	75.6	90.1	82.2
MLP	76.1	75.5	87.4	81.0	76.4	76.0	87.0	81.1	76.3	75.9	86.9	81.0	76.3	75.7	87.5	81.2	76.1	75.6	87.2	81.0

Modelo	Pandemia (2020-2021)																			
	Atributos De Choudhury				Atributos De Choudhury + Caracteres Orientais				Atributos De Choudhury + Emojis				Atributos De Choudhury + Links				Atributos De Choudhury + Mídia			
	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1
GB	75.5	74.4	90.0	81.5	75.5	74.4	90.0	81.5	76.0	74.8	90.3	81.8	75.6	74.5	90.0	81.5	75.4	74.5	89.7	81.4
MLP	74.5	74.5	87.4	80.4	74.2	74.0	87.8	80.3	75.1	75.1	87.5	80.8	74.6	74.6	87.4	80.5	74.5	74.2	88.1	80.6

Nenhum dos novos atributos teve alto desempenho significativo com os atributos da pesquisadora Choudhury et al. [De Choudhury et al. 2013] nos modelos *Gradient Boosting* (GB) e Perceptron Multi-camadas (MLP), tendo somente pouquíssimo desempenho em relação ao f1-score dos modelos do experimento com os atributos da De Choudhury et al. [De Choudhury et al. 2013]. Então devemos complementar que os 5 novos atributos (caracteres orientais, emojis, mídias, frequência de links e número de curtidas) não são úteis para construção de um modelo promissor capaz de detectar sinais de padrão de comportamento depressivo e não depressivo.

5.4. Avaliação dos modelos - Teste

Nesta seção serão apresentados os resultados dos experimentos de testes (somente com os atributos da De Choudhury et al. [De Choudhury et al. 2013]) com o f1-score médio dos experimentos dos modelos *Gradient Boosting* (GB) e Perceptron Multi-camadas (MLP) das bases de dados pré-pandemia (2018-2019) e pandemia (2020-2021).

Na Tabela 6, o modelo *Gradient Boosting* (GB) na base pré-pandemia (2018-2019) teve desempenho de f1-score de 81.6% e na base pandemia (2020-2021) teve desempenho de f1-score de 79.9%. O modelo Perceptron Multi-camadas (MLP) na base pré-pandemia (2018-2019) teve desempenho de *f1-score* de 80.6% e na base pandemia (2020-2021) teve desempenho de *f1-score* de 79.4%.

Tabela 6. Experimentos de Teste (Pré-Pandemia e Pandemia)

Modelo	Pré-Pandemia (2018-2019)				Pandemia (2020-2021)				Média F1-Score
	Experimento - 10 Atributos De Choudhury				Experimento - 10 Atributos De Choudhury				Experimento -
	Acurácia	Precisão	Recall	F1-Score	Acurácia	Precisão	Recall	F1-Score	10 Atributos De Choudhury
Gradient Boosting (GB)	76.8	75.5	88.9	81.6	74.2	73.2	87.8	79.9	80.75
Perceptron Multicamadas (MLP)	75.8	75.4	86.5	80.6	74.2	74.2	85.4	79.4	80.0

Conclui-se que os modelos *Gradient Boosting* (GB) com *f1-score* médio de 80.75% e Perceptron Multicamadas com *f1-score* médio de 80.0% nos experimentos de teste, obtiveram desempenho promissores em relação a pesquisadora De Choudhury et al. [De Choudhury et al. 2013] que obteve desempenho de *f1-score* médio de 68.0%, de acordo com a literatura. Logo, a incrementação desses novos atributos podem ser úteis para melhoramento de modelos de aprendizagem de máquina no objetivo na detecção de possíveis sinais de padrão de comportamento depressivos ou não depressivos de novos usuários do *Twitter* [Géron 2019].

6. Discussão/Conclusão

Neste projeto demonstrou-se o potencial de usar a rede social *Twitter* como uma ferramenta para medir e detectar sinais de padrões de comportamento depressivo em postagens e atividades dos usuários no Twitter, tendo como base os trabalhos de De Choudhury et al. [De Choudhury et al. 2013] e Coppersmith et al. [Coppersmith et al. 2014]. Usando-se o *framework Snsrape* para filtrar e coletar *tweets* em português de usuários do Twitter (de forma anônima) que se autodeclararam depressivos e aplicando-se métodos e técnicas de Mineração de Dados construíram-se 2 bases de dados de 2 períodos compostos de 10 atributos da De Choudhury et al. [De Choudhury et al. 2013] (volume de *tweets*, índice de insônia, estilo linguístico de 1ª, 2ª e 3ª pessoa, emoções de valência e ativação, termos depressivos, grafo social e medicamentos antidepressivos) de 5 novos atributos (caracteres orientais, *emojis*, mídia, frequência de links e número de curtidas): o período pré-pandemia (01/01/2018 a 31/12/2019) foram coletados um total de 71.232 usuários (depressivos = 37.985, controle = 33.247) e pandemia (01/01/2020 a 31/12/2021) período pandemia foram coletados um total de 70.370 usuários (depressivos = 36.801, controle = 33.569).

Foi feita uma análise comparativa nas duas bases de dados para verificar se houve aumento ou diminuição dos casos de depressão por consequência da pandemia *COVID-19* no Brasil. Realizamos uma análise exploratória quantitativa dos usuários depressivos nas 2 bases de dados, observamos que no período da pandemia o número de relatos de usuários se autodeclarando depressivo (ex. "eu tenho depressão") é mais que dobro do número de relatos de usuários no período da pré-pandemia. A *COVID-19* pode ter sido a causa de depressão nas pessoas. Não reunimos informações suficientes para comprovar se foi a *COVID-19* ou não.

Na compreensão e exploração dos dados nos *tweets* coletados, através do ponto de vista da pré-pandemia encontramos 116 usuários que não estavam depressivos antes do início da *COVID-19*, mas que se tornaram depressivos na pandemia. Não reunimos informações suficientes para comprovar se foi a *COVID-19* ou não. Por outro lado, do ponto de vista da pandemia encontramos 66 usuários que já eram depressivos antes da existência da pandemia *COVID-19*, isso pode ter ocorrido de acordo com a literatura sobre psicologia humana, psiquiatria, neurociência e sociolinguística sobre saúde humana que mapearam vários fatores que podem desencadear um quadro depressivo.

Na simulação de teste o modelo *Gradient Boosting (GB)* obteve *f1-score* médio 81.0% e o modelo Perceptron Multicamadas obteve *f1-score* médio de 80.2%. Esses resultados demonstram que a adição de novos aumentou a qualidade da eficiência das predições em comparação com os modelos da pesquisadora De Choudhury et al. [De Choudhury et al. 2013] que obteve uma média *f1-score* de 68.0%. Então esses dois modelos são ótimos para serem incorporados numa ferramenta de análise de sentimentos com o intuito de auxiliar as pessoas que necessitam de ajuda na detecção de possíveis sinais de padrões relevantes de comportamento depressivo ou não depressivo.

Como trabalhos futuros, podem-se acrescentar novos atributos correlacionados a aspectos sociais, comportamentais e outros tipos de transtornos mentais (ansiedade, trauma pós-parto, suicídio, etc.) de outras redes sociais nas bases de dados. Assim, sugere-se a aplicação de outros algoritmos como por exemplo, aprendizagem de redes neurais *RNN (Recurrent Neural Network)* ou *LSTM (Long short term memory)* ou apren-

dizagem profunda *Transformers* para aumentar o desempenho da predição de sinais de padrão comportamental, tendo como o resultado suporte para tomada de decisão para governantes e gestores públicos com intuito de criar novas políticas públicas para amparar a população que tem esses transtornos mentais.

Agradeço ao Programa de Iniciação Científica 2021-2022 da Universidade de Brasília (UnB) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudos e auxílio financeiro que possibilitou a dedicação integral ao projeto do estudo.

Referências

- Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Corbanezi, E. (2018). Transtornos Depressivos e Capitalismo Contemporâneo. *Caderno CRH*, 31:335–353. Publisher: Universidade Federal da Bahia - Faculdade de Filosofia e Ciências Humanas - Centro de Recursos Humanos.
- De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting Depression via Social Media. In Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P., and Soboroff, I., editors, *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press.
- Dos Santos, F. L. and Ladeira, M. (2014). The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset. In *2014 Brazilian Conference on Intelligent Systems*, pages 50–54.
- Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books, 2 edition.
- JustAnotherArchivist (2022). Snscape. original-date: 2018-09-09T20:16:31Z.
- Kristensen, C. H., Gomes, C. F. d. A., Justo, A. R., and Vieira, K. (2011). Normas brasileiras para o Affective Norms for English Words. *Trends in Psychiatry and Psychotherapy*, 33:135–146. Publisher: Associação de Psiquiatria do Rio Grande do Sul.
- Park, M., Cha, C., and Cha, M. (2012). Depressive moods of users portrayed in Twitter. *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012*, pages 1–8.
- Schröer, C., Kruse, F., and Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181:526–534.
- von Sperling, O. (2019). UnB Sense : a web application to probe for signs of depression from user profiles on social media. Accepted: 2021-02-01T16:59:26Z.
- von Sperling, O. and Ladeira, M. (2019). Mining Twitter Data for Signs of Depression in Brazil. In *Anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 25–32. SBC. ISSN: 2763-8944.