UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

# ROTULAÇÃO DE SINTOMA DE DEPRESSÃO UTILIZANDO APRENDIZADO ATIVO E PROCESSAMENTO DE LINGUAGEM NATURAL

**Rafael Vinicius Polato Passador**

**Orientadora: Profa. Dra. Helena de Medeiros Caseli**

São Carlos - SP
16 de janeiro de 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

# ROTULAÇÃO DE SINTOMA DE DEPRESSÃO UTILIZANDO APRENDIZADO ATIVO E PROCESSAMENTO DE LINGUAGEM NATURAL

**Rafael Vinicius Polato Passador**

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Federal de São Carlos, para a obtenção do título de bacharel em Ciência da Computação.
Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos - SP
16 de janeiro de 2024

*Dedico este trabalho a toda minha família e aos meus amigos, que me apoiaram nessa caminhada, e aos meus professores, pelo incontável conhecimento compartilhado.*

*"Devemos mudar nossa atitude tradicional em relação à construção de programas. Em vez de imaginar que nossa principal tarefa é instruir o computador sobre o que ele deve fazer, vamos imaginar que nossa principal tarefa é explicar a seres humanos o que queremos que o computador faça".*

*(Donald E. Knuth)*

# AGRADECIMENTOS

Ao Prof. Dr. Ivandré Paraboni (EACH/USP) pela co-orientação deste trabalho e pela disponibilização do *corpus* (SANTOS; FUNABASHI; PARABONI, 2020) e dos rótulos gerados pelo GPT-3.5 (SANTOS; PARABONI, 2024) utilizados nos experimentos apresentados neste trabalho.

Ao Augusto Mendes por compartilhar o modelo de classificação denominado nesse trabalho de BERTimbau-Amive, usado nas rotulações.

# RESUMO

A abordagem de cuidado dos distúrbios psicológicos, especialmente a depressão e a ansiedade, é vista como uma das maiores preocupações atuais em saúde mental. As redes sociais, como Twitter, não apenas permitem que indivíduos mantenham contato e promovam apoio mútuo, como também são frequentemente objeto de pesquisas que visam identificar indivíduos com potenciais perfis depressivos ou classificar postagens depressivas. Nesse contexto, observa-se que grande parte dos sistemas atuais de Processamento de Linguagem Natural (PLN) opera com base em modelos cujo sucesso está intimamente ligado à qualidade e ao montante de dados de treinamento específicos disponíveis. Contudo, adquirir quantidades substanciais de dados anotados é geralmente um processo custoso, sobretudo considerando a natureza árdua e complexa de rotulação para atividades de PLN. Neste cenário, algumas abordagens foram propostas na tentativa de mitigar o custo de geração de conjuntos de treinamento de qualidade. O Aprendizado Ativo (*Active Learning*) procura alcançar elevada precisão com um número reduzido de dados anotados, permitindo que um algoritmo de aprendizado sugira quais observações o especialista deve rotular para serem utilizados no processo de treinamento. Neste projeto, realiza-se uma investigação inicial das estratégias de aprendizado ativo utilizando um comitê de modelos na classificação automática de um sintoma de depressão (tristeza ou humor depressivo) em postagens do Twitter, utilizando o *corpus* SetembroBR, alcançando um valor F1 de até 10 pontos percentuais superior com a *query* de *Consensus Entropy* em relação à amostragem aleatória.

**Palavras-chave**: depressão, Twitter, saúde mental, *active learning*, aprendizado ativo

# ABSTRACT

The approach to caring for psychological disorders, especially depression and anxiety, is seen as one of the current major concerns in mental health. Social networks, such as Twitter, not only allow individuals to maintain contact and promote mutual support, but are also often the subject of research aiming to identify individuals with potential depressive profiles or to classify depressive posts. In this context, it is observed that most of the current Natural Language Processing (NLP) systems operate based on models whose success is closely linked to the quality and amount of specific training data available. However, acquiring substantial amounts of annotated data is generally a costly process, especially considering the arduous and complex nature of labeling for NLP activities. In this scenario, some approaches have been proposed in an attempt to mitigate the cost of generating quality training sets. Active Learning seeks to achieve high accuracy with a reduced number of annotated data, allowing a learning algorithm to suggest which observations the expert should label to be used in the training process. In this project, an initial investigation of active learning strategies using a model commitee in the automatic classification of a symptom of depression (sadness or depressive mood) in Twitter posts is carried out, using the SetembroBR corpus, achieving an F1 value up to 10 percentage points higher with the Consensus Entropy query compared to random sampling

**Keywords**: depression, Twitter, mental health, active learning

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# SUMÁRIO

# 1 INTRODUÇÃO

A abordagem e cuidado dos distúrbios psicológicos, especialmente a depressão e ansiedade, notórios entre eles, é vista como uma das mais cruciais preocupações atuais em saúde mental. Conforme a Pesquisa Nacional de Saúde de 2019[1], no Brasil, 10,2% dos adultos reportaram ter sido diagnosticados com depressão por um especialista em saúde mental.

Neste panorama, a interação em redes sociais online (RSO) ganha importância especial. Estas plataformas não apenas permitem que indivíduos mantenham contato e promovam apoio mútuo, mas também oferecem um espaço vital para expressar emoções e pensamentos, funcionando como um termômetro emocional para detectar tendências de saúde mental em larga escala.

Para se estabelecer um diagnóstico de depressão maior, é necessário que a pessoa exiba, no mínimo, cinco sintomas durante um período de duas semanas ou mais. Um destes sintomas deve ser necessariamente um estado de ânimo deprimido ou um desinteresse ou falta de prazer marcantes. Os outros sintomas podem variar entre alterações significativas no apetite ou no peso, problemas de sono (como dormir demais ou de menos), inquietação ou lentidão observável nos movimentos, uma sensação persistente de cansaço ou falta de energia, sentimentos de inutilidade ou culpa exagerada, dificuldade para se concentrar ou se decidir, e a presença de pensamentos frequentes sobre a morte. De acordo com o DSM-5 (American Psychiatric Association et al., 2014), esses sintomas devem ser intensos o suficiente para provocar sofrimento real ou prejudicar o funcionamento normal do indivíduo em âmbitos sociais, de trabalho ou outros aspectos críticos da vida dele.

A detecção precoce da depressão é fundamental para a intervenção e o acompanhamento por especialistas em saúde mental. Em um contexto mais amplo, perceber padrões depressivos em populações poderia ajuda a entender o problema de forma mais abrangente, oferecendo subsídios para formular políticas de saúde mental mais eficazes.

Na atualidade, o acompanhamento da evolução dos sintomas depressivos é frequentemente realizado por meio de questionários clínicos como o PHQ-9 (KROENKE; SPITZER; WILLIAMS, 2001) e a Escala de Hamilton, utilizados por médicos. Estes instrumentos, contudo, enfrentam restrições como a necessidade de o paciente buscar ativamente por ajuda e a confiabilidade na recordação dos próprios sentimentos e emoções ao longo do tempo. Um método adicional e complementar para identificar indícios de depressão pode ser a análise de conteúdos postados em redes sociais, que muitas vezes refletem nuances do estado emocional dos usuários.

[1] Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=29270&t=resultados>

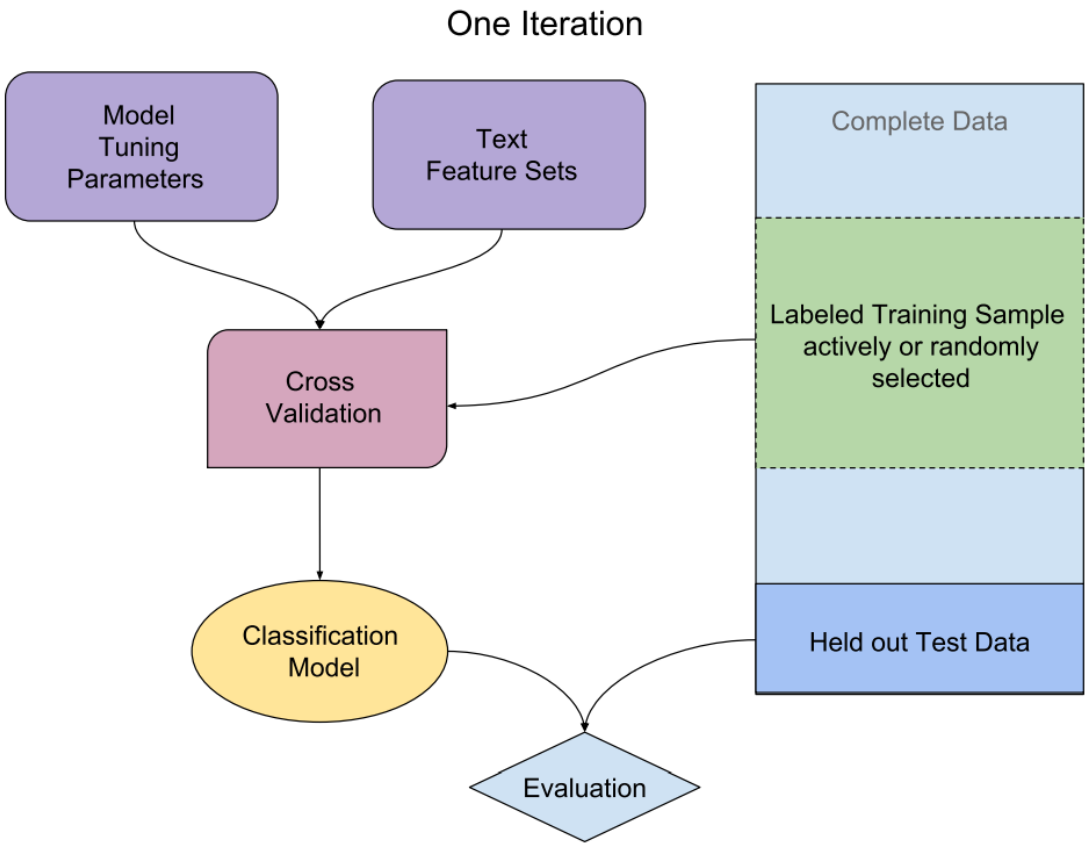
Assim, as RSO se revelam como ferramentas valiosas no reconhecimento de indivíduos com possível perfil depressivo (PPD), contribuindo eficazmente na alocação de recursos para o tratamento de saúde mental. Este aspecto é destacado em várias pesquisas (CHOUDHURY; COUNTS; HORVITZ, 2013; MOWERY et al., 2016; JAMIL et al., 2017; YADAV et al., 2020; ZIWEI; CHUA, 2019; YADAV et al., 2020; SANTOS; FUNABASHI; PARABONI, 2020; GISSIN; SHALEV-SHWARTZ, 2019; MANN; PAES; MATSUSHIMA, 2020; FATIMA et al., 2018) que não só apontam as RSO como meios mais eficientes de identificar PPD em comparação com métodos tradicionais utilizados em saúde mental, como questionários de autoavaliação, como também atuam na classificação de postagens depressivas.

As RSO, como Twitter e Reddit, são frequentemente objeto de estudo dessas pesquisas, conforme evidenciado pelos trabalhos mencionados anteriormente. Em sua maioria, os pesquisadores adotam técnicas de Processamento de Linguagem Natural (PLN) e algoritmos de aprendizado de máquina com o objetivo de analisar e interpretar o conteúdo postado pelos usuários. Estas abordagens permitem detectar nuances no texto que podem indicar padrões associados a perfis ou sintomas depressivos. A vastidão e a natureza dinâmica do conteúdo gerado pelos usuários nas plataformas de redes sociais oferecem uma oportunidade única para o monitoramento em tempo real, contribuindo para uma intervenção mais precisa e em tempo hábil. Dessa forma, é possível afirmar que o estado da arte já comprova a viabilidade de usar as técnicas de PLN combinadas a algoritmos de aprendizado de máquina para realização dessas tarefas.

Grande parte dos sistemas atuais de PLN opera com base em modelos cujo sucesso está intimamente ligado à qualidade e ao montante de dados de treinamento específicos disponíveis. Embora estes sistemas possam exibir bons desempenhos com supervisão apropriada, adquirir quantidades substanciais de dados anotados é geralmente um processo custoso, sobretudo considerando a natureza árdua e complexa de rotulação para atividades de PLN (ZHANG; STRUBELL; HOVY, 2022). Neste cenário, algumas abordagens foram propostas na tentativa de mitigar o custo de geração de conjuntos de treinamento de qualidade, entre elas o método de aprendizado ativo (do inglês, "*active learning*" ou AL). O AL procura alcançar elevada precisão com um número reduzido de amostras anotadas, permitindo que um algoritmo de aprendizado sugira quais instâncias o especialista deve rotular para serem utilizadas no processo de treinamento, reduzindo a rotulação de instâncias que não são informativas para o algoritmo de classificação, otimizando o uso dos recursos humanos dispendiosos (MILLER; LINDER; MEBANE, 2020). A Figura 1 exemplifica a aplicação dessa técnica.

Dessa forma, nesse exemplo de interação de aprendizado ativo, que é uma abordagem semi-supervisionada para treinamento de modelos de aprendizado de máquina, um modelo inicial é treinado com um conjunto pequeno de dados rotulados. Este modelo é, então, utilizado para fazer previsões sobre um conjunto de dados não rotulados. As previsões para as quais o modelo tem menos certeza são selecionadas para rotulação manual.

Figura 1 – Exemplo de uma interação do Active Learning




Fonte: (MILLER; LINDER; MEBANE, 2020)

Neste fluxograma, os parâmetros de afinação do modelo e os conjuntos de características de texto são as entradas iniciais para um processo de validação cruzada (*cross validation*), que é uma técnica para avaliar a generalização do modelo. A partir da validação cruzada, o modelo de classificação é treinado e, então, aplicado à amostra de treinamento rotulada ativamente ou selecionada aleatoriamente, utilizando uma fração do conjunto de dados completo, o qual é subdividido em amostra de treinamento e amostra de teste.

A performance do modelo é, então, avaliada, o que encerra o ciclo da iteração e fornece *insights* para a próxima iteração do aprendizado ativo, como a seleção de novos exemplos para serem rotulados ou ajustes nos hiperparâmetros. Este ciclo iterativo continua até que o modelo alcance uma performance satisfatória ou que o orçamento, ou seja, a oferta de recursos disponíveis, de rotulação seja esgotado.

Para realização desse estudo, foi utilizado o *corpus* SetembroBR (SANTOS; OLIVEIRA; PARABONI, 2023), disponibilizado pelo Prof. Dr. Ivandré Paraboni (EACH/USP). O conjunto de dados contém todos os *tweets* públicos escritos por 3.900 usuários únicos (ou seja, excluindo mensagens escritas por outros usuários, conhecidas como *retweets*) que auto relataram um diagnóstico ou tratamento para um transtorno mental, mas que foram postados antes de um diagnóstico ou tratamento auto-relatado para depressão e/ou ansiedade. O SetembroBR foi

gerado com o objetivo de fornecer dados de treinamento e teste para modelos de aprendizado de máquina supervisionados. Nesse contexto, os *tweets* dos usuários depressivos não possuem rotulação de sintomas, de maneira que cada usuário pode, ou não, ter realizado postagens na mídia social com conteúdo que caracterizaria-se como um sintoma de depressão.

Dessa forma, o intuito dessa pesquisa foi avaliar como o aprendizado ativo pode ser utilizado em um grande conjunto de dados para construção e otimização de um comitê de modelos de classificação binária capaz de prever sintomas depressivos em postagens de mídias sociais. Assim como reiterado em Ren et al. (2021), a aquisição de uma grande quantidade de dados anotados de alta qualidade consome muita mão de obra, tornando esse processo algo inviável em campos que requerem altos níveis de especialização (como reconhecimento de fala, extração de informações, imagens médicas etc.). Paralelamente, a crescente utilização do *Deep Learning* e de modelos de linguagem pré-treinados tem seu sucesso fundamentado na grande quantidade de dados anotados publicamente disponíveis. Por conseguinte, é natural investigar se o AL pode ser usado para reduzir o custo de anotação.

Dessa forma, a estratégia adotada foi um modelo *Human-in-the-loop*, ou seja, um processo no qual a intervenção humana é integrada a um sistema de aprendizado de máquina, em que um especialista humano avalia e fornece rótulos ou anotações para os dados selecionados pelo modelo para o retreinamento dos classificadores. A seleção das amostras é feita a partir de estratégias de consulta baseadas em informatividade, as quais atribuem uma medida de informatividade a cada instância não rotulada individualmente, como probabilidade e entropia.

Essa estratégia de seleção (denominada *query*), é uma das mais comumente utilizadas, baseando-se na incerteza das previsões dos modelos de classificação Zhang, Strubell e Hovy (2022). Nesse contexto, construiu-se um comitê de modelos para levar em consideração a saída de classificadores distintos, ao invés de apenas considerar um único classificador, aplicando a técnica de *query-by-committee*.

Para realizar tal investigação preliminar, fez-se necessário uma primeira rodada de rotulação inicial, a fim de obter um conjunto de dados rotulados com algum sintoma depressivo para treinamento dos modelos iniciais de classificação. Nesse ínterim, foi utilizado o modelo de classificação binária disponibilizado pelo Augusto Mendes (MENDES; CASELI, submitted), que detecta presença de um dos sintomas definidos para o diagnóstico de depressão-maior usada no PHQ-9: tristeza/humor depressivo. Assim, com as predições geradas por tal modelo sendo utilizados como rótulos do *corpus* de treinamento, foi possível criar um comitê de modelos iniciais para classificação desse sintoma e diferentes técnicas de aprendizado ativo foram investigadas na redução dos custos de anotação manual.

Assim, este trabalho de conclusão de curso tem como objetivo avaliar a efetividade do aprendizado ativo (*active learning*) no auxílio à rotulação manual de *tweets* contendo o sintoma tristeza/humor depressivo no *corpus* SetembroBr.

Este trabalho está dividido em 5 capítulos, incluindo esta introdução. No Capítulo 2 são descritos os principais trabalhos relacionados a esta pesquisa. O Capítulo 3 descreve o *framework* proposto, os modelos utilizados e as técnicas de aprendizado ativo investigadas. O Capítulo 4 traz resultados de performance dos modelos gerados, interpretação desses resultados e análise comparando as diferentes metodologias de seleção de amostragem utilizadas. Por fim, o Capítulo 5 apresenta conclusões resultantes dos experimentos, bem como direcionamento de possíveis aspectos a serem explorados em trabalhos futuros.

# 2 TRABALHOS RELACIONADOS

Este Capítulo descreve os trabalhos mais relacionados à pesquisa desenvolvida neste trabalho de conclusão de curso, resumidos na Tabela 1 e com sua principais características listadas na Tabela 2. Como critério de seleção, buscou-se trabalhos relativamente recentes (últimos 4 anos) e que possuiam o objetivo semelhante ao desse estudo, reduzir o grande custo de anotação acarretado pela grande quantidade de dados necessária para o treinamento dos modelos de Deep Learning e de linguagem.

Em um estudo sobre a aplicação de AL em modelos de linguagem baseados em transformadores BERT (DEVLIN et al., 2019), Ein-Dor et al. (2020) realizam uma comparação de estratégias de AL considerando um cenário real, no qual dispõe-se de um conjunto de dados altamente desbalanceado e com escassez de dados rotulados. Para isso, os autores descrevem a aplicação de uma técnica chamada aprendizado ativo baseado em "pool" usando modelos BERT para classificação.

O aprendizado ativo baseado em pool, realizado em lotes (*batch mode*), é um método onde se seleciona um conjunto de amostras de um grande "pool" (conjunto `U`) de dados não rotulados para rotulação. O objetivo é escolher amostras que, uma vez rotuladas, ofereçam o maior benefício possível para o treinamento do modelo de aprendizado de máquina. Dessa forma, tal modelo é treinado inicialmente com um conjunto limitado de 100 exemplos rotulados (a semente inicial `L`). Em seguida, em uma série de iterações, estratégias de aprendizado ativo – como confiança mínima (*Least Confidence*, LC), Monte Carlo Dropout (MD), Ensemble de Perceptron (PE), Comprimento do Gradiente Esperado (EGL), Core-Set e aprendizado ativo Discriminativa (DAL) – são usadas para escolher lotes adicionais de exemplos não rotulados (do conjunto `U`) para serem rotulados e adicionados ao conjunto de treinamento. Isso é feito para melhorar progressivamente a performance do modelo. Como comparação, foi realizada também a seleção aleatória de amostras do conjunto de dados sem rotulação.

Dentre as estratégias aplicadas pelo estudo, destacam-se:

- **Confiança Mínima (LEWIS; GALE, 1994):** Seleciona instâncias para as quais o modelo tem menos certeza da previsão de acordo com a regra de decisão de entropia máxima.
- **Monte Carlo Dropout (GAL; GHAHRAMANI, 2016):** Similar ao LC, mas a incerteza da instância é calculada usando Dropout de Monte Carlo[1] em 10 ciclos de inferência,

[1] Ao realizar inferência, aplica-se o dropout (desligamento aleatório de neurônios) repetidamente, gerando múltiplas "passagens"e previsões. Analisando a variação dessas previsões, é possível obter uma medida de incerteza. No contexto de active learning, essa incerteza ajuda a selecionar os dados mais informativos para rotular, otimizando o processo de aprendizado

com a função de aquisição de entropia máxima.

- **Conjunto de Perceptrons (PE):** Seleciona instâncias com a maior incerteza – de maneira semelhante ao LC – mas fazendo uma média sobre um conjunto de modelos, semelhante à um comitê.

- **Comprimento do Gradiente Esperado (HUANG et al., 2016):** Seleciona instâncias com o maior valor de gradiente, pois espera-se que exerçam grande influência sobre o modelo.

- **Core-Set (SENER; SAVARESE, 2018):** Seleciona instâncias que melhor cobrem o conjunto de dados no espaço de representação aprendido (CLS), usando o método guloso descrito em Sener e Savarese (2018).

- **Aprendizado Ativo Discriminativo (GISSIN; SHALEV-SHWARTZ, 2019):** Esta abordagem visa selecionar instâncias que tornem L mais representativo de todo o conjunto, seguindo o método utilizado em (GISSIN; SHALEV-SHWARTZ, 2019).

Ein-Dor et al. (2020) avaliaram 10 conjuntos de dados com temáticas diversas e considerando diferentes tarefas de classificação, utilizando as métricas de acurácia e F1. Como resultado, foi aferido que as estratégias de aprendizado ativo melhoraram o F1 da base aleatória de 4 a 8% em média. Esses resultados demonstraram que a AL pode de fato aprimorar os resultados do BERT quando o orçamento para anotações é pequeno, especialmente para conjuntos de dados altamente desbalanceados que têm uma baixa prioridade para exemplos positivos, como é o caso em muitas configurações do mundo real, inclusive a usada nessa pesquisa.

Paralelamente, Mamooler et al. (2022) reforçam a dificuldade de rotular grande quantidades de dados que pertencem a domínios específicos, necessitando de especialistas e de grande quantidade de recursos e esforço para realizar tal tarefa. Dessa forma, os autores demonstram a viabilidade da utilização do Active Learning para melhorar o aprendizado dos modelos de linguagem com uma quantidade menor de dados anotados, criando um pipeline que visa o solucionar o entrave da realização de ajuste fino (*fine-tuning*) em modelos pré-treinados, como BERT e RoBERTa, que tendem a apresentar uma performance não satisfatória em pequenos conjuntos de dados de treinamento.

Nesse contexto, evidencia-se que as estratégias de AL existentes frequentemente exigem um conjunto de amostras rotuladas para iniciar, o que, ainda sim, é caro para adquirir. Em vista de superar este alto custo, Mamooler et al. (2022) propuseram uma estratégia baseada em agrupamento para reduzir o esforço necessário para criar o conjunto inicial de amostras anotadas, considerando um cenário no qual tem-se um baixo orçamento disponível para anotação de especialistas e um conjunto de dados desbalanceado.

A estratégia em questão consta na adaptação do modelo RoBERTa para a tarefa específica de classificação de textos legais, seguindo por um processo denominado de destilação de conhecimento (*knowledge distillation*), no qual o conhecimento de um modelo maior e mais complexo (o professor), que foi treinado a partir do *corpus* Cer et al. (2017) *STS Benchmark*, é transferido para um modelo menor (o aluno). O modelo resultante, chamado de DisTAPT RoBERTa, é capaz de gerar *embeddings* de sentenças com significados semânticos mais precisos e comparáveis com a tarefa desejada. Posteriormente, os autores propõem usar o algoritmo KMeans para agrupar (*clustering*) as amostras de texto não rotuladas. O KMeans é um algoritmo de *clustering* que agrupa dados tentando separar amostras em $n$ grupos de igual variância, minimizando um critério conhecido como inércia, ou soma de quadrados dentro do grupo[2].

Diante desse pipeline, são utilizadas quatro técnicas de seleção do aprendizado ativo: seleção aleatória de amostras, seleção baseada na incerteza do modelo (Hard-Mining e Perceptron Dropout) e uma abordagem que usa classificação binária para escolher amostras representativas (DAL).

Os resultados indicam que o pipeline proposto atinge uma eficácia comparável à de métodos totalmente supervisionados, porém com um custo significativamente reduzido em termos de anotação. No dataset *Contract-NLI* Koreeda e Manning (2021), o modelo DisTAPT RoBERTa atinge um F1-score de 0,6508 com apenas 40 amostras rotuladas, enquanto o modelo totalmente supervisionado usando todo conjunto alcança 0,6990. Paralelamente, no conjunto de dados *LEDGAR* Sener e Savarese (2018) o F1-score com ajuste fino totalmente supervisionado é de 0,9538 para RoBERTa-base, enquanto DisTAPT RoBERTa obtém um F1-score de 0,9321 com apenas 60 amostras rotuladas.

No contexto das redes sociais, Farinneya et al. (2021) abordam a identificação de boatos em mídias sociais e apontam que esta tarefa é um desafio especialmente dada a natureza de não moderação dessas plataformas e a escassez de dados rotulados para treinar modelos. Os autores, então, propõem uma estratégia de Aprendizado Ativo Transferido (ATL) para identificar boatos com dados anotados limitados, explorando o impacto de várias abordagens de aprendizado de máquina e representações contextuais.

Nesse contexto, os autores atacam o problema de classificar rumores em *tweets* através de uma estratégia *Human-in-the-loop*, na qual combina-se cenários de aprendizado ativo com diversos modelos de classificação textuais. Dentre as estratégias aplicadas pelo estudo, destacam-se a de confiança mínima, que afere a incerteza das previsões dos modelos, *Query by Committee* (QBC), que mede a discordância entre o comitê dos modelos e o Epsilon-Guloso, uma estratégia utilizada em Aprendizado por Reforço para equilibrar a exploração (experimentar

[2] Ao invés de escolher amostras iniciais aleatoriamente de todo o conjunto de dados, os autores sugerem selecionar amostras a partir dos centroides dos *clusters* (medoides) gerados pelo KMeans, reduzindo o conjunto de candidatos a serem rotulados e aumentando a representatividade.

coisas novas) e a explotação (usar o que já se sabe) através do cálculo de probabilidade epsilon.

Nos modelos utilizados, os autores combinam modelos linguísticos de transformação vetorial do estado da arte, como BERT, GloVe e TweetBERT Qudar e Mago (2020), sendo este último um modelo de linguagem treinado em 680 milhões de *tweets*, com classificadores comumente utilizados na literatura, como Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Random Frests (RF), Logistic Regression (LR), Árvores de Decisão Ada-Boosted (ADA), K-Nearest Neighbors (KNN) e Classificadores Gaussianos (GP). Assim, os modelos foram treinados em um *dataset* que continha 6425 *tweets* contendo 2402 rumores e 4023 não rumores, iniciando-se com 20 amostras escolhidas aleatoriamente e adicionando, em cada iteração, 50 amostras.

Para a avaliação, os autores utilizaram a métrica F1 para comparar as performances dos modelos, em que a combinação de TweetBERT com Regressão Logística (TweetBERT+LR) e a estratégia de Lote Classificado por confiança mínima (BATCH-LC) foi identificada como a melhor configuração. Esta configuração alcançou, em média, 78,93% de F1, utilizando 75% dos dados disponíveis para amostragem, superando em 2% o valor alcançado usando amostras aleatórias. A maioria dos modelos mostrou um grande ganho com 100-200 amostras de dados escolhidas através do Aprendizado Ativo, e um pequeno relativamente menor quando a quantidade de amostras é superior à 200.

Além disso, Farinneya et al. (2021) também investigaram o impacto de introduzir *tweets* de outros tópicos na performance dos modelos, concluindo que essa prática é prejudicial, diminuindo a média em até 19,1% de F1, enquanto utilizam apenas 25% dos dados.

Em Sahan, Smidl e Marik (2021), os autores buscam uma alternativa para contornar o entrave da complexidade e custo da anotação supervisionada de textos por meio de um estudo comparativo de diversas estratégias de aprendizado ativo para diferentes conjuntos de dados textuais. A tarefa investigada foi a de detecção de notícias falsas, utilizando *word embeddings*, focando em métodos de aprendizado ativo bayesianos, devido a sua capacidade de representar a incerteza de um modelo na realização de predições. Tal abordagem é focada na rotulação de dados para os quais o classificador prevê grande incerteza. A incerteza é quantificada usando funções de aquisição, como variância preditiva ou entropia preditiva. Embora funções de aquisição diferentes frequentemente forneçam resultados semelhantes, diferentes representações da distribuição preditiva produzem resultados mais diversos.

Os experimentos foram realizados para duas tarefas: categorização de texto (Kaggle News Category e Twitter Sentiment140) e detecção de notícias falsas (Kaggle Fake News e Fake News Detection datasets). O experimento inicia com uma escolha aleatória de 10 amostras de um conjunto de 1000 documentos de texto. Para cada experimento, são simulados 200 pedidos de anotação. Diante desse cenário, são comparados métodos como SGLD, que adiciona ruído adicional ao gradiente no descente de gradiente estocástico; Dropout MC, uma extensão do dropout comum que amostra a máscara binária na etapa de previsão; ensembles

profundos, consistindo de várias redes treinadas em paralelo de condições iniciais diferentes; e incerteza Softmax, um método simples usando apenas uma rede para estimar um valor único de uma previsão para maximizar a entropia. Não obstante, também é realizada uma seleção aleatória de amostras para reiterar a eficácia do uso de *active learning*.

Dessa forma, foram comparadas três metodologias, através da observação da AUC (Area Under The Curve), para geração de *embeddings*: Fast Text, LASER e o modelo BERT modificado RoBERTa. Como resultado, os autores demonstram que a abordagem convencional de Dropout Monte Carlo fornece bons resultados para a maioria das tarefas. Os métodos de ensemble oferecem uma representação mais precisa da incerteza, o que permite manter o ritmo de aprendizado de um problema complexo com o aumento do número de solicitações, superando o Dropout a longo prazo. No entanto, para a maioria dos conjuntos de dados, a estratégia ativa usando Dropout MC e Deep Ensembles alcançou um desempenho satisfatório mesmo com um número muito baixo de amostras. Os melhores resultados foram obtidos com as *embeddings* oriundas do modelo RoBERTa, contudo todas as estratégias de aprendizagem ativo superaram a seleção aleatória de amostras, alcançando valores de AUC superiores de 0.015 até 0.080 em relação a estratégia de amostra aleatória.

Paralelamente, Jacobs et al. (2021) tinham como objetivo investigar se o AL poderia ser utilizado para reduzir os custos e esforços de anotação dos dados, mantendo uma performance similar ao modelo treinando no conjunto de dados por completo. Os resultados mostraram que o aprendizado ativo baseado em incerteza pode proporcionar um desempenho melhor do que a amostragem aleatória em conjuntos de dados reduzidos. No entanto, essa diferença não foi consistente em toda a curva de treinamento: em alguns pontos, o aprendizado ativo superou a amostragem aleatória e em outros alcançou um desempenho semelhante. O BALD foi a função de consulta com o melhor desempenho geral, possivelmente porque é a única que mede a incerteza do modelo. Dessarte, os autores compararam diversas estratégias de *query* de aprendizado ativo, testando-as no modelo do estado da arte BERT.

Especificamente, o artigo explora estratégias de aprendizado ativo baseadas em incerteza. Essas estratégias, como Razão de Variação, Entropia Preditiva e Aprendizado Ativo Bayesiano por Desacordo (BALD), são projetadas para identificar os pontos de dados mais informativos para rotulação. Ao focar em pontos de dados que o modelo tem mais incerteza, a eficiência do processo de aprendizado é aprimorada, potencialmente reduzindo a quantidade total de dados necessária para o treinamento.

Sobre o conjunto de dados e a metodologia, os pesquisadores começam com uma amostra inicial de dados de 5% do conjunto total. Dois *datasets* foram utilizados para validar e comparar o desempenho das diferentes implementações de AL. O primeiro conjunto de dados utilizado foi o Stanford Sentiment Treebank (SST) (SOCHER et al., 2013). SST consiste em 215.154 frases de filmes com rótulos de sentimento refinados na faixa de 0 a 1. Essas frases estão contidas nas árvores de análise de 11.855 frases. O segundo conjunto de dados utilizado

consiste nas descrições de empresas localizadas em Utrecht e contém 2.212 exemplos divididos em 15 classes. Este ponto de partida é usado para iniciar o processo de aprendizado ativo. O estudo investiga diferentes tamanhos de pool de consulta, variando de 0,5% a 5% do conjunto de dados. Essa variabilidade no tamanho do pool permite examinar como diferentes volumes de seleção de dados impactam no processo de aprendizagem. No entanto, o número total de rodadas no ciclo de aprendizado ativo não é explicitamente declarado.

A eficácia das estratégias aplicadas é avaliada usando duas métricas: precisão e uma versão modificada da métrica de deficiência. A precisão é uma medida direta do desempenho do modelo, enquanto a métrica de deficiência oferece uma visão de quanto o desempenho do modelo está aquém de um cenário ideal. Dentre as estratégias, a BALD foi a que obteve menor deficiência e acurácia semelhante às demais técnicas.

Schröder, Niekler e Potthast (2022) realizaram um estudo cujo objetivo principal foi explorar a eficiência do AL em contextos onde coletar dados rotulados para aprendizado de máquina é custoso e demorado. O foco estava em minimizar os custos de rotulação ao empregar uma abordagem de aprendizado ativo. Isso envolve consultar um oráculo (como um especialista humano) para rotular instâncias de problema selecionadas que são consideradas mais informativas para a próxima iteração do algoritmo de aprendizagem. O estudo visou avaliar o desempenho dessa abordagem usando transformadores em benchmarks amplamente utilizados para classificação de texto e explorar se certas abordagens de *query* baseadas em incerteza poderiam superar a estratégia de consulta de entropia de previsão, que é amplamente utilizada nesse contexto.

Assim, observou-se 5 estratégias de aprendizado ativo: *Prediction Entropy (PE)*, que seleciona instâncias com a maior entropia na distribuição de rótulos previstos, *Breaking Ties (BT)*, analisando as instâncias com a menor margem entre as duas probabilidades mais prováveis, considerando um contexto multiclasse, ou seleciona as amostras semelhante a PE, em um contexto binário, *Least Confidence (LC)*, selecionando as instancias com menor confiança de acordo com o modelo, *Contrastive Active Learning (CA)*, que contabiliza a média máxima de Kullback-Leibler (KL) e *Random Sampling (RS)*, que aleatoriza a seleção. Nesse ínterim, a rotina do AL consiste em utilizar 25 instâncias de treinamento para treinar o primeiro modelo, seguido por 20 iterações de aprendizado ativo, com cada iteração envolvendo a consulta e rotulação de 25 instâncias adicionais.

Foram utilizados 5 conjunto de dados de classificação distintos já comumente utilizados nessa tarefa: AG's News (ZHANG; ZHAO; LECUN, 2016), Customer Reviews (HU; LIU, 2004), Movie Reviews (PANG; LEE, 2005a), Subjectivity (PANG; LEE, 2005b), e TREC-6 (LI; ROTH, 2002). Para a avaliação da tarefa, foram consideradas as acurácias de modelos oriundos do ajuste fino de arquiteturas BERT e DistilRoBERTa, utilizando as diversas técnicas de aprendizado ativo, bem como o treinamento com o conjunto de dados completo – denominada classificação passiva.

Como resultado, a maior discrepância entre a aprendizado ativo e a classificação de texto passiva é observada no AG's News, que também é o maior conjunto de dados do qual os modelos de aprendizado ativo usam menos de 1% para treinamento. Nos demais casos, todos os modelos obtiveram resultados próximos ou até superam o estado da arte, usando apenas entre 0,4% e 14% dos dados. Vale ressaltar que o LC alcança o melhor resultado de precisão para dois conjuntos de dados, enquanto a PE e a abordagem CA apresentam o melhor desempenho em apenas um conjunto de dados cada.

Em Corvino et al. (2022), os autores apresentam uma aplicação para melhorar a eficiência dos modelos de Reconhecimento de Entidades Nomeadas (NER) integrando conhecimento e experiência humanos através de uma abordagem Human-in-the-Loop (HITL). Eles focam no método Query by Committee (QBC) para aprendizado ativo, propondo uma heurística para relaxamento do QBC para equilibrar entre o desacordo do modelo e algum nível de acordo para rotulação de dados mais eficiente.

Neste contexto, o Query by Committee (QBC) utiliza o conceito de desacordo: Se uma determinada instância é classificada de maneira diferente pela grande maioria dos modelos M, então determinar o rótulo correto para essa instância ajuda a reduzir a entropia. Na prática, essa instância é informativa e uma boa candidata para rotulação humana.

Dessa maneira, Corvino et al. (2022) para avaliar experimentalmente o QBC, utilizam três conjuntos de dados. Os conjuntos Laptops e Restaurants, em inglês, são de Pontiki et al. (2015), adaptados para tarefas de entidades nomeadas em análises de sentimentos baseadas em aspectos. As entidades representam aspectos de produtos e serviços, com categorias indicando o sentimento (neutro, positivo, negativo). Já o conjunto de dados de Gazetas do Governo, em português, do jornal governamental brasileiro, inclui nomes de pessoas e cargos em agências governamentais. Cada conjunto tem divisões pré-definidas para treino e teste.

Utilizou-se um comitê de modelos de NER baseados em cinco modelos de linguagem neural pré-treinados, com estratégia de *fine-tuning* e simulação de processo Human-in-the-loop em várias iterações para treinamento, em que foram anotados 125 instâncias para o dataset de laptops, 200 para restaurantes e 250 para o governamental. Os modelos para os conjuntos em inglês e português são diferentes, incluindo BERT, RoBERTa, DistilBERT, BERTimbau, entre outros.

Como resultado, foi aferido F1-score dos modelos. Todos os modelos do comitê obtiveram maiores valores de F1 em relação aos modelos bases treinados, em que, no dataset em português, o modelo BERTimbau atingiu uma métrica de 90%, superando em 10% o valor base. Além disso, os autores decorrem sobre uma limitação de uma simulação de QBC, na qual, apesar do viés dos modelos ser reduzido através do uso de comitês, ainda há presença do erro humano, que pode discordar das anotações das instâncias.

Em um estudo recente, Lemmens e Daelemans (2023) combinam aprendizado ativo

com base na incerteza de previsão com adaptação de tarefa não supervisionada por meio de Modelagem de Linguagem com Máscara (MLM), com o objetivo de reduzir os custos de anotação em *corpus* de mídias sociais.

Em relação aos conjuntos de dados, o primeiro consiste em postagens do Facebook (LJUBEšIć; FIšER; ERJAVEC, 2019) relacionadas a discursos de ódios de dois tópicos específicos: imigrantes e comunidade LGBTQIAP+. O segundo consta com postagens do Reddit Demszky et al. (2020) divididas em 28 emoções e anotadas de maneira multiclasse. O último consta com postagens de notícias divididas em 20 tópicos diversos. Já dentre as estratégias de aprendizado ativos aplicadas, utilizam-se as mesmas medidas baseadas em incerteza aplicadas em Schröder, Niekler e Potthast (2022), comparando-as com o modelo base treinado em todo conjunto de dados e com a inclusão de amostras aleatórias.

Os resultados nesses três conjuntos de dados diferentes (dois *corpora* de mídias sociais, um conjunto de dados de referência) mostram que, com apenas uma fração dos dados de treinamento disponíveis, esta abordagem atinge pontuações F1 similares às alcançadas pelo modelo treinado com todos os dados utilizando, respectivamente, 29,7% e 54% de todos os dados anotados nos conjuntos de dados FRENK e 20 News Groups. Além disso, superam a amostragem aleatória em até 2% de F1.

A Tabela 1 resume os trabalhos relacionados, enquanto a Tabela 2 resume as principais características das estratégias de AL empregadas por eles. Por fim, a Tabela 3 explicita quais foram as melhores estratégias de AL e a melhora de desempenho em relação a amostragem aleatória, em cada um desses trabalhos.

Tabela 1 – Trabalhos relacionados

| ID | Citação | Título | Ano |
|---|---|---|---|
| 1 | (FARINNEYA et al., 2021) | Active Learning for Rumor Identification on Social Media | 2021 |
| 2 | (MAMOOLER et al., 2022) | An Efficient Active Learning Pipeline for Legal Text Classification | 2022 |
| 3 | (EIN-DOR et al., 2020) | Active Learning for BERT: An Empirical Study | 2020 |
| 4 | (SAHAN; SMIDL; MARIK, 2021) | Active Learning for Text Classification and Fake News Detection | 2022 |
| 5 | (JACOBS et al., 2021) | Active learning for reducing labeling effort in text classification tasks | 2021 |
| 6 | (SCHRöDER; NIEKLER; POTTHAST, 2022) | Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers | 2022 |
| 7 | (CORVINO et al., 2022) | On the use of query by committee for human-in-the-loop named entity recognition | 2022 |
| 8 | (LEMMENS; DAELEMANS, 2023) | Combining Active Learning and Task Adaptation with BERT for Cost-Effective Annotation of Social Media Datasets | 2023 |

Tabela 2 – Principais características dos trabalhos relacionados

| ID | Qtde. Dados | Conjunto Inicial | Batch | Modelo | Query |
|---|---|---|---|---|---|
| 1 | 6.425 | 1000 | 50 | SVM, RF, KNN | LC, QBC, Batch-LC, EG, MD |
| 2 | 44.294 | 10 | 10 | BERT, RoBERTa | Random, LC, Dropout, DAL |
| 3 | 5.556 - 21.000 | 100 | 50 | BERT | LC, Dropout, EG, DAL |
| 4 | 1000 | 10 | 200 | Fast Text, RoBERTa | Dropout, SGLD, Softmax |
| 5 | 11800 | 590 | 11, 118, 590 | BERT | Razão de Variação, Entropia Preditiva e BALD |
| 6 | 6000 | 25 | 25 | BERT, RoBERTa | LC, PE, BT |
| 7 | 254 - 1315 | - | 125 - 200 | BERT, RoBERTa, BERTimbau | QBC |
| 8 | 8404 - 43,414 | 10% | 500 | BERT | Razão de Variação, Entropia Preditiva e BALD |

Tabela 3 – Comparação das Estratégias de Query e suas Melhorias em Desempenho

| ID | Melhor Estratégia de Query | Melhoria em Desempenho |
|---|---|---|
| 1 | Monte Carlo Dropout e LC | 8% de F1 |
| 2 | DAL | - |
| 3 | LC | 2% de F1 |
| 4 | SGLD | 8% de AUC |
| 5 | BALD | - |
| 6 | LC e PE | 4% de Acurácia |
| 7 | QBC | 10% de F1 |
| 8 | AL | 2% de F1 |

# 3 MATERIAIS E MÉTODOS

Este Capítulo apresenta os materiais usados neste trabalho, com uma breve descrição do *corpus* (seção 3.1.1) e do modelo automático previamente treinado para rotulação do sintoma "Tristeza/humor depressivo" (seção 3.1.2). Em seguida, a seção 3.2 descreve os passos realizados para permitir a investigação do uso do AL para rotulação desse sintoma de depressão.

## 3.1 Materiais

### 3.1.1 O *corpus* SetembroBR

Para realização desse estudo, foi utilizado o *corpus* SetembroBR (SANTOS; OLIVEIRA; PARABONI, 2023), disponibilizado pelo Prof. Dr. Ivandré Paraboni (EACH/USP). O conjunto de dados contém todos os *tweets* públicos escritos por 3.900 usuários únicos (ou seja, excluindo mensagens escritas por outros usuários, conhecidas como *retweets*) que auto relataram um diagnóstico ou tratamento para um transtorno mental, mas que foram postados antes de um diagnóstico ou tratamento auto-relatado para depressão e/ou ansiedade. O SetembroBR foi gerado com o objetivo fornecer dados de treinamento e teste para modelos de aprendizado de máquina supervisionados (*dataset*). Nesse contexto, os *tweets* dos usuários depressivos não possuem rotulação de sintomas, de maneira que cada usuário pode, ou não, ter realizado postagens na mídia social com conteúdo que caracterizaria-se como um sintoma de depressão.

A priori, é importante mencionar que o SetembroBR não possui nenhum tipo de rotulação quanto a sintomas depressivos, sendo separado somente entre indivíduos auto-diagnosticados como depressivos e um grupo controle. Como o intuito da tarefa selecionada para esta pesquisa é a realização de uma classificação binária de sintomas depressivos em *tweets* em português do Brasil, optou-se por trabalhar apenas com o conjunto de dados referentes aos usuários auto declarados depressivos que totalizam 2.427.499 postagens.

Por se tratar de dados de conteúdo sensível (saúde mental) e em respeito às diretrizes de uso do SetembroBR, este documento não reproduz exemplos de *tweets* presentes no *corpus*.

### 3.1.2 O modelo inicial de rotulação do sintoma "Tristeza/humor depressivo"

Para permitir o uso do aprendizado ativo, faz-se necessária a criação de rótulos iniciais que servirão como a classe positiva no treinamento dos modelos, ou seja, *tweets* que indiquem a presença de um sintoma depressivo. Seguindo os conceitos do PHQ-9 e considerando-se a relevância do sintoma "Tristeza/humor depressivo" para o diagnóstico da depressão, este foi selecionado para a realização dos experimentos.

O sintoma de "Tristeza/humor depressivo" foi definido pelo Comitê de Especialistas[1]

A partir dessa definição, os especialistas do Comitê do Amive anotaram um *corpus* de postagens anônimas da comunidade universitária, coletadas a partir de páginas públicas da rede social Facebook. Os textos foram anotados de acordo com diretrizes estabelecidas pelo Comitê a partir de análise de amostras das postagens coletadas, e com base na experiência clínica dos especialistas em saúde mental, incluindo trabalho voltado ao público universitário. A partir de 278 postagens rotuladas com o sintoma "Tristeza/humor depressivo" foi realizado um *fine-tuning* do modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), denominado neste trabalho como BERTimbau-Amive, que obteve uma acurácia de aproximadamente 75% em 89 instâncias de teste.

Por fim, vale mencionar que, em um cenário ideal, seria interessante anotar todo o *corpus* quanto a presença desse sintoma, podendo realizar uma análise completa de como as técnicas de aprendizado ativo impactam na performance do classificador, comparando o resultado do treinamento utilizando o conjunto de dados completo e uma amostragem de tamanho inferior com as postagens selecionadas pelos algoritmos de seleção. Todavia, o volume de dados (mais de 2 milhões de *tweets*) impossibilita tal perspectiva em decorrência da limitação de recursos e tempo disponíveis para realização dessa tarefa. Ainda assim, a criação dos modelos preditivos demanda a presença de rótulos positivos e negativos nas postagens.

### 3.1.3 *Corpus* filtrado pelo GPT-3.5

Para aumentar a assertividade dos modelos usados no comitê de classificadores para o *Active Learning*, uma estratégia adotada neste trabalho foi utilizar os indícios de relevância de *tweets* para classificação de saúde mental conforme (SANTOS; PARABONI, 2024). Neste trabalho, os autores utilizaram o GPT-3.5 para avaliar a relevância de *tweets* para a previsão de saúde mental.

Um *prompt* baseado na descrição clínica de depressão foi definido pelos autores do estudo e cada *tweet* do *corpus* SetembroBR foi categorizado em termos de relevância para a saúde mental. Com base nessa categorização, os *tweets* foram classificados como de alta, média ou baixa relevância. Essa abordagem focou na semântica, como sintomas e sinais clínicos de depressão, mas também considerou outros indicadores linguísticos. O método combinou o prompting do GPT 3.5, como mostrado na Figura 2, com a classificação de texto *bag-of-words* para resultados rápidos e computacionalmente baratos.

No total, considerando os *tweets* categorizados como possuindo alta relevância para saúde mental, totaliza-se 39 mil postagens no grupo de usuários diagnosticados como depressivos

[1] Composto por especialistas em psiquiatria, psicologia, terapia ocupacional e PLN. do projeto Amive[2] como: "Caracterizado pela tristeza, que pode durar mais do que o habitual. Sentir-se para baixo, desanimado, sem vontade. Negatividade, pessimismo. Melancolia. É importante que a gravidade e/ou a duração do sintoma de tristeza/humor depressivo sejam considerados na anotação."

Figura 2 – Prompt de instrução para o GPT 3.5

Considering tweet X below, which of the following three options would be most likely?

**Option 1**: Tweet X has strong indications that the individual who wrote it may be suffering from some type of depression or anxiety disorder:

1.a - This may occur because the tweet explicitly mentions intense feelings of depression, despair, intense anxiety, or related symptoms.

1.b - The tweet strongly and explicitly suggests that the user is experiencing high levels of depression or anxiety, even if it is not explicitly stated.

**Option 2**: Tweet X has moderate indications that the individual who wrote it may be suffering from some type of depression or anxiety disorder:

2.a - This may be because the tweet mentions feelings of depression, anxiety, stress, or related symptoms, but the indicators are not as strong as in the `high' category.

2.b - The tweet indirectly suggests that the user is experiencing low levels of depression or anxiety based on the content, even if it is not explicitly stated.

**Option 3**: Tweet X has little or no indication that the individual who wrote it may be suffering from any type of depression or anxiety disorder:

3.a - Messages that have no relation to the topic.

3.b - Messages that use language in a colloquial manner without indicating that it has a medical basis.

Return only the option value (1, 2 or 3).

Tweet X is {*tweet text here*}

Fonte: (SANTOS; PARABONI, 2024)

(total de 2.427.499 postagens).

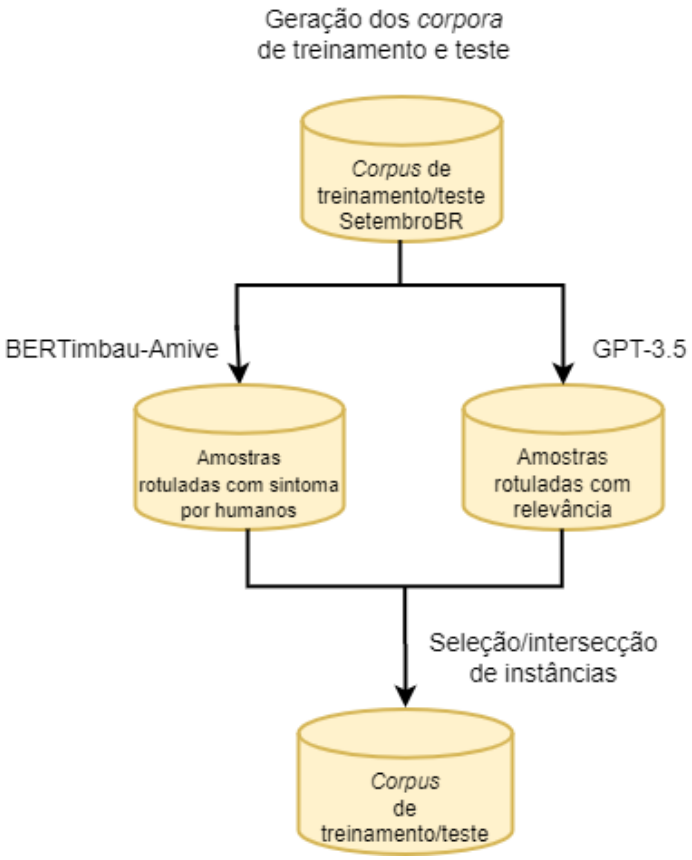
## 3.2 Aprendizado Ativo (*Active Learning*)

A Figura 3 explicita a primeira etapa do desenvolvimento: a construção do *corpus*. Nesta etapa, os materiais listados na seção anterior – *corpus* SetembroBR, modelo BERTimbau-Amive e *corpus* filtrado pelo GPT-3.5 – são usados para gerar os *corpora* de treinamento e teste.

A primeira etapa do pipeline de *Active Learning* consiste na rotulação automática dos *tweets* em duas classes: positiva, em que há presença do sintoma "Tristeza/humor depressivo", e negativa, em que não há presença deste sintoma. Como mencionado anteriormente, para a realização dessa tarefa, adotou-se o modelo BERTimbau-Amive (seção 3.1.2). Esse modelo foi usado para criar rótulos positivos e negativos para todos os *tweets* presentes nos conjuntos de treinamento e teste do SetembroBR. Como resultado, obteve-se um total de 26.443 *tweets* rotulados positivamente quanto a presença da classe Tristeza/Humor depressivo (do total original de 2.427.449 postagens), sendo 20.907 no conjunto de treino e 5.536 de teste.

Em seguida, os *tweets* identificados com a presença do sintoma foram confrontados com aqueles de alta relevância para a saúde mental conforme Santos e Paraboni (2024) (seção 3.1.3). Para tanto, criou-se um fluxo de seleção de *tweets* para o conjunto de treinamento, ilustrado na Figura 4. Considerando os 26.443 *tweets* classificados pelo BERTimbau-Amive

Figura 3 – Pipeline para a geração dos *corpora* de treino e teste

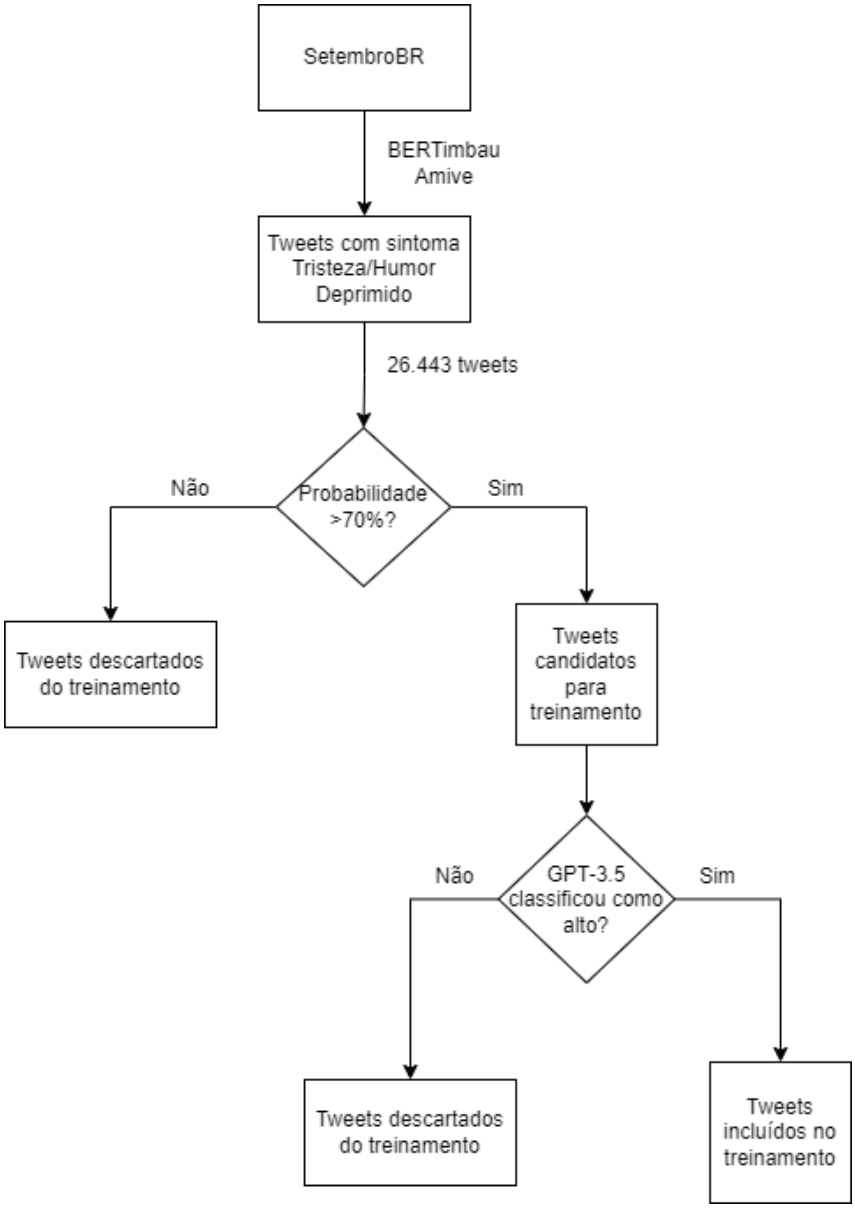



Fonte: Figura própria

com probabilidade retornada pelo modelo de pelo menos 70% de assertividade, realizou-se uma intersecção com os *tweets* rotulados pelo GPT-3.5. Os *tweets* rotulados como altamente relevantes para previsão da saúde mental foram comparados com os classificados com a presença do sintoma de "Tristeza/Humor depressivo", e os rotulados como baixa relevância foram comparados com os classificados com a ausência desse sintoma. Esse sintoma foi escolhido em decorrência de ser o principal sintoma depressivo de acordo com o American Psychiatric Association (2013) e por possuír a maior quantidade de instâncias rotuladas automáticamente no pelo BERTimbau-Amive no *corpus* SetembroBR.

Ao final, tem-se em um conjunto de treinamento com 6000 *tweets*, sendo 3000 da classe positiva e 3000 da classe negativa e um conjunto de teste com 2500 *tweets*, sendo 1250 na classe positiva e 1250 na classe negativa. Além disso, como garantia de confiabilidade no conjunto de teste, assegurou-se que nenhum usuário que aparecesse no conjunto de treino, estivesse, também, presente no conjunto de teste. Assim, é possível garantir que não há influência das maneiras distintas de escrita dos usuários nos resultados obtidos. O conjunto de treinamento foi usado para gerar os modelos que compuseram o comitê de classificadores, conforme descrito na seção 3.2.1.

Figura 4 – Script de construção do conjunto de treinamento




Fonte: Figura própria

### 3.2.1 Construção do Comitê de Classificadores

A partir do conjunto de treinamento gerado como descrito anteriormente, treinou-se 3 modelos de classificação indicados no estado da arte, conforme ilustrado na Figura 5:

1. O primeiro modelo, foi gerado a partir do *fine tuning* do BERTweet.BR[3] que é um modelo pré-treinado em um *corpus* de aproximadamente 1 milhão de *tweets* em português, seguindo a mesma estrutura do modelo de (NGUYEN; VU; NGUYEN, 2020).

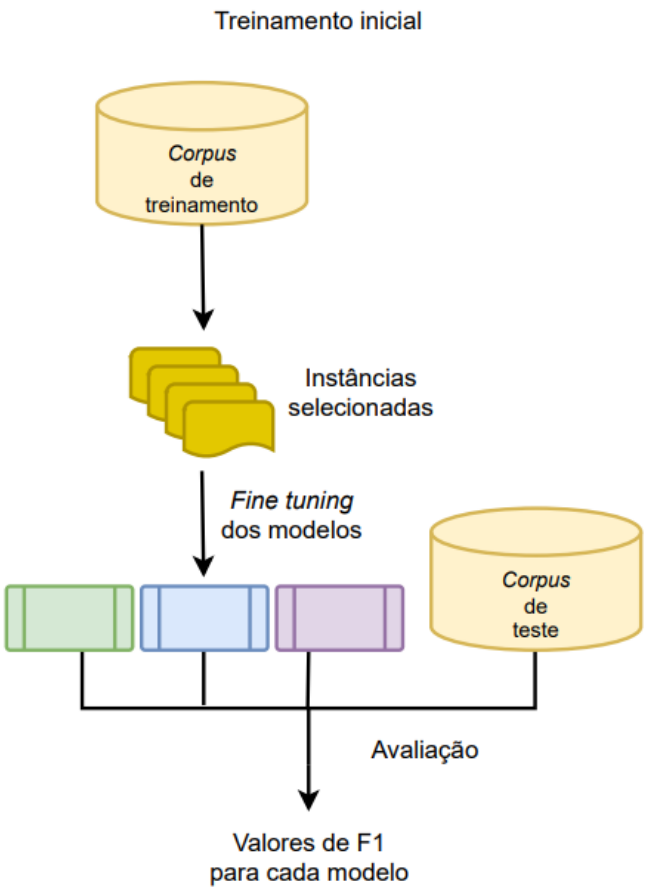
2. O segundo modelo é um *fine tuning* do BERTabaporu[4] (COSTA et al., 2023), um modelo

[3] Disponível em: https://huggingface.co/melll-uff/bertweetbr
[4] Disponível em: https://huggingface.co/pablocosta/bertabaporu-base-uncased

BERT no domínio do Twitter em português brasileiro, sendo construído a partir de uma coleção de 238 milhões de tweets escritos por mais de 100 mil usuários únicos do Twitter totalizando mais de 2,9 bilhões de *tokens*.

3. O terceiro modelo é um *fine tuning de* um modelo XLM-RoBERTa[5] (CONNEAU et al., 2020), uma versão multilíngue do modelo RoBERTa que conta com 8,4 bilhões de *tokens* em português.

Figura 5 – Pipeline do treinamento inicial




Fonte: Figura própria

Foi aplicado um preprocessamento dos dados para retirar os acentos agudos, circunflexos e as crases. Além disso, os dados brutos já estavam anonimizados e separados por usuários, de maneira que os usuários do conjunto de treino e de teste são distintos. Durante o treinamento, foram utilizados os *tokenizadores* originais disponibilizados pela HuggingFace. Além disso, foi utilizado o otimizador AdamW com o *learning rate* de 5e-56 e um agendador linear com aquecimento (*warmup*) para ajustar dinamicamente a taxa de aprendizagem ao longo das iterações de treinamento nos primeiros 10% de amostras. Em cada etapa, foram realizadas 10

[5] Disponível em: https://huggingface.co/docs/transformers/xlm-roberta

épocas de processamento através de um ambiente virtual do Google Colab[6] com um *batch size=8*, com um tempo de processamento de, aproximadamente, 40 minutos por época.

Esses três modelos formam o comitê dos modelos para as *queries* de seleção das amostras baseadas no aprendizado ativo, conforme descrito na próxima seção.

### 3.2.2 Query e Retreinamento

A Figura 6 ilustra a aplicação do aprendizado ativo, a qual foi repetida 3 vezes.

Considerando os trabalhos relacionados, as técnicas baseadas na seleção de amostras a partir da incerteza de modelos é a mais utilizada na literatura do aprendizado ativo. Como reiterado em (CORVINO et al., 2022), a utilização de um comitê de modelos pode ajudar na redução do viés nas seleções de amostras, uma vez que as *queries* são baseadas nas saídas de diversos classificadores, e, ao mesmo tempo, pode ser aplicado para calcular a incerteza dos modelos (ZHANG; STRUBELL; HOVY, 2022), similarmente a técnica de menor confiança.

Para seleção dos *tweets* mais relevantes para serem anotados por um humano, foram empregadas três estratégias de *query* baseadas em incertezas de modelos, considerando o desacordo entre modelos do comitê. Para suporte a essa tarefa foi utilizada a biblioteca "modAL" (DANKA; HORVATH, 2018), que consiste em um *framework* de *Active Learning* de código-fonte aberto, construído a partir do scikit-learn (PEDREGOSA et al., 2011), uma biblioteca comumente utilizada em Processamento de Linguagem Natural e Aprendizado de Máquina.

A primeira estratégia, denominada *Max Disagreement* (Discordância Maior) seleciona os exemplos onde há o máximo desacordo entre os modelos do comitê. Ao contrário dos métodos baseados em entropia, que consideram a distribuição geral de votos ou média de probabilidades, o "max disagreement" foca nos casos onde pelo menos dois modelos têm opiniões fortemente divergentes sobre a classificação de um exemplo. Exemplificando, se um dos classificadores do comitê discorda fortemente dos outros na categorização de um *tweet*, a amostragem de desacordo maior escolhe este dado para ser rotulada pelo oráculo.

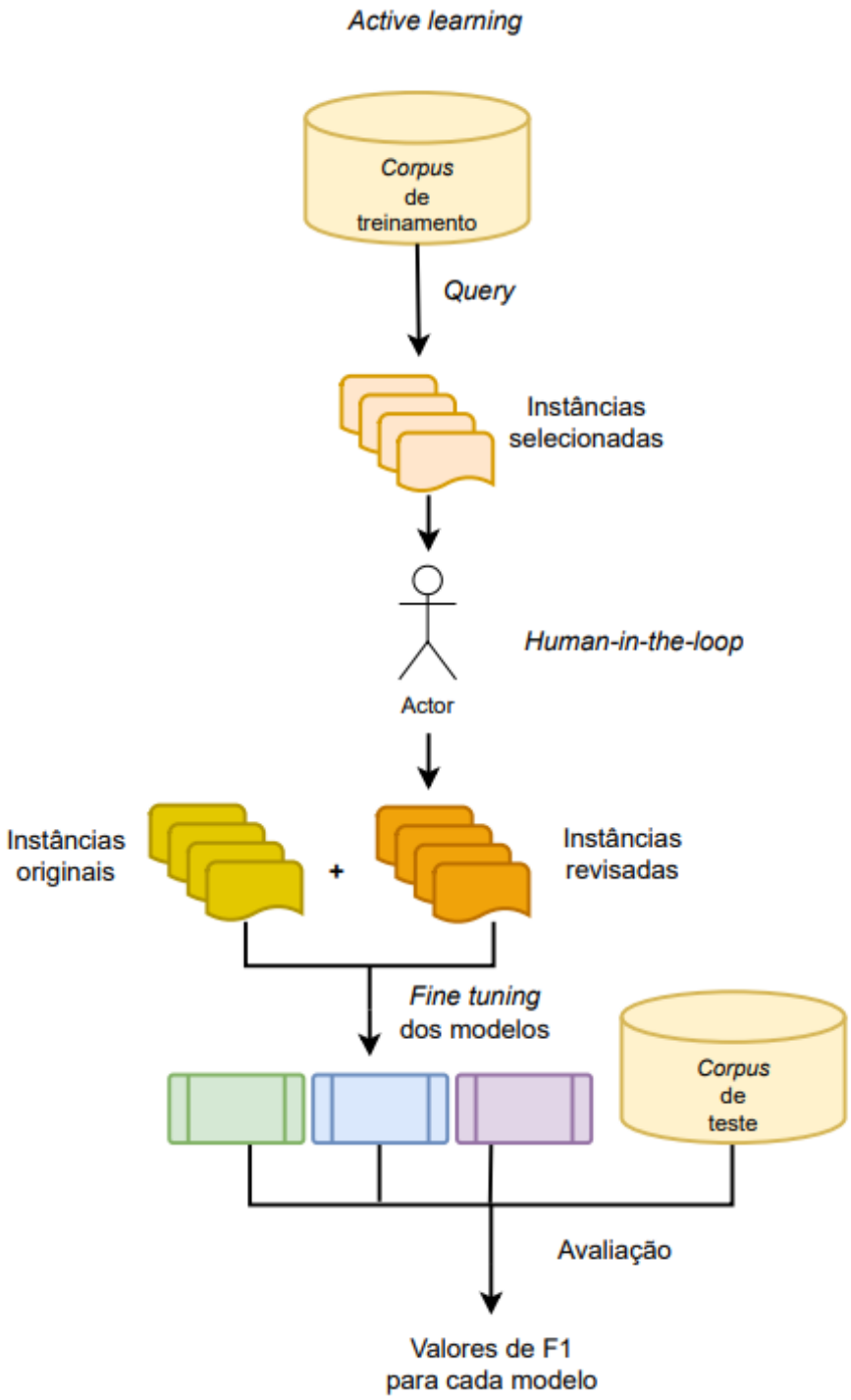
A divergência maior pode ser calculada como:

$$D_{KL}(P||Q) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right) \tag{3.1}$$

Onde $D_{KL}(P||Q)$ representa a divergência de Kullback-Leibler entre as distribuições de probabilidade P e $Q, P(i)$ é a probabilidade do evento $i$ na distribuição $P$, e $Q(i)$ é a probabilidade do evento $i$ na distribuição $Q$.

[6] Disponível em: https://colab.research.google.com/

Figura 6 – Pipeline do *Active Learning*




Fonte: Figura própria

A segunda estratégia avaliada é a *Vote Entropy* (Voto de Entropia). Essa estratégia seleciona a instância que apresenta a maior incerteza entre os classificadores do comitê, baseada na distribuição de votos, no qual cada modelo no comitê vota em uma classe para cada instância, e a entropia dos votos é usada para medir a incerteza. As instâncias com as maiores entropias de votos são escolhidas, pois são consideradas as mais informativas ou desafiadoras para o comitê, podendo proporcionar um aprendizado mais eficiente ao ser rotulada.

Por fim, a terceira estratégia de *query* investigada foi a *Consensus Entropy* (Consenso de Entropia), na qual, ao em vez de calcular a distribuição dos votos, a medida de desacordo da entropia de consenso calcula primeiro a média das probabilidades de classe de cada classificador. Isso é chamado de probabilidade de consenso. Em seguida, a entropia da probabilidade de consenso é calculada e a instância com a maior entropia de consenso é selecionada, uma vez que, matematicamente, a entropia é mais alta quando a distribuição de probabilidade é mais uniforme (ou seja, quando o modelo está mais incerto), e é mais baixa quando a distribuição é tendenciosa para uma classe específica (ou seja, quando o modelo está mais certo).

A entropia pode ser calculada como:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log(P(x_i)) \tag{3.2}$$

Onde H(X) é a entropia do conjunto de dados X, n é o número de classes ou estados possíveis, e $P(x_i)$ representa a probabilidade da ocorrência da classe ou estado $i$.

Dessa forma, o *Human-in-the-loop* foi aplicado para as três estratégias de *query*. Cada *query* seleciona as 250 amostras mais relevantes, de acordo com a sua respectiva estratégia, e tais amostras são anotadas pelo oráculo. Após a anotação, as amostras são incluídas no conjunto de treinamento e retiradas do conjunto de seleção de amostras mais importantes. Dessa forma, é possível treinar novamente os modelos e reavaliar as performances.

Além disso, a fim de comparar os impactos das técnicas do aprendizado ativo, foram treinados modelos atráves da técnica de amostragem aleatória, na qual seleciona-se 250 instâncias aleatórias do *corpus* de *query*, adicionando-as no conjunto de treinamento. Dessa maneira, é possível quantificar o ganho de performance do QBC.

Ao decorrer das rodadas de *query*, preconiza-se que a inclusão de novas amostras pode refletir no desbalanceamento do *dataset*. Como a classe dominante representa aproximadamente 0,1% do conjunto de dados, considerando a rotulação identificada pelo classificador BERTimbau-Amive, é factível afirmar que existe chance de que a maioria das postagens selecionadas seja representante da classe majoritária. A fim de amenizar o impacto desse fenômeno, incluiu-se um peso entre as classes na função de perda (*Cross-Entropy*). Dessa forma, possibilita-se a compensação dessa disparidade, atribuindo maior importância ao erro da classe minoritária.

### 3.2.3 Anotação de Amostras

A anotação da amostras selecionadas pelas *queries*, no processo *Human-in-the-loop* foi realizada por um especialista humano que também fez parte do Comitê de Especialistas do projeto Amive, totalizando 250 instâncias em cada rodada de query. A tarefa de anotação segue a definição do sintoma "Tristeza/Humor depressivo" conforme definido pelo Comitê de Especialistas do projeto Amive e apresentada na seção 3.1.2.

Além de seguir a definição do sintoma, o oráculo também adotou outras diretrizes no processo de anotação das instâncias selecionadas: *tweets* no quais o autor não era diretamente o sujeito da oração ou do sentimento expresso em questão não foram considerados como pertencentes ao sintoma; nem postagens nas quais relatava-se, diretamente, outros sintomas (Ex: insônia, ansiedade) ou dores específicas (Ex: dor de cabeça).

Para a anotação das instâncias selecionadas pelas *queries*, o oráculo utilizou a interface de anotação especificada na Figura 7 na qual a previsão do modelo era apresentada em uma das colunas, sendo 0 para ausência do sintoma e 1 para a presença do sintoma. Nesta interface, o oráculo deveria marcar se a previsão dada pelo modelo estava certa ou errada.


Figura 7 – Interface de anotação usada pelo oráculo

Todo esse processo de seleção de instâncias pelas *queries*, anotação pelo oráculo e retreinamento dos modelos foi repetido um total de 3 vezes, atingindo a condição de parada. Essa definição foi feita a partir do cálculo de recurso disponível para a anotação humana em detrimento do tempo restante para realização do trabalho. Em outros cenários, observa-se diversas condições de parada, podendo considerar, inclusive, a performance dos modelos. Todavia, por ser um estudo inicial com limitações, esse critério não foi considerado na condição de parada.

No total, foram anotadas 1334 amostras únicas, sendo que as *query* atuam no domínio da incerteza das predições dos modelos, ou seja, podem haver sobreposições de tal modo que duas estratégias distintas podem selecionar as mesmas instâncias na mesma rodada.

### 3.2.4 Métrica de avaliação

Como métrica de avaliação utilizou-se a média F1, amplamente utilizada nos trabalhos relacionados e especialmente útil em situações onde os *datasets* são desbalanceados, como é o caso do *corpus* SetembroBR (SANTOS; OLIVEIRA; PARABONI, 2023).

Dessa forma, a média F1 pode ser calculada como:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.3}$$

enquanto a precisão é calculada como:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{3.4}$$

e a revocação:

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{3.5}$$

onde $TP$ representa os verdadeiros positivos, $FP$ os falsos positivos, e $FN$ os falsos negativos.

Por se tratar de um *corpus* desbalanceado, foi utilizado um peso entre os valores da classe positiva e da classe negativa, calculando uma média ponderada. Esse cálculo é realizado automaticamente pelo parâmetro *average* = *weighted*, disponível no *Scikit-Learn* [7].

[7] Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1$_score.html$

# 4 RESULTADOS E DISCUSSÃO

Esta seção descreve os resultados dos experimentos realizados para a tarefa de verificar a usabilidade do aprendizado ativo na tarefa de classificação automática de sintomas depressivos em *tweets* escritos em português do Brasil.

## 4.1 Avaliação dos modelos iniciais

A Tabela 4 apresenta os valores iniciais dos treinamentos do comitê, anteriores à adição das amostras selecionadas pelas estratégias de *query*, ou seja, os valores de F1 calculados para os modelos gerados como ilustrado na Figura 5.

| **Modelo** | **F1** |
|---|---|
| BERTweet.BR | 52,97% |
| BERTabaporu | 49,08% |
| XLM-RoBERTa | 43,05% |

Tabela 4 – Resultados Iniciais dos Modelos do Comitê

## 4.2 Avaliação do Active Learning

As Tabelas 5, 6 e 7 apresentam, respectivamente, os resultados das rodadas 1, 2 e 3 do aprendizado ativo, conforme ilustrado na Figura 6. Dessa forma, é possível observar os resultados da média F1 mediante a configuração dos modelos do comitê com as estratégias de *query* nas 3 rodadas realizadas.

Em cada rodada, são adicionadas 250 instâncias no treinamento de cada modelo, de acordo com a seleção realizada pela estratégia utilizada. As tabelas também trazem os valores de F1 para a abordagem de amostragem aleatória na qual são adicionadas 250 instâncias através da função *sample()* da biblioteca Pandas[1] que retorna, aleatoriamente, essa mesma quantidade de dados.

### 4.2.1 Análise dos resultados

Analisando os resultados, é possível verificar que o aprendizado ativo apresenta ganhos de performance na maioria dos experimentos realizados. Entre os melhores resultados, o modelo BERTweet.BR atingiu um F1 médio de 86,14% na estratégia de*query* de *Consensus Entropy*

[1] Disponível em: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html

| **Modelo** | **Estratégia de Query** | **F1** |
|---|---|---|
| BERTabaporu | Max Disagreement | 59,45% |
| XLM-RoBERTa | Max Disagreement | 55,24% |
| BERTweet.BR | Max Disagreement | 66,39% |
| BERTabaporu | Vote Entropy | 54,55% |
| XLM-RoBERTa | Vote Entropy | 50,24% |
| BERTweet.BR | Vote Entropy | 60,87% |
| BERTabaporu | Consensus Entropy | 67,07% |
| XLM-RoBERTa | Consensus Entropy | 61,83% |
| **BERTweet.BR** | **Consensus Entropy** | **75,26%** |
| BERTabaporu | Amostragem Aleatória | 54,30% |
| XLM-RoBERTa | Amostragem Aleatória | 53,62% |
| BERTweet.BR | Amostragem Aleatória | 65,26% |

Tabela 5 – Resultados dos Experimentos na rodada 1

| **Modelo** | **Estratégia de Query** | **F1** |
|---|---|---|
| BERTabaporu | Max Disagreement | 74,65% |
| XLM-RoBERTa | Max Disagreement | 69,23% |
| BERTweet.BR | Max Disagreement | 73,45% |
| BERTabaporu | Vote Entropy | 64,65% |
| XLM-RoBERTa | Vote Entropy | 62,23% |
| BERTweet.BR | Vote Entropy | 68,74% |
| BERTabaporu | Consensus Entropy | 78,60% |
| XLM-RoBERTa | Consensus Entropy | 73,69% |
| **BERTweet.BR** | **Consensus Entropy** | **83,14%** |
| BERTabaporu | Amostragem Aleatória | 71,10% |
| XLM-RoBERTa | Amostragem Aleatória | 67,41% |
| BERTweet.BR | Amostragem Aleatória | 71,45% |

Tabela 6 – Resultados dos Experimentos na rodada 2

| **Modelo** | **Estratégia de Query** | **F1** |
|---|---|---|
| BERTabaporu | Max Disagreement | 77,45% |
| XLM-RoBERTa | Max Disagreement | 72,23% |
| BERTweet.BR | Max Disagreement | 77,23% |
| BERTabaporu | Vote Entropy | 68,34% |
| XLM-RoBERTa | Vote Entropy | 66,09% |
| BERTweet.BR | Vote Entropy | 71,42% |
| BERTabaporu | Consensus Entropy | 84,39% |
| XLM-RoBERTa | Consensus Entropy | 75,11% |
| **BERTweet.BR** | **Consensus Entropy** | **86,84%** |
| BERTabaporu | Amostragem Aleatória | 74,91% |
| XLM-RoBERTa | Amostragem Aleatória | 69,15% |
| BERTweet.BR | Amostragem Aleatória | 76,98% |

Tabela 7 – Resultados dos Experimentos na rodada 3

no round 3, superando em quase 10 pontos percentuais o resultado obtido pela amostragem aleatória na mesma rodada.

Além disso, é possível analisar que os modelos BERTweet.BR e BERTabaporu apresentaram resultados semelhantes. Uma possível explicação seria que ambos os modelos são oriundos do mesmo domínio, sendo treinandos em um conjunto de dados de *tweets*. Paralelamente, o modelo XML-RoBERTa, que pertence a outro domínio de dados, obteve resultados inferiores aos demais, apesar de também superarem a amostragem aleatória.

Dentre as técnicas de *query* aplicadas, a *Vote Entropy* foi a que obteve os piores resultados. Como essa metodologia baseia-se na rotulação final dos modelos, ao contrário das demais que consideram, também, a probabilidade daquela instância pertencer à classe prevista, a Vote Entropy pode priorizar amostras onde há um conflito moderado entre os modelos (por exemplo, uma distribuição de votos 2 contra 1), em vez de focar em amostras onde todos os modelos estão altamente incertos. Esses conflitos moderados nem sempre são os mais informativos para melhorar o desempenho geral do sistema.

Por outro lado, a *Consensus Entropy* foi a que obteve os melhores valores de F1. Em *datasets* com desequilíbrio de classes, como o SetembroBR, é crucial que as amostras selecionadas para o treinamento sejam representativas e desafiadoras. A Consensus Entropy, ao exigir um consenso sobre a incerteza, pode ser mais eficaz em escolher amostras que realmente ajudem a balancear o aprendizado entre as classes.

Por fim, a *query* de *Max Disagreement* também obteve bons resultados, alcançando um valor de 77% de F1 na configuração *BERTweet.BR + Max Disagreement*. Amostras que geram discordância máxima entre modelos frequentemente revelam áreas onde pelo menos um modelo está significativamente errado ou mal calibrado. Isso pode ser valioso para rapidamente identificar e corrigir deficiências nos modelos, podendo ser particularmente útil em *datasets* complexos e variados, como o cenário desse estudo, onde entender e reconciliar diferentes interpretações dos modelos pode ser chave para alcançar uma compreensão mais profunda e um desempenho melhorado.

## 4.3 Avaliação da aplicabilidade em outro *corpus*

Para investigar a aplicabilidade do melhor modelo derivado do Active Learning em um *corpus* diferente, os classificadores treinados usando a estratégia de Active Learning foram utilizados para predizer as classes das instâncias do *corpus* utilizado no treinamento do modelo BERTimbau-Amive, descrito na seção 3.1.2. Vale lembrar que este *corpus* possui instâncias derivadas de postagens do Facebook.

Foram coletadas as 63 instâncias do conjunto de teste rotuladas com o sintoma de "Tristeza / Humor depressivo" e mais 63 instâncias aleatórias que não possuíam esse sintoma.

O intuito deste experimento foi avaliar uma transferência de conhecimento de domínio nos melhores modelos treinados no *corpus* do SetembroBR, bem como utilizar amostras anotadas por especialistas humanos em um conjunto de teste.

A Tabela 8 explicita os valores de F1 alcançados pelos modelos.

| **Modelo** | **Estratégia de Query** | **Rodada** | **F1** |
|---|---|---|---|
| **BERTweet.BR** | **Consensus Entropy** | **3** | **85,94%** |
| BERTabaporu | Consensus Entropy | 3 | 85,74% |
| XLM-RoBERTa | Consensus Entropy | 3 | 60,11% |

Tabela 8 – Resultados dos Experimentos com os melhores modelos gerados usando Active Learning no *corpus* Amive-Facebook

Dessa forma, verifica-se que os modelos BERTweet.BR e BERTabaporu, com a *query* de *Consensus Entropy* no último round dos experimentos atingiram valores de F1 próximos a 86%, indicando uma boa transferência de conhecimento. Todavia, o modelo XML-RoBERTa com a mesma configuração não atingiu o mesmo resultado.

## 4.4 Limitações desta pesquisa

Algumas limitações na aplicação do aprendizado ativo foram enfrentadas no decorrer do desenvolvimento deste trabalho.

Primeiramente, não foi possível recorrer a um comitê de anotadores diverso para rotulação das instâncias selecionadas pelas estratégias de *query*. Dessa forma, cabe pontuar um possível viés de anotação humana, uma vez que não há como medir o nível de discordância entre os anotadores.

Concomitantemente, devido ao volume de dados do SetembroBR, não havia recursos humanos disponíveis para rotular um conjunto de treinamento e de teste inicial. Para enfrentar esse problema, recorreu-se ao modelo previamente desenvolvido do BERTimbau-Amive, que, apesar de ter alcançado resultados favoráveis no *corpus* que foi treinado (aproximadamente 75% de F1), levou a valores de F1 dos modelos iniciais bem abaixo: 52,97% (BERTweet.BR), 49,08% (BERTabaporu) e 43,05% (XLM-RoBERTa).

Como o objetivo desse trabalho era avaliar o impacto das aplicações de aprendizado ativo na construção de classificadores textuais, buscou-se aplicar um cenário que controlasse esses vieses, como o uso da filtragem por relevância das instâncias (GPT-3.5), a testagem dos modelos em outro *corpus*, a seleção de amostras de testes de usuários diferentes e a construção de um comitê com modelos refinados de domínios distintos do estado da arte.

# 5 CONCLUSÕES

Neste Trabalho de Conclusão de Curso foi realizada uma investigação inicial das estratégias de aprendizado ativo na classificação de um sintoma de depressão (tristeza ou humor depressivo) em postagens do Twitter, no *corpus* SetembroBR.

Como reiterado na literatura, a tarefa de anotação de um conjunto de dados demanda uma grande alocação de recursos financeiros e de tempo. No domínio da saúde mental, esse entrave fica ainda mais evidente, uma vez que identificar sintomas em postagens de mídias sociais é uma tarefa altamente complexa, em decorrência da necessidade de especialistas na área e da escassez de dados disponíveis.

Dessa forma, considerando tal *corpus* que possui um volume de dados de 2.427.499 postagens, a utilização de técnicas alternativas à anotação manual demonstra-se altamente relevante. Nesse contexto, foi possível construir um comitê de modelos a partir da criação de um *dataset* de treinamento, unindo esforços entre as rotulações previstas pelo BERTimbau-Amive e o *prompt* do GPT-3.5 (SANTOS; PARABONI, 2024).

A fim de mitigar o viés instaurado pela rotulação automática através da transferência de conhecimento, optou-se pela escolha de treinar três modelos linguísticos de *Deep Learning* distintos, medindo, assim, a incerteza e a discordância conjunta do comitê, buscando deixar mais robusta e acurada a seleção de instâncias a serem anotadas pelo oráculo (Human-in-the-loop).

Dentre as estratégias baseadas na incerteza, que são as mais comumente utilizadas, destacou-se a *Consensus Entropy*, que seleciona amostras em que todos os modelos do comitê concordam que estão incertos, com o objetivo de identificar e aprender com as amostras que são inerentemente ambíguas ou difíceis para todo o sistema. Assim, diante de um *corpus* altamente diverso, no qual estima-se que o número de *tweets* com a presença do sintoma depressivo "Tristeza / Humor depressivo" represente menos de 10% do conjunto total, reconhecer quais amostras são mais representativas é uma tarefa árdua, porém crucial.

Contudo, mesmo diante de tal cenário, foi possível atingir métricas favoráveis de F1, alcançando valores próximos a 86% com o modelo BERTweet.BR, 84% com o modelo BERTabaporu e 75% com o XLM-RoBERTa, utilizando a *query* de consenso de entropia. Vale destacar também que, com exceção da última estratégia de *query*, todas as outras conseguiram atingir valores acima da amostragem aleatória em cerca de 10 pontos percentuais.

Além disso, foi avaliado, também, o desempenho desses mesmos modelos no *corpus* de postagens anônimas coletadas do Facebook no projeto Amive, no qual os mesmos modelos atingiram 85%, 85% e 60%, respectivamente.

Em síntese, foi possível realizar uma análise inicial do aprendizado ativo na classificação

de postagens de mídias sociais no contexto de saúde mental, através da construção de um comitê de modelos, atingindo o objetivo deste trabalho e indicando a viabilidade da tarefa.

Como continuação desse trabalho, possíveis caminhos seriam explorar outras estratégias de query, como as baseadas em valores de gradiente, densidade, ou formas híbridas, combinando-as com as baseadas em incerteza. Além disso, realizar uma análise de saturação, avaliando a relação entre a quantidade de rodadas do *Human-in-the-loop* e o ganho de performance dos modelos e verificar, também, a aplicabilidade da técnica no contexto multiclasse, utilizando outros sintomas depressivos e reduzindo o custo de anotação.

# REFERÊNCIAS

American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed.. ed. Washington, DC: Autor, 2013. Citado na página 36.

American Psychiatric Association et al. DSM-5: Manual diagnóstico e estatístico de transtornos mentais. [S.l.]: Artmed Editora, 2014. Citado na página 17.

CER, D. et al. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, 2017. Disponível em: <http://dx.doi.org/10.18653/v1/S17-2001>. Citado na página 25.

CHOUDHURY, M. D.; COUNTS, S.; HORVITZ, E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference. New York, NY, USA: Association for Computing Machinery, 2013. (WebSci '13), p. 47–56. ISBN 9781450318891. Disponível em: <https://doi.org/10.1145/2464464.2464480>. Citado na página 18.

CONNEAU, A. et al. Unsupervised Cross-lingual Representation Learning at Scale. 2020. Citado na página 38.

CORVINO, G. et al. On the use of query by committee for human-in-the-loop named entity recognition. In: Anais do X Symposium on Knowledge Discovery, Mining and Learning. Porto Alegre, RS, Brasil: SBC, 2022. p. 106–113. ISSN 2763-8944. Disponível em: <https://sol.sbc.org.br/index.php/kdmile/article/view/24975>. Citado 3 vezes nas páginas 29, 30 e 39.

COSTA, P. B. et al. BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. In: MITKOV, R.; ANGELOVA, G. (Ed.). Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023. p. 217–223. Disponível em: <https://aclanthology.org/2023.ranlp-1.24>. Citado na página 37.

DANKA, T.; HORVATH, P. modAL: A modular active learning framework for Python. 2018. Available on arXiv at <https://arxiv.org/abs/1805.00979>. Disponível em: <https://github.com/modAL-python/modAL>. Citado na página 39.

DEMSZKY, D. et al. GoEmotions: A dataset of fine-grained emotions. In: JURAFSKY, D. et al. (Ed.). Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. p. 4040–4054. Disponível em: <https://aclanthology.org/2020.acl-main.372>. Citado na página 30.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. Citado na página 23.

EIN-DOR, L. et al. Active Learning for BERT: An Empirical Study. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. p. 7949–7962. Disponível em: <https://aclanthology.org/2020.emnlp-main.638>. Citado 3 vezes nas páginas 23, 24 e 30.

FARINNEYA, P. et al. Active learning for rumor identification on social media. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 4556–4565. Disponível em: <https://aclanthology.org/2021.findings-emnlp.387>. Citado 3 vezes nas páginas 25, 26 e 30.

FATIMA, I. et al. Analysis of user-generated content from online social communities to characterise and predict depression degree. J. Inf. Sci., Sage Publications, Inc., USA, v. 44, n. 5, p. 683–695, oct 2018. ISSN 0165-5515. Disponível em: <https://doi.org/10.1177/0165551517740835>. Citado na página 18.

GAL, Y.; GHAHRAMANI, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2016. Citado na página 23.

GISSIN, D.; SHALEV-SHWARTZ, S. Discriminative Active Learning. 2019. Citado 2 vezes nas páginas 18 e 24.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2004. (KDD '04), p. 168–177. ISBN 1581138881. Disponível em: <https://doi.org/10.1145/1014052.1014073>. Citado na página 28.

HUANG, J. et al. Active Learning for Speech Recognition: the Power of Gradients. 2016. Citado na página 24.

JACOBS, P. F. et al. Active learning for reducing labeling effort in text classification tasks. 2021. Citado 2 vezes nas páginas 27 e 30.

JAMIL, Z. et al. Monitoring tweets for depression to detect at-risk users. In: HOLLINGSHEAD, K.; IRELAND, M. E.; LOVEYS, K. (Ed.). Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality. Vancouver, BC: Association for Computational Linguistics, 2017. p. 32–40. Disponível em: <https://aclanthology.org/W17-3104>. Citado na página 18.

KOREEDA, Y.; MANNING, C. D. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. 2021. Citado na página 25.

KROENKE, K.; SPITZER, R. L.; WILLIAMS, J. B. The phq-9: validity of a brief depression severity measure. Journal of general internal medicine, Wiley Online Library, v. 16, n. 9, p. 606–613, 2001. Citado na página 17.

LEMMENS, J.; DAELEMANS, W. Combining active learning and task adaptation with BERT for cost-effective annotation of social media datasets. In: BARNES, J.; CLERCQ, O. D.; KLINGER, R. (Ed.). Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Toronto, Canada: Association for Computational Linguistics, 2023. p. 237–250. Disponível em: <https://aclanthology.org/2023.wassa-1.22>. Citado 2 vezes nas páginas 29 e 30.

LEWIS, D. D.; GALE, W. A. A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 1994. (SIGIR '94), p. 3–12. ISBN 038719889X. Citado na página 23.

LI, X.; ROTH, D. Learning question classifiers. In: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1. USA: Association for Computational Linguistics, 2002. (COLING '02), p. 1–7. Disponível em: <https://doi.org/10.3115/1072228.1072378>. Citado na página 28.

LJUBEšIć, N.; FIšER, D.; ERJAVEC, T. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. 2019. Citado na página 30.

MAMOOLER, S. et al. An Efficient Active Learning Pipeline for Legal Text Classification. 2022. Citado 2 vezes nas páginas 24 e 30.

MANN, P.; PAES, A.; MATSUSHIMA, E. H. See and read: Detecting depression symptoms in higher education students using multimodal social media data. In: Proceedings of the International AAAI Conference on Web and Social Media. [S.l.: s.n.], 2020. v. 14, p. 440–451. Citado na página 18.

MENDES, A. R.; CASELI, H. de M. Identifying fine-grained depression signs in social media posts. Submitted to LREC 2024. submitted. Citado na página 20.

MILLER, B.; LINDER, F.; MEBANE, W. R. Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches. Political Analysis, Cambridge University Press, v. 28, n. 4, p. 532–551, 2020. Citado 2 vezes nas páginas 18 e 19.

MOWERY, D. L. et al. Towards automatically classifying depressive symptoms from Twitter data for population health. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES). Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 182–191. Disponível em: <https://aclanthology.org/W16-4320>. Citado na página 18.

NGUYEN, D. Q.; VU, T.; NGUYEN, A. T. BERTweet: A pre-trained language model for English Tweets. 2020. Citado na página 37.

PANG, B.; LEE, L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 2005. (ACL '05), p. 115–124. Disponível em: <https://doi.org/10.3115/1219840.1219855>. Citado na página 28.

PANG, B.; LEE, L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 2005. (ACL '05), p. 115–124. Disponível em: <https://doi.org/10.3115/1219840.1219855>. Citado na página 28.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Citado na página 39.

PONTIKI, M. et al. SemEval-2015 task 12: Aspect based sentiment analysis. In: NAKOV, P. et al. (Ed.). Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, 2015. p. 486–495. Disponível em: <https://aclanthology.org/S15-2082>. Citado na página 29.

QUDAR, M. M. A.; MAGO, V. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis. 2020. Citado na página 26.

REN, P. et al. A survey of deep active learning. ACM Comput. Surv., Association for Computing Machinery, New York, NY, USA, v. 54, n. 9, oct 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3472291>. Citado na página 20.

SAHAN, M.; SMIDL, V.; MARIK, R. Active learning for text classification and fake news detection. In: 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC). [S.l.: s.n.], 2021. p. 87–94. Citado 2 vezes nas páginas 26 e 30.

SANTOS, W.; FUNABASHI, A.; PARABONI, I. Searching brazilian twitter for signs of mental health issues. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020. p. 6111–6117. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.750>. Citado 2 vezes nas páginas 7 e 18.

SANTOS, W. R. dos; OLIVEIRA, R. L. de; PARABONI, I. SetembroBR: a social media corpus for depression and anxiety disorder prediction. Language Resources and Evaluation, 2023. Citado 3 vezes nas páginas 19, 33 e 42.

SANTOS, W. R. dos; PARABONI, I. Prompt-based mental health screening from social media text. 2024. Citado 4 vezes nas páginas 7, 34, 35 e 49.

SCHRöDER, C.; NIEKLER, A.; POTTHAST, M. Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers. 2022. Citado 2 vezes nas páginas 28 e 30.

SENER, O.; SAVARESE, S. Active Learning for Convolutional Neural Networks: A Core-Set Approach. 2018. Citado 2 vezes nas páginas 24 e 25.

SOCHER, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: YAROWSKY, D. et al. (Ed.). Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. p. 1631–1642. Disponível em: <https://aclanthology.org/D13-1170>. Citado na página 27.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.], 2020. Citado na página 34.

YADAV, S. et al. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 696–709. Disponível em: <https://aclanthology.org/2020.coling-main.61>. Citado na página 18.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level Convolutional Networks for Text Classification. 2016. Citado na página 28.

ZHANG, Z.; STRUBELL, E.; HOVY, E. A survey of active learning for natural language processing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 6166–6190. Disponível em: <https://aclanthology.org/2022.emnlp-main.414>. Citado 3 vezes nas páginas 18, 20 e 39.

ZIWEI, B. Y.; CHUA, H. N. An application for classifying depression in tweets. In: Proceedings of the 2nd International Conference on Computing and Big Data. New York, NY, USA: Association for Computing Machinery, 2019. (ICCBD 2019), p. 37–41. ISBN 9781450372909. Disponível em: <https://doi.org/10.1145/3366650.3366653>. Citado na página 18.