UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ANDRÉ MEDIOTE DE SOUSA

# SSSD: Explorando Modelos Pré-treinados e Busca Semântica para Detecção de Posicionamentos no Twitter

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Karin Becker

Porto Alegre
2023

**CIP — CATALOGAÇÃO NA PUBLICAÇÃO**

de Sousa, André Mediote

SSSD: Explorando Modelos Pré-treinados e Busca Semântica para Detecção de Posicionamentos no Twitter / André Mediote de Sousa. – Porto Alegre: PPGC da UFRGS, 2023.

83 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2023. Orientador: Karin Becker.

1. Modelos Pré-treinados. 2. Busca Semântica. 3. Detecção de Posicionamentos. 4. Few-Shot Learning. 5. Twitter. I. Becker, Karin. II. Título.

## AGRADECIMENTOS

Primeiramente, gostaria de expressar minha sincera gratidão à UFRGS pela oportunidade de realizar meu mestrado em uma das instituições mais prestigiadas do Brasil. É uma honra imensa ter um título emitido por esta renomada instituição constando em meu currículo.

Estendo minha profunda gratidão à minha orientadora, Dra. Karin Becker, que não apenas me incentivou a extrair o melhor de mim mesmo, mas também demonstrou uma paciência e compreensão excepcionais, guiando-me pelas melhores trajetórias durante o programa de mestrado para que eu pudesse alcançar meus objetivos. Minha jornada foi enriquecida e meu caminho foi iluminado graças à sua sabedoria e apoio constantes.

Gostaria também de agradecer aos meus colegas de pesquisa, Régis Ebeling, Jesus Yepes e Leonardo Andrade. A colaboração e o auxílio deles foram fundamentais para o meu progresso e sucesso. A jornada acadêmica foi mais gratificante e enriquecedora por ter tido a oportunidade de aprender e crescer ao lado de tais indivíduos talentosos e dedicados.

Sou eternamente grato aos meus pais, que foram os arquitetos da minha educação e os primeiros a me ensinar a batalhar pelos meus sonhos. Eles sempre me inspiraram com exemplos de trabalho duro, honestidade e resiliência, fundamentais para enfrentar os desafios da vida. A base que construíram para mim tem sido a minha força e guia.

Por fim, mas certamente não menos importante, agradeço do fundo do meu coração à minha companheira, Jociane Schardong. Seu incentivo, paciência, dedicação, atenção aos detalhes e amor foram luzes guiando-me nesta jornada desafiadora. Sem dúvida, sua presença e apoio foram pilares sem os quais eu não teria conseguido superar os obstáculos dessa etapa da minha vida. Seu amor não apenas me fortaleceu, mas também me deu a paz e a confiança necessárias para continuar perseguindo meus sonhos, mesmo nos momentos mais difíceis.

## RESUMO

Neste trabalho, é apresentado o SSSD (Semantic Search Stance Detection), um método inovador baseado no paradigma de Aprendizado de Máquina *Few-shot Learning* para Detecção de Posicionamentos (DP). Esta técnica emprega Modelos Pré-treinados (MPTs) para otimizar DP em *tweets* através da Busca Semântica. O SSSD tem a capacidade de interpretar o contexto e classificar o conteúdo dos *tweets* de maneira eficiente, requerendo apenas um conjunto pequeno de exemplos rotulados, o que contribui substancialmente para a redução do esforço manual de rotulagem e dos recursos necessários para o treinamento de modelos de DP. A estratégia proposta aprimora a precisão da DP ao filtrar conteúdos irrelevantes e focar nas postagens mais pertinentes. O SSSD destaca-se por ser pioneiro na integração de MPTs e Busca Semântica, facilitando a superação de desafios relacionados à escassez de dados rotulados e promovendo a melhoria da DP em mídias sociais. Em experimentos que tomaram como referência a competição SemEval-2016 Tarefa 6, o SSSD superou todos os *benchmarks* estabelecidos, evidenciando um potencial significativo na economia de recursos. Foi realizada ainda uma análise qualitativa para avaliar a eficácia do SSSD na detecção de posicionamentos relacionados à campanha de vacinação no Brasil durante a pandemia de COVID-19. Os resultados confirmam que o SSSD apresenta bons resultados mesmo com um volume limitado de dados rotulados, diferenciando-o positivamente em comparação com outras metodologias.



**Palavras-chave:** Modelos Pré-treinados. Busca Semântica. Detecção de Posicionamentos. Few-Shot Learning. Twitter.

**SSSD: Leveraging Pre-trained Models and Semantic Search for Stance Detection**


## ABSTRACT

In this work, SSSD (Semantic Search Stance Detection) is presented, an innovative method based on the Few-shot Learning paradigm for Stance Detection (SD). This technique employs Pre-trained Models (PTMs) to optimize SD in tweets through Semantic Search. SSSD is capable of interpreting context and efficiently classifying tweet content, requiring only a small set of labeled examples. This substantially reduces the manual labeling effort and resources necessary for training SD models. The proposed strategy enhances SD precision by filtering irrelevant content and focusing on the most pertinent posts. SSSD stands out as a pioneer in integrating PTMs and Semantic Search, facilitating the overcoming of challenges related to the scarcity of labeled data and enhancing SD in social media. In experiments referencing the SemEval-2016 Task 6 competition, SSSD surpassed all established benchmarks, showcasing significant potential in resource savings. A qualitative analysis was also conducted to evaluate the efficacy of SSSD in detecting stances related to the vaccination campaign in Brazil during the COVID-19 pandemic. The results confirm that SSSD achieves good results even with a limited volume of labeled data, distinguishing itself positively compared to other methodologies.



**Keywords:** Pre-trained Models, Semantic Search, Stance Detection, Few-Shot Learning, Twitter.

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ACC | Acurácia |
| AM | Aprendizado de Máquina |
| AP | Aprendizado Profundo |
| BERT | Bidirecional Encoder Representations from Transformers |
| BiLSTM | Bidirectional Long Short-Term Memory |
| BoW | Bag of Words |
| CGU | Conteúdo Gerado pelo Usuário |
| CNN | Convolutional Neural Networks |
| DP | Detecção de Posicionamentos |
| GloVe | Global Vectors for Word Representation |
| GPT | Generative Pre-trained Transformer |
| GRU | Gated Recurrent Units |
| KNN | K-Nearest Neighbors" |
| LSTM | Long Short-Term Memory |
| MPT | Modelos Pré-Treinados |
| NB | Naive Bayes |
| PLN | Processamento de Linguagem Natural |
| RF | Randon Forest |
| RL | Logistc Regression |
| RoBERTa | Robustly optimized BERT |
| SBERT | Sentence-BERT |
| SemEval | Semantic Evaluation |
| SSSD | Semantic Search Stance Detection |
| SVM | Suport Vector Machine |
| TF-IDF | Term Frequency-Inverse Document |

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Detecção de Posicionamento (DP) é a tarefa que determina automaticamente se o autor de um texto está a favor, contra ou não se manifesta em relação a um determinado alvo. Os alvos podem ser empresas, movimentos, pessoas ou ideias (MOHAMMAD et al., 2016b). Inicialmente aplicada na análise de debates políticos em fóruns *online*, a DP se mostrou altamente atrativa para medir a opinião pública em redes sociais, especialmente no Twitter, atualmente conhecida como "X" (ALDAYEL; MAGDY, 2019).

No campo de Aprendizado de Máquina (AM), a DP é frequentemente abordada como um problema de classificação supervisionado (ALTURAYEIF; LUQMAN; AHMED, 2023) e isso implica desafios típicos desse tipo de paradigma, especialmente os relacionados à escassez de dados rotulados. Rotular dados é uma tarefa onerosa e demorada, o que normalmente leva a conjuntos de treinamento pequenos (AL-GHADIR; AZMI; HUSSAIN, 2021). Diante dessa restrição, os modelos convencionais de DP frequentemente se apóiam em métodos complexos de engenharia de características (*features*), o que afeta negativamente sua generalização e reprodutibilidade (ALTURAYEIF; LUQMAN; AHMED, 2023). Além disso, em cenários que empregam modelos de Aprendizado Profundo (AP), essa deficiência pode intensificar a propensão ao superajuste, comprometendo a robustez e a confiabilidade dos modelos (HAN et al., 2021).

Como alternativa, estratégias de AM não-supervisionado (DARWISH et al., 2020; RASHED et al., 2021; WEI; MAO; CHEN, 2019) emergem como uma solução promissora, mas, apesar de seu potencial, foram pouco exploradas no campo de DP. Na mesma direção, técnicas de AM fracamente supervisionadas foram exploradas mas não apresentaram desempenho satisfatório. Recentemente, técnicas de AM por transferência de aprendizado baseadas em aprendizado com poucos exemplos (*Few-shot Learning*) (LUO et al., 2022; ALLAWAY; SRIKANTH; MCKEOWN, 2021; CONFORTI et al., 2021) surgem como fortes candidatas no combate às limitações impostas pela falta de dados rotulados. No entanto, as abordagens atuais são de difícil reprodução, pois requerem arquiteturas complexas para atingir um nível satisfatório de generalização, e além disso, não apresentam bons níveis de desempenho quando comparados a abordagens de aprendizado supervisionado (ALTURAYEIF; LUQMAN; AHMED, 2023).

Nesse contexto, modelos de linguagem Pré-Treinados (MPTs) tais como o BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2018) e o GPT (Generative Pre-Trained Transformer) (RADFORD et al., 2019), representam um

avanço significativo no campo de PLN (Processamento de Linguagem Natural). Ao serem treinados a partir de extensos conjuntos de dados, os MPTs capturam não só o contexto, mas também são capazes de extrair propriedades sintáticas e semânticas das palavras em documentos de texto. O conhecimento previamente adquirido pode ser transferido para outros domínios reduzindo significativamente a complexidade no processo de treinamento e implantação de modelos de AM (HAN et al., 2021). Esta característica se torna especialmente valiosa em aplicações como a Busca Semântica, onde os MPTs demonstram capacidade em identificar, com agilidade e precisão, múltiplas postagens que compartilham posicionamentos semelhantes quando fornecidos alguns exemplos como argumentos de busca (MUENNIGHOFF, 2022).

Visando explorar eficientemente as vantagens dos MPTs em capturar o conteúdo semântico e contextual em *tweets* através da Busca Semântica, esta dissertação introduz o método *Sematic Search Stance Detection* (SSSD) (Detecção de Posicionamento por Busca Semântica). Esta estratégia adota uma abordagem de *Few-shot Learning* para a DP, estruturada em duas etapas essenciais. A primeira, denominada Rotulagem Semântica, envolve a rotulagem automática de um vasto corpus de domínio específico a partir de pequenas amostras de *tweets* já rotulados. Esta fase estabelece as fundações para a segunda etapa, na qual se utiliza o conjunto de dados, filtrado e ampliado para o treinamento efetivo de modelos de DP. Resultados preliminares foram reportados em (SOUSA; BECKER, 2023).

A principal vantagem do SSSD é seu potencial para economizar recursos, ao reduzir drasticamente a necessidade de rotulação manual de *tweets*. Utilizando poucos exemplos, o processo de Rotulagem Semântica consegue, rotular automaticamente um grande volume de *tweets*, ao mesmo tempo em que melhora significativamente o desempenho e a qualidade das técnicas de DP. O método não só atenua os desafios trazidos pela falta de dados rotulados, como também atua como um sistema de filtragem, focando nas postagens mais relevantes e desprezando conteúdos ruidosos ou não relacionados aos alvos. Adicionalmente, o uso de MPTs permite que a DP seja realizada em domínios de diferentes linguagens. No melhor do nosso conhecimento, o SSSD é a primeira abordagem de DP que combina MPTs e Busca Semântica para enfrentar os desafios impostos pela limitação de dados rotulados.

Nesta dissertação, o SSSD foi avaliado sob duas perspectivas distintas: qualitativa e quantitativa. Na abordagem quantitativa, são empregados alvos, métricas e conjuntos de dados predefinidos para as Tarefas A e B do SemEval-2016 Tarefa 6 (MOHAMMAD

et al., 2016b), em que o SSSD superou o desempenho dos atuais sistemas estado da arte propostos por Zhao e Yang (2021) (Tarefa A) e Lai et al. (2017) (Tarefa B). Dentro deste contexto, também foram avaliados os fatores que mais influenciam a quantidade de *tweets* rotulados e o desempenho na detecção de posicionamentos.

A análise qualitativa se concentrou no alvo "Vacinação", recorrendo a *tweets* coletados durante o período da pandemia de COVID-19 no Brasil, conforme estudo desenvolvido em (SOUSA; BECKER, 2021). O foco central foi examinar a aplicabilidade e desempenho do SSSD em ambientes que carecem de conjuntos de dados previamente rotulados. Nos testes realizados, foi demonstrada uma forma prática de aplicar o SSSD, recorrendo a recursos de Modelagem de Tópicos para identificar posicionamentos relevantes. O desempenho alcançado confirmou a premissa de que o SSSD produz resultados superiores com um menor número de instâncias de treino, além de ter mostrado habilidade em gerenciar inconsistências e ruídos nos dados.

A organização desta dissertação segue a seguinte estrutura: O Capítulo 2 explora os elementos fundamentais das metodologias aplicadas no processo de DP no Twitter usando o SSSD. O Capítulo 3 fornece um panorama dos estudos existentes em DP, caracterizando o estado da arte e as lacunas abordadas por este trabalho. O Capítulo 4 detalha o método SSSD, descrevendo as suas etapas, componentes e dados necessários. Uma avaliação quantitativa do SSSD é apresentada no Capítulo 5, enfocando as métricas e alvos definidos para as Tarefas A e B do SemEval-2016 Tarefa 6. O Capítulo 6 amplia a discussão com uma análise qualitativa, empregando o SSSD no contexto da "Vacinação" durante a pandemia de COVID-19 no Brasil. O Capítulo 7, destaca as conclusões desta dissertação e direções de trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo detalha aspectos fundamentais para entender as metodologias empregadas no processo de DP utilizando o SSSD, incluindo *Embeddings*, MPTs, Busca Semântica e Classificação de Texto.

## 2.1 Embbedings

*Embeddings* são representações vetoriais amplamente empregadas em diversas tarefas de PLN, desde classificação de texto e detecção de posicionamentos, tradução automática e análise de sentimento (LI et al., 2022). Eles capturam nuances semânticas e contextuais de palavras, frases e até documentos inteiros. Derivados de avançadas técnicas de AM, como redes neurais profundas, os *embeddings* mapeiam complexidades linguísticas em um espaço vetorial denso e contínuo, otimizando o processamento e realçando informações cruciais (HAN et al., 2021).

O grande diferencial dos *embeddings* está na sua capacidade de oferecer representações densas e de baixa dimensionalidade, permitindo que os modelos de PLN operem de forma mais eficiente com textos, o que é fundamental para a análise aprofundada e o reconhecimento de padrões em grandes conjuntos de dados textuais. Ao representar as palavras em um espaço vetorial, é possível identificar e utilizar a proximidade semântica entre as palavras para realizar tarefas complexas de PLN com maior precisão e eficiência.

Os métodos de criação de *embeddings* estáticos, como Word2Vec (MIKOLOV et al., 2013) ou Doc2Vec (LE; MIKOLOV, 2014) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) revolucionaram o campo de PLN, alcançando uma ampla popularidade. No entanto, essas abordagens têm a limitação significativa de gerar uma única representação global para cada palavra, o que ignora as nuances contextuais e, assim, não pode lidar adequadamente com a polissemia, i.e., a ocorrência de múltiplos sentidos para uma única palavra, uma característica intrínseca da linguagem natural (CHEN et al., 2018; HAN et al., 2021).

Essa deficiência foi substancialmente mitigada pelo advento dos modelos contextuais, como o ELMo (ILIĆ et al., 2018), e aqueles baseados na revolucionária arquitetura de Transformers (VASWANI et al., 2017). Esses modelos, ao contrário de seus predecessores, são capazes de gerar *embeddings* que levam em consideração o contexto em que uma palavra está inserida, permitindo assim representações mais ricas e precisas que re-

fletem o verdadeiro uso das palavras em diferentes contextos. Esses avanços tecnológicos abriram portas para uma compreensão muito mais profunda da linguagem, potencializando a eficiência e a eficácia dos sistemas de PLN.

## 2.2 Modelos Pré-treinados

Nas abordagens mais recentes de PLN, a tendência predominante é representar as características intrínsecas dos textos por meio de *embeddings* contextuais (ALTURAYEIF; LUQMAN; AHMED, 2023). Esses *embeddings* são gerados principalmente a partir de MPTs baseados em Transformers, como o BERT (DEVLIN et al., 2019) e o GPT (RADFORD et al., 2019).

MPTs são pré-treinados em extensas quantidades de dados, como textos, imagens e áudio por meio de algoritmos avançados de AM. Sua essência reside na capacidade de extrair representações ricas e abrangentes dos dados durante o treinamento em larga escala. Essas representações são capturadas pelos parâmetros do modelo, que podem ser considerados como um conhecimento prévio ou uma "intuição" sobre o domínio dos dados (HAN et al., 2021).

Desta forma, graças à rica compreensão da linguagem já estabelecida durante o pré-treinamento, os MPTs atenuam a necessidade de vastos conjuntos de dados e do alto consumo de recursos computacionais que seriam imprescindíveis ao se desenvolver um modelo competente a partir do zero. Isso não só otimiza o processo de implementação de soluções de AM, mas também amplia significativamente a eficácia e a acessibilidade dessas tecnologias avançadas em diversos campos de aplicação (HAN et al., 2021).

Existem muitos MPTs, entre eles o "all-MiniLM-L6-v2" (WANG et al., 2020a) e o Microsoft MPNET (SONG et al., 2020), disponíveis no *framework* SentenceTransformers[1] (SBERT) (REIMERS; GUREVYCH, 2019). SBERT modifica a estrutura de rede do BERT e seus derivados, empregando redes siamesas capazes de gerar *embeddings* de frases que podem ser comparadas por similaridade de cosseno. Este avanço metodológico representa uma evolução significativa pois permitiu que MPTs baseados no BERT fossem usados em tarefas que até então não eram viáveis antes, incluindo Agrupamento e Busca Semântica (REIMERS; GUREVYCH, 2019).

Ambos MPNET e "all-MiniLM-L6-v2" foram projetados para ser utilizados como codificadores de frases e parágrafos curtos. A concatenação de múltiplos conjuntos de

[1] https://www.sbert.net/docs/pretrained_models.html

dados[2] é utilizada para o ajuste fino dos modelos.

Nesta dissertação os MPTs "all-MiniLM-L6-v2" e MPNET foram utilizados como base da Busca Semântica no SSSD e na Modelagem de Tópicos discutida na Seção 6.3.1 respectivamente.

## 2.3 Busca Semântica

A Busca Semântica é um termo usado para definir os mecanismos de busca que empregam técnicas de AM para discernir o significado e a intenção por trás das palavras e frases utilizadas nas consultas (MUENNIGHOFF, 2022). A Busca Semântica compreende dois elementos fundamentais:

- ***Busca:*** envolve o processo de recuperar as $k$ respostas mais semelhantes de um *corpus* documental com base em uma consulta específica, conforme ilustrado na Figura 2.1.
- ***Semântica:*** relacionado à compreensão dos documentos e das consultas, considerando também seu contexto, indo além do mero reconhecimento de palavras-chave.

Neste sentido, os MPTs emergiram como uma das principais abordagens empregadas para Busca Semântica, proporcionando uma alternativa aos modelos probabilísticos que adotam abordagens não semânticas, como o BM25 (MUENNIGHOFF, 2022; ROBERTSON; ZARAGOZA et al., 2009).

A essência desse processo consiste em mapear as entradas do corpus, que podem ser frases, parágrafos ou documentos inteiros, para um espaço vetorial através de *embeddings* usando MPTs. Nesse espaço, a proximidade entre vetores sugere similaridade semântica. Durante a Busca Semântica, a consulta é transformada e posicionada nesse mesmo espaço vetorial. A partir desse ponto, são identificados e apresentados os *embeddings* mais próximos no corpus, usando alguma medida de distância (e.g., cosseno, produto escalar), que possuem uma forte congruência semântica com a consulta (MUENNIGHOFF, 2022).

Para cada iteração de busca neste processo, os *Cross-Encoders*, tais como o BERT ou RoBERTa (ZHUANG et al., 2021), codificam cada consulta em relação ao todos os documentos do corpus simultaneamente, separando-os com um *token* [SEP] (MUENNIGHOFF, 2022). No entanto, isso gera uma sobrecarga computacional massiva (REI-

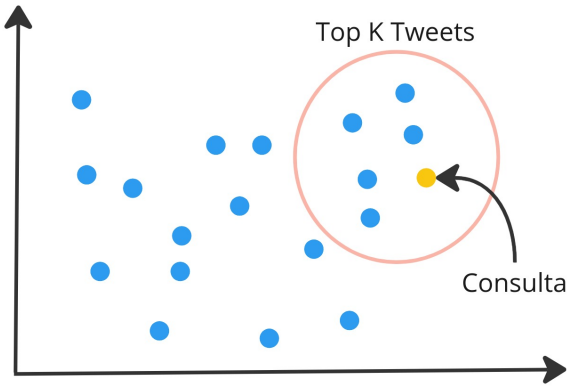
[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

MERS; GUREVYCH, 2019). Por outro lado, os *Bi-Encoders* baseados no SBERT, como o MPNET e MiniLM, codificam a consulta e os documentos separadamente, permitindo que os *embeddings* do corpus sejam armazenados em *cache* e reaproveitados posteriormente a cada consulta, reduzindo consideravelmente o tempo necessário para encontrar os pares mais similares, mantendo um nível de precisão comparável (REIMERS; GUREVYCH, 2019).

A Busca Semântica é dita simétrica quando o texto da consulta e do corpus têm tamanhos e conteúdos semelhantes. Neste caso, é possível inverter a consulta e as respostas para encontrar as mesmas correspondências. Por outro lado, a busca semântica é assimétrica quando a consulta é curta e se deseja encontrar documentos mais extensos que respondam à consulta. Nesse caso, inverter a consulta e as respostas provenientes do corpus não é apropriado, pois a consulta é geralmente uma pergunta ou palavras-chave, enquanto os documentos resultantes devem fornecer informações mais abrangentes (MUENNIGHOFF, 2022).

A distinção entre Busca Semântica simétrica e assimétrica é importante, pois afeta o processo de pesquisa e os algoritmos utilizados para encontrar as melhores correspondências semânticas entre a consulta e as entradas do corpus. A escolha adequada entre os dois tipos de busca depende da natureza da consulta e dos objetivos da pesquisa. É crucial considerar essas diferenças ao aplicar abordagens de busca semântica para garantir resultados relevantes e precisos de acordo com o contexto específico.

Figura 2.1: Processo de Busca Semântica




Fonte: o autor.

Nesta dissertação, foi empregada Busca Semântica simétrica explorando MPTs baseados em SBERT em conjunto com similaridade de cosseno.

## 2.4 Classificação de Texto

A Classificação de Texto é uma técnica de AM que atribui categorias ou rótulos pré-definidos a trechos de texto, sendo essencial em diversas aplicações de PLN. Ela tem aplicações variadas, desde a detecção de emoções e opiniões na análise de sentimento até sistemas especializados em perguntas e respostas, conforme mencionado por Li et al. (2022).

Os modelos de AM empregados para esta tarefa se segmentam, de forma geral, em dois grandes grupos. O primeiro grupo compreende os modelos tradicionais de AM como Naïve Bayes (NB), o K-Nearest Neighbor (KNN) e Support Vector Machine (SVM). Esses modelos são conhecidos por sua robustez e eficácia em *datasets* de tamanho moderado (LI et al., 2022). No segundo grupo, os modelos baseados AP, são particularmente eficazes em lidar com grandes volumes de dados e capturar nuances e relações complexas no texto. Dentre eles, destacam-se as as Convolutional Neural Network (CNNs), Gated Recurrent Unit (GRUs) e até mesmo customização (*fine-tunning*) de MPTs como o BERT.

O processo de classificação de texto começa com uma etapa essencial de pré-processamento que prepara o texto para a classificação, removendo elementos não essenciais ou transformando o texto em uma forma mais tratável. Ações comuns nesse estágio incluem a remoção de *stop-words*, i.e., termos frequentes que geralmente não contribuem para o significado do texto, como “e”, “ou” e “mas”. Também é comum a lematização, que reduz palavras a sua forma canônica, assim, por exemplo, “correndo” e "corredor"podem ser reduzidos à mesma forma, “corre”. Além disso, caracteres especiais ou elementos específicos que podem não ter relevância analítica são frequentemente removidos (LI et al., 2022). Em postagens sociais são comuns outras práticas tais como a remoção de URLs ou se menções.

Quando são aplicados algoritmos tradicionais de AM, após o pré-processamento, o texto também passa por uma etapa de engenharia de *features* onde é convertido em representações vetoriais utilizando técnicas como Bag-Of-Words (BOW) (ZHANG; JIN; ZHOU, 2010), N-gram (CAVNAR; TRENKLE et al., 1994) e Term Frequency-Inverse Document Frequency (TF-IDF) (BAEZA-YATES; RIBEIRO-NETO et al., 1999). Por outro lado, em abordagens baseadas em AP, a etapa de engenharia de *features* é desnecessária pois esses modelos aprendem as representações vetoriais (*embeddings*) diretamente do texto bruto (LI et al., 2022).

Na fase de treinamento, os modelos de classificação são alimentados pelos veto-

res associados a seus respectivos rótulos (conjunto de treino). Segue-se uma fase de teste, durante a qual o modelo tem seu desempenho é medido com base em métricas como F1, Acurácia, Precisão e Revocação utilizando um conjunto de testes. O método SSSD proposto nesta dissertação admite qualquer abordagem de classificação de texto, tradicional ou AP. Nos experimentos, foram usadas abordagens tradicionais de AM com engenharia de *features*, e as métricas supracitadas para avaliação dos modelos de DP.

# 3 TRABALHOS RELACIONADOS

Este capítulo descreve um panorama dos estudos existentes em DP, caracterizando o estado da arte e as lacunas abordadas por este trabalho.

## 3.1 Detecção de Posicionamentos

Com o avanço tecnológico proporcionado pela internet, diversas plataformas online emergiram, impulsionando a geração de Conteúdo Gerado pelo Usuário (CGU) em variados formatos, abrangendo texto, áudio e imagens. As redes sociais desempenham um papel fundamental, permitindo a comunicação e expressão de opiniões. Esse cenário resultou em um aumento significativo na disponibilidade de dados, fomentando a demanda pelo processamento automatizado do CGU, principalmente por meio de técnicas de AM e PLN(ALTURAYEIF; LUQMAN; AHMED, 2023).

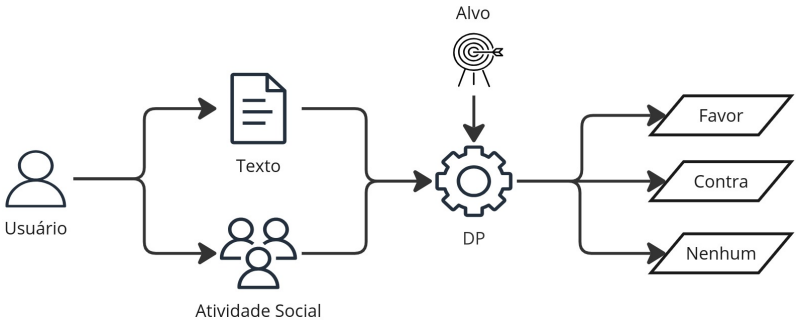
A DP emerge como um método poderoso para extrair conhecimento dessa rica fonte de dados, avaliando opiniões, posturas e atitudes das pessoas em relação a produtos, eventos, serviços e outros tópicos relevantes principalmente a partir de conteúdo em texto (ALTURAYEIF; LUQMAN; AHMED, 2023). A DP tem potencial para desempenhar um papel crucial em tomadas de decisões para empresas, formuladores de políticas, políticos e indivíduos em geral. Isso inclui a detecção de posicionamentos expressos em textos, como a verificação de rumores e a identificação de notícias falsas (NGUYEN et al., 2016; KÜÇÜK; CAN, 2020).

Embora as pesquisas iniciais sobre DP tenham começado com dados provenientes de debates políticos em fóruns online, o Twitter rapidamente se estabeleceu como a principal fonte de dados para estudos de DP, destacando a importância e popularidade das redes sociais para pesquisa e desenvolvimento nesta área (ALDAYEL; MAGDY, 2021). A predominância do uso do Twitter também pode ser atribuída às suas considerações mais flexíveis em termos de acessibilidade e ética na extração de dados por meio de suas APIs, quando comparado com outras plataformas de mídia social, como o Facebook, que impõem obstáculos mais significativos no processo de coleta de dados (ALTURAYEIF; LUQMAN; AHMED, 2023).

No âmbito de AM, o termo “Detecção de Posicionamento” é utilizado para definir um problema de classificação (ALDAYEL; MAGDY, 2021). Esta abordagem visa determinar, a partir de um texto e/ou da atividade social associada, se o autor expressa uma

posição a favor, contra ou nenhuma em relação a um alvo específico, como esquematizado na Figura 3.1. Alvos podem abranger pessoas, organizações ou ideias (MOHAMMAD et al., 2016b). Embora alguns estudos incluem a categoria "Neutro", sugerindo que o autor possui uma posição neutra em relação ao alvo (GRIMMINGER; KLINGER, 2021), argumenta-se que uma posição neutra não existe, visto que as pessoas tendem a se posicionar a favor ou contra uma proposição (JAFFE, 2009). Além disso, há um consenso na literatura de que, quando o posicionamento de um texto em relação a um alvo não é favorável nem contrário, a categoria apropriada seria "Nenhum" ao invés de "Neutro", pois não é possível obter informações de posicionamento do texto (ALTURAYEIF; LUQMAN; AHMED, 2023).

Figura 3.1: Fluxo de Detecção de Posicionamentos




Fonte: adaptada de Alturayeif, Luqman e Ahmed (2023).

A tarefa de DP ganhou notoriedade com o surgimento de uma série de competições internacionais dedicadas à avaliação e avanço dos mais recentes métodos em PLN e Linguística Computacional, conhecidos como SemEval (Avaliação Semântica) (ALTURAYEIF; LUQMAN; AHMED, 2023). O SemEval tornou-se uma referência, proporcionando um ambiente onde pesquisadores e profissionais podem avaliar, testar e comparar suas metodologias e algoritmos frente a uma diversidade de desafios da PLN. Notavelmente, a Tarefa 6 do SemEval em 2016, centrada em DP, ganhou destaque e reconhecimento na comunidade científica. Esta competição foi dividida em dois segmentos: a Tarefa A, voltada para técnicas supervisionadas, e a Tarefa B, destinada a abordagens não supervisionadas ou fracamente supervisionadas.

O corpus disponibilizado no SemEval-2016 Tarefa 6 é relativo à língua inglesa, o que torna a maioria dos trabalhos nesta área focados nesta língua (ALTURAYEIF; LUQMAN; AHMED, 2023). Outras seis competições voltadas para DP já foram realizadas, contribuindo para o avanço da pesquisa nesta área ao oferecerem conjuntos de dados anotados em diferentes idiomas, diretrizes de anotação, métricas de avaliação e uma visão geral das equipes participantes. Os detalhes dessas competições são apresentados a seguir

em ordem cronológica:

1. **NLPCC-2016 Tarefa 4:** Competição de detecção de posicionamento em microblogs chineses com duas subtarefas semelhantes ao SemEval-2016 Tarefa 6 (XU et al., 2016).
2. **IberEval-2017:** Tarefa compartilhada foi realizada para detecção de posicionamento e gênero em *tweets* em espanhol e catalão (TAULÉ et al., 2017).
3. **SemEval-2017 Tarefa 8:** Tarefa compartilhada para identificar rumores e o posicionamento dos usuários do Twitter por meio de suas respostas textuais (DERCZYNSKI et al., 2017).
4. **SemEval-2019 Tarefa 7:** Tarefa compartilhada composta por duas atividades: verificação de rumor e previsão de posicionamento sobre o rumor em publicações no Twitter e Reddit (GORRELL et al., 2019).
5. **EVALITA-2020 (SardiStance)**: O *SardiStance*, realizado durante a conferência EVALITA-2020 (CIGNARELLA et al., 2020), foi a primeira tarefa compartilhada para detecção de posicionamento na língua italiana. Esta competição também compreendeu duas sub-tarefas: A Tarefa A está relacionada à detecção de posicionamento textual, e a Tarefa B é baseada na detecção de posicionamento contextual que utiliza informações adicionais da rede social do usuário e *tweets*, bem como informações sobre o perfil do usuário.

## 3.2 Features para Detecção de Posicionamentos em Redes Sociais

As *features* empregadas em modelos para DP em redes sociais, geralmente se enquadram em duas categorias:

- **Conteúdo Textual:** Derivadas de elementos linguísticos, sintáticos, semânticos, estruturas gramaticais, sentimentos expressos e particularidades do estilo de escrita presentes nas postagens.
- **Usuário:** Oriundas de interações sociais, conexões de rede e metadados dos usuários.

Estudos recentes (WEI; XU; MAO, 2019; ZHANG et al., 2020; LIU et al., 2021; LIANG et al., 2021) focam principalmente no nível do conteúdo. As *features* são tipicamente extraídas do texto usando métodos de engenharia de *features* (e.g. *n-grams*, *tf-idf*) ou explorando *embeddings*, sejam eles estáticos (e.g., GloVe, Word2Vec) ou contextuais, oriundos de MPTs sofisticados como o BERT. Adicionalmente, visando enriquecer o conteúdo textual, é comum a incorporação de informações externas, tais como léxicos de

sentimentos e emoções (e.g., (DEY; SHRIVASTAVA; KAUSHIK, 2018; AL-GHADIR; AZMI; HUSSAIN, 2021)).

Comparativamente, abordagens que se baseiam unicamente em *features* de usuário (ALDAYEL; MAGDY, 2019; DARWISH et al., 2020) são menos frequentes em pesquisas de DP, utilizando interações, preferências e vínculos de rede para determinar similaridades entre usuários. A tendência de usar *features* de usuário para detectar posicionamentos é sustentada pelo princípio da homofilia, que sugere que indivíduos tendem a se associar àqueles que são semelhantes entre eles (BESSI et al., 2016).

## 3.3 Abordagens para Detecção de Posicionamentos

Nesta seção, foram descritas as abordagens de DP que serviram de inspiração para o desenvolvimento do SSSD. As metodologias de AM discutidas, são categorizadas de maneira abrangente em Supervisionadas, Não Supervisionadas, Fracamente Supervisionadas e baseadas em Transferência de Aprendizado. Cada categoria é discutida detalhadamente no restante desta seção.

### 3.3.1 Abordagens Supervisionadas

A maioria dos estudos sobre DP empregam técnicas tradicionais de classificação usando AM supervisionado. Nessa abordagem, o foco é treinar modelos utilizando dados previamente rotulados para alvos ou domínios específicos, com a expectativa de que esses modelos apresentem um desempenho satisfatório para classificar novos dados dentro do mesmo contexto (ALTURAYEIF; LUQMAN; AHMED, 2023). Dentre os algoritmos mais adotados, o SVM destaca-se como uma escolha predominante (LAI et al., 2020; PAMUNGKAS; BASILE; PATTI, 2019; MOHAMMAD; SOBHANI; KIRITCHENKO, 2017; ALDAYEL; MAGDY, 2019; HACOHEN-KERNER; IDO; YA'AKOBOV, 2017; LAI et al., 2017; SUN et al., 2019b).

No estudo proposto por Dey, Shrivastava e Kaushik (2017), os autores desenvolveram uma abordagem para DP em *tweets* utilizando um classificador SVM. Inicialmente, fizeram uso de léxicos de sentimentos para identificar posicionamentos neutros nos *tweets*. Posteriormente, os *tweets* não neutros resultantes da primeira etapa, foram categorizados como sendo a favor ou contra os alvos. Já Al-Ghadir, Azmi e Hussain (2021) utilizam

léxicos de sentimentos e listas classificadas de palavras ponderadas por *tf-idf* para treinar classificadores KNN. Apesar de apresentar um dos melhores desempenhos no SemEval-2016 Tarefa 6, seus detalhes operacionais são obscuros, dificultando sua reprodutibilidade (GÓMEZ-SUTA; ECHEVERRY-CORREA; SOTO-MEJÍA, 2023). Outros estudos (ALDAYEL; MAGDY, 2019; LYNN et al., 2019; DARWISH et al., 2018) usam *features* de usuário para melhorar o desempenho do classificador. Contudo, essas abordagens requerem dados adicionais sobre o comportamento do usuário, o que limita sua aplicabilidade fora das plataformas de mídia social.

Lai et al. (2017) propõem uma abordagem supervisionada para DP em *tweets*, especificamente em relação aos alvos "Hillary Clinton" e "Donald Trump" usando os conjuntos de dados do SemEval-2016 Tarefa 6. Foram considerados três grupos de características para realizar a classificação: baseadas em sentimentos, rede e baseadas em contexto. Para validar a eficácia da abordagem, os autores utilizaram um classificador Gaussian Naive Bayes, experimentando diferentes combinações de características. O modelo que obteve os melhores resultados foi aquele que combinou características de usuário análise de sentimentos e compreensão do contexto político específico com pontuações de 74,5% e 71,2 % para os alvos "Hillary Clinton" e "Donald Trump" respectivamente.

Como alternativa às abordagens tradicionais de classificação, muitos trabalhos exploram o AP. As RNNs, são um destaque neste cenário, dada a sua capacidade intrínseca de processar e reconhecer padrões em dados sequenciais, como os encontrados na linguagem natural (SUN et al., 2019a; BORGES; MARTINS; CALADO, 2019; PODDAR et al., 2018). As LSTMs, uma variação das RNNs, são projetadas especialmente para lidar com longas sequências de texto e suas dependências temporais, destacam-se como a arquitetura de AP mais frequentemente adotada em estudos supervisionados para DP (HOSSEINIA; DRAGUT; MUKHERJEE, 2020; KOCHKINA; LIAKATA; AUGENSTEIN, 2017; LAI et al., 2020; SUN et al., 2018; ZHU; HE; ZHOU, 2020).

Outros estudos optaram por variações destas arquiteturas de AP, como as BiLSTMs (ZHANG et al., 2020; AHMED; CHY; CHOWDHURY, 2020; YANG et al., 2020; CHEN; YE; CUI, 2021; LIANG et al., 2021). Estas analisam sequências de textos tanto na direção original quanto na inversa, garantindo uma visão contextual mais completa. Em paralelo, as GRUs surgiram como uma opção valiosa devido à sua estrutura mais simplificada, com um menor número de portas (*gates*), as GRUs facilitam o treinamento e frequentemente superam em desempenho, sobretudo em tarefas como DP (HOSSEINIA; DRAGUT; MUKHERJEE, 2020; ZHOU; CRISTEA; SHI, 2017; WEI;

MAO; ZENG, 2018; ZHU; HE; ZHOU, 2020; BHATT et al., 2018). Apesar da eficiência das LSTMs e GRUs em séries temporais, elas não extraem *features* espaciais do texto tão bem quanto as CNNs (EL-ALFY; LUQMAN, 2022). Para superar essa deficiência, alguns pesquisadores usam abordagens hibridas combinando CNNs com LSTMs ou GRUs para extração de características (LI; XU; WANG, 2019; MOHTARAMI et al., 2018).

As técnicas supervisionadas de DP apresentam como principal vantagem seu desempenho, desde que contem com uma anotação de dados de qualidade e algoritmos adequados. Sua principal limitação está na dependência de um volume considerável de dados rotulados, especialmente quando se adota AP (HAN et al., 2021). A rotulação de dados, além de ser um processo oneroso, é também demorado, o que frequentemente resulta em bases de tamanho reduzido além de ser dependente de linguagem (AL-GHADIR; AZMI; HUSSAIN, 2021).

Diante desse desafio, uma das questões centrais em AM é desenvolver modelos de AP que sejam robustos mesmo com uma quantidade limitada de dados rotulados (HAN et al., 2021). Além disso, considerando a diversidade de idiomas e a natureza dos problemas de PLN, ter dados rotulados por humanos para cada situação específica torna-se uma meta impraticável (ALTURAYEIF; LUQMAN; AHMED, 2023). Uma síntese das técnicas de DP discutidas nesta seção que obtiveram os melhores resultados é ilustrada na Tabela 3.1, ressaltando features, métodos de AM, datasets usados e desempenho. Esta tabela é discutida em mais detalhes na Seção 3.4.

### 3.3.2 Abordagens Não-Supervisionadas

Técnicas não-supervisionadas emergem como uma solução promissora ao desafio da escassez de dados rotulados. Apesar do seu potencial, ainda foram pouco exploradas no campo da DP (GÓMEZ-SUTA; ECHEVERRY-CORREA; SOTO-MEJÍA, 2023). Entre as estratégias não-supervisionadas, é notável a ênfase dada às técnicas de agrupamento usando *features* de rede (AL-GHADIR; AZMI; HUSSAIN, 2021).

Darwish et al. (2020), aplicaram redução de dimensionalidade para projetar *features* de redes em um espaço de baixa dimensão com UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction), seguido de agrupamento com DBSCAN (Density-Based Spatial Clustering of Applications with Noise), assumindo que os usuários em cada grupo teriam o mesmo posicionamento em relação aos alvos. Posteriormente calculou-se uma variante escore de valência usado por Conover et al. (2011) a

partir das contas mais mencionadas e *hashtags* mais relevantes para rotular automaticamente os *tweets* de cada grupo com base nas características mais frequentes, alcançando uma pontuação F1-Macro de 90.4% em um conjunto de dados relacionados a Tópicos sobre eleições nos EUA.

Expandindo a abordagem anterior, Samih e Darwish (2021) identificaram os usuários mais ativos por tópico e calcularam a semelhança entre eles com base nas *hashtags* que utilizadas ou *retweets*. Em seguida, os usuários são projetados em um espaço de menor dimensão usando UMAP, de uma maneira em que usuários semelhantes são aproximados e agrupados usando HDBSCAN, enquanto usuários dissimilares são distanciados. Com essa estratégia, eles obtiveram uma pontuação F1-Macro de 92,1%.

De forma semelhante, Rashed et al. (2021) utilizaram o codificador de sentenças universais multilíngue baseado em CNNs do Google para projetar usuários em um espaço de baixa dimensão. Eles também usaram UMAP para a redução de dimensionalidade e HDBSCAN para o agrupamento, aplicando essa abordagem a um conjunto de dados relacionado às eleições turcas. Essa metodologia permitiu-lhes agrupar usuários com posicionamentos semelhantes, obtendo uma pontuação F1 de 74,0%.

Todos os trabalhos citados empregaram *features* de rede para agrupar usuários e, em seguida, classificá-los como "Favor" ou "Contra" aos alvos em questão. Dessa forma, as métricas e os classificadores não foram treinados ou avaliados considerando a classe "Neutro", sugerindo que os resultados mostrados por esses estudos, apesar de bons, não podem ser comparados com trabalhos avaliados com as métricas do SemEval-2016 Tarefa 6.

Além disso, métodos de DP focados em características do usuário se limitam a plataformas de CGU que permitem acesso aos dados dos usuários. Com o aumento das preocupações sobre privacidade, as plataformas de mídia social vêm impondo restrições cada vez mais severas ao compartilhamento dessas informações, restringindo a eficácia de técnicas que dependem desses dados (ALTURAYEIF; LUQMAN; AHMED, 2023). Um resumo dos trabalhos é apresentado na Tabela 3.2.

### 3.3.3 Abordagens Fracamente-Supervisionadas

Nesta abordagem, o modelo aprende tanto a partir de dados rotulados quanto não rotulados. Alguns estudos anotaram dados automaticamente empregando um classificador baseado em regras (AUGENSTEIN et al., 2016; DIAS; BECKER, 2016; WEI; MAO;

CHEN, 2019). O método proposto por Dias e Becker (2016) recebe como entrada *n-grams* que representam posicionamento em relação alvos específicos e termos comuns usados para denotar posição i.e. *hashtags*, juntamente com léxicos de sentimento, para compor automaticamente um grande corpus de treinamento. Em seguida, aplica um algoritmo de aprendizado supervisionado para desenvolver um modelo de DP.

Wei, Mao e Chen (2019) propuseram uma técnica que compreende duas fases. A primeira envolve a anotação automática de um conjunto de *tweets* de domínio, do qual os tópicos são extraídos. Esses tópicos são usados para treinar uma rede de detecção (TDNet) para identificar os posicionamentos (“Favor”, “Contra”) que os diferentes *tweets* expressam em relação a diferentes alvos. Paralelamente, uma segunda rede (SRNet) é usada para refinar os dados de treinamento, removendo instâncias ruidosas ou dados que são confusos ou enganosos para o modelo (“Nenhum”) através de um processo de aprendizado de reforço, onde “recompensas” são dadas com base na precisão das predições da TDNet, incentivando a SRNet a eliminar dados que estão levando a predições incorreta.

Embora os métodos fracamente-supervisionados abordam a questão da falta de dados rotulados, eles não apresentam bons desempenhos quando comparados com algoritmos de DP supervisionados, não-supervisionados ou de transferência de aprendizado, conforme mostrado na Tabela 3.3. Além disso, o longo tempo de treinamento e a má generalização e sua complexidade são as principais limitações dessa abordagem de aprendizado (ALTURAYEIF; LUQMAN; AHMED, 2023).

### 3.3.4 Abordagens baseadas em Transferência de Aprendizado

No domínio da AM e, extensivamente em PLN, as técnicas de transferência de aprendizado têm se destacado como a estratégia mais eficaz para abordar a escassez de dados rotulados (RUDER, 2019). Pesquisas conduzidas por Kawintiranon e Singh (2021) exploraram o ajuste fino de MPTs em domínios específicos a fim de desenvolver um modelo de DP apto a operar em cenários com poucos dados rotulados. Ao recorrer ao conhecimento anteriormente adquirido, particularmente em termos de características semânticas e sintáticas inerentes ao contexto-alvo, torna-se possível treinar o classificador utilizando um conjunto reduzido de exemplos rotulados (HAN et al., 2021).

Muitos estudos exploram mecanismos de atenção como estratégia de classificação (XU et al., 2018; SUN et al., 2022; KAWINTIRANON; SINGH, 2021). No trabalho proposto por Xu et al. (2018), os autores introduziram um modelo baseado em auto-

atenção que demonstrou um desempenho superior a diversos métodos de referência em várias áreas, conforme evidenciado pelos resultados experimentais. Na mesma direção, Sun et al. (2022) desenvolveram uma rede de atenção que deduz a correlação entre postagens associadas a diferentes alvos. Esta análise se baseia nos sentimentos expressos em cada *tweet* para determinar e compreender seu posicionamento. Em contrapartida, Kawintiranon e Singh (2021) ajustaram o BERT em um conjunto de dados não rotulados relacionado às eleições americanas de 2020 para prever a postura em relação aos candidatos presidenciais de 2020, a saber, Joe Biden e Donald Trump. Eles também usaram mecanismos de atenção para detectar posicionamentos ao focar nas palavras distintivas de cada postura.

Notavelmente, Zhao e Yang (2021) propõem uma Rede de Cápsula Hierárquica chamada PE-HCN que adota o modelo RoBERTa para geração de *embbedings* contextuais usando características de conteúdo que alcançou o atual estado-da-arte para DP baseado na Tarefa A do SenEval 2016. Os resultados obtidos na avaliação demonstram um avanço considerável ao atingir uma pontuação de 78,4% conforme a métrica aplicada na referida competição. Também usando o SemEval-2016 Tarefa 6, Ghosh et al. (2019) demonstraram que a utilização do BERT para detecção de posicionamento oferece um desempenho altamente eficaz. Liu et al. (2022) apresentaram um novo modelo de linguagem de grande escala chamado POLITICS ao ajustar o RoBERTa a um conjunto de dados de grande escala composto por artigos de notícias políticas oferecendo um método de propósito geral para DP em conteúdo ideológico.

No entanto, além dessas abordagens serem complexas e difíceis de reproduzir (ALTURAYEIF; LUQMAN; AHMED, 2023), nenhuma atacou diretamente o problema da escassez de dados rotulados. Nesta direção, recentemente uma variação da Transferência de Aprendizado chamada Aprendizado com Poucos Exemplos (*Few-shot Learning*) tem sido empregada em alguns estudos (ALLAWAY; MCKEOWN, 2020; LIU et al., 2021; LUO et al., 2022; ALLAWAY; SRIKANTH; MCKEOWN, 2021; CONFORTI et al., 2021). O objetivo desse paradigma de aprendizado é facilitar a generalização de modelos de PLN em contextos que contam com conjuntos de treinamento restritos, seja por ausência total ou pela presença de apenas alguns exemplos rotulados, utilizando-se de conhecimento MPTs (WANG et al., 2020b).

Allaway e McKeown (2020) criaram um grande *dataset* chamado VAST composto por milhares de tópicos abrangendo diversos domínios. A partir da seleção de poucos exemplos de treinamento a partir de tópicos variados, seu objetivo foi treinar mo-

delos de DP baseado em uma rede neural *feed-forward*. O VAST ainda foi aproveitado nos trabalhos de Liu et al. (2021) e Luo et al. (2022) em uma abordagem de aumento de dados (*data-augumentation*). Conforti et al. (2021) propuseram o uso de um *framework* fracamente supervisionado para DP, que usa dados anotados sinteticamente para melhorar o desempenho em novos alvos. Utilizando dados rotulados conforme a metodologia estabelecida em (CONFORTI et al., 2020), os autores empregaram MLP (Multi Layer Perceptron) para anotar automaticamente um extenso corpus de *tweets*. Esses dados anotados foram usados para treinar modelos de DP usando vetores *tf-idf* derivados do texto. Este estudo específico tem grande similaridade com a abordagem utilizada pelo SSSD. No entanto, obteve um desempenho insatisfatório, registrando apenas 37,6% na métrica Macro-F1 Geral do SemEval-2016 Tarefa 6.

Embora as técnicas de *Few-shot Learning* sejam desenvolvidas para mitigar problemas de escassez de dados rotulados, elas ainda não atingem níveis de desempenho satisfatórios, ficando atrás das abordagens de Transferência de Aprendizado tradicionais, conforme mostrado na Tabela 3.4.

## 3.4 Considerações Finais

O campo de DP em redes sociais tem experimentado um crescimento notável nos últimos anos. Embora as técnicas de DP tenham avançado, ainda existem muitos desafios e oportunidades de pesquisa que devem ser priorizados no futuro. Dentre esses desafios, a escassez de dados rotulados destaca-se como uma preocupação central (ALDAYEL; MAGDY, 2021; ALTURAYEIF; LUQMAN; AHMED, 2023).

As Tabelas 3.1, 3.2, 3.3 e 3.4 resumem respectivamente os métodos Supervisionados, Não-Supervisionados, Fracamente Supervisionados e de Transferência de Aprendizado discutidos na Seção 3.3. Os trabalhos referentes a essas tarefas foram resumidos segundo os seguintes critérios:

- **Features**: Tipo de *features* usadas para modelar a tarefa de DP. As abreviações usadas nesta coluna são LSE: léxicos de sentimentos, LSU: léxicos de subjetividade, LEM (léxicos de emoções), EE (*embeddings* estáticos), EC (*embeddings* contextuais),
- **Modelos**: Modelos usados para DP. Nesta coluna, todas as abreviações referem-se a algoritmos de AM, MPTs ou AP, exceto pela sigla AT, que denota mecanismos de atenção.

- **Datasets**: Nome dos conjuntos de dados utilizados para avaliar os modelos de DP. Dentre os *datasets* é destaque o Semeval 2016, descrito na Seção 3.1, representado pelo sigla SE16-T6. Exceto quando indicado explicitamente, os resultados na coluna Pontuação se referem à tarefa A.
- **Pontuação**: Pontuação Macro-F1 Geral (Seção 5.2) do melhor modelo de ML de cada estudo. Vale observar que alguns estudos não apresentaram seus resultados utilizando essa métrica; nesses casos, exibimos os resultados com base na acurácia sinalizando com a abreviação ACC.

Ao analisar as abordagens de DP apresentadas ao longo da Seção 3.3 e conforme detalhado na Tabela 3.1, fica evidente que as estratégias fundamentadas em AM supervisionada destacam-se por sua consistência em obter resultados superiores. Contudo, a efetividade desses métodos é frequentemente limitada pela necessidade de extensos conjuntos de dados devidamente rotulados, com dependência do alvo do posicionamento e da língua usada na expressão do posicionamento. Isso implica que vários dos métodos de DP que seguem este paradigma dependem de técnicas complexas de engenharia de *features*, o que pode tornar sua reprodução desafiadora (ALTURAYEIF; LUQMAN; AHMED, 2023).

Os métodos que alcançaram os resultados mais expressivos em conformidade com o SemEval-2016 Tarefa 6 foram propostos por Zhao e Yang (2021) (utilizando Transferência de Aprendizado) para a Tarefa A e por Lai et al. (2017) (abordagem Supervisionada) para a Tarefa B. Essas abordagens são reconhecidas como estado da arte no contexto de DP em redes sociais. Elas atingiram um Macro-F1 Geral de 78,4% e 74,5% nas Tarefas A e B, respectivamente, conforme detalhado nas Tabelas 3.4 e 3.1.

Embora os métodos não-supervisionados se apresentem como uma alternativa viável, como demonstrado na Tabela 3.2, necessitam tipicamente de *features* de usuários, cujo o acesso tem sido limitado pelas plataformas. Como uma evolução das técnicas de Técnicas de Transferência de Aprendizado (Tabela 3.3, as abordagens que empregam *Few-shot Learning*, sumarizadas na Tabela 3.4, são bastante promissoras. Contudo, esses métodos ainda necessitam de refinamento para alcançar performances que se equiparem às das estratégias supervisionadas. Uma barreira particular é a dificuldade de reprodução dessas técnicas, já que elas dependem de arquiteturas complexas para obter uma generalização eficaz. Isso implica que, apesar de seu potencial, sua reprodutibilidade é consideravelmente afetada (ALDAYEL; MAGDY, 2021).

O SSSD diferencia-se dos trabalhos relacionados ao integrar de maneira eficiente

MPTs com Busca Semântica. Esta abordagem alcança resultados comparáveis a métodos supervisionados utilizando exclusivamente o conteúdo textual de postagens no Twitter. Além disso, é aplicável a novos domínios e diferentes idiomas, exigindo apenas um conjunto reduzido de dados anotados. Adicionalmente é robusto para tratar problemas de sanidade de dados tipicamente presentes em dados textuais coletados a partir de redes sociais, em particular o Twitter.

Tabela 3.1: Métodos Supervisionados (Ordenado por Pontuação e Datasets)

| Estudo | Features | Modelos | Datasets | Pontuação |
|---|---|---|---|---|
| HaCohen-Kerner, Ido e Ya'akobov (2017) | n-gram, LSE | LibSVM | SE16-T6 | **77.1** |
| Al-Ghadir, Azmi e Hussain (2021) | tf-idf, LSE | KNN | SE16-T6 | 76.4 |
| Gómez-Suta, Echeverry-Correa e Soto-Mejía (2023) | tf-idf, n-gram<br>LSE | SVM, RL<br>AdaBoost | SE16-T6 | 74.6 |
| Lai et al. (2017) | usuário<br>LSE | Gaussian Naive Bayes | SE16-T6 (B) | **74.5** |
| Dey, Shrivastava e Kaushik (2017) | n-gram<br>sintáticas<br>LSE, LSU | SVM | SE16-T6 | 74.4 |
| Chen, Ye e Cui (2021) | n-gram, EC | RoBERTa<br>BiLSTM<br>AT | SE16-T6 | 73.7 |
| Aldayel e Magdy (2019) | usuário<br>n-gram | SVM | SE16-T6 | 71.8 |
| Wei, Mao e Zeng (2018) | EE | BiGRU<br>AT | SE16-T6 | 71.0 |
| Ahmed, Chy e Chowdhury (2020) | estatísticas<br>EE | MLP, CNN<br>BiLSTM<br>Random Forest | SE16-T6 | 70.4 |
| Mohammad, Sobhani e Kiritchenko (2017) | n-gram, LSE<br>EE | SVM | SE16-T6 | 70.3 |
| Siddiqua, Chy e Aono (2018) | sintáticas | SVM | SE16-T6 | 70.0 |
| Sun et al. (2019a) | EE | RNN | SE16-T6 | 69.4 |
| Zhou, Cristea e Shi (2017) | EE | BiGRU, CNN | SE16-T6 | 67.4 |
| Lai et al. (2020) | n-gram, bow<br>LEM | biLSTM<br>SVM, RL<br>CNN, LSTM | SE16-T6<br>IberEval | 64.5 |
| Yang et al. (2020) | EE | BiLSTM<br>AT | NLPCC2016<br>SE16-T6 | 74.1<br>69.2 |
| Sun et al. (2018) | LSE, EE | LSTM<br>AT | SE16-T6 | 61.0 |
| Sobhani, Mohammad e Kiritchenko (2016) | n-gram, LSE | SVM | SE16-T6 | 59.2 |

Continua na próxima página.

Tabela 3.1: Métodos Supervisionados (Ordenado por Pontuação e Datasets)

| Estudo | Features | Modelos | Datasets | Pontuação |
|---|---|---|---|---|
| Zhang et al. (2020) | EE, LSE, LEM | GCN, BiLSTM | SE16-T6 | 53.6 |
| Hosseinia, Dragut e Mukherjee (2020) | LEM, EE, EC | GRU, BERT | Procon20 | 76.9 |
| Chen e Ku (2016) | usuário, EE | CNN, LDA | CreateDebate<br>FBFans | 75.5 |
| Liang et al. (2021) | sintáticas<br>pragmáticas | BiLSTM, GCN<br>AT | WT-WT?<br>SE16-T6 | 74.2<br>59.5 |
| Zhu, He e Zhou (2020) | usuário, EE | LSTM, GRU<br>LDA, AT | US Election-2016<br>Brexit | 72.0<br>65.0 |
| Lai et al. (2020) | estruturais<br>bow, n-gram<br>LSE | SVM | TW-BREXIT | 67.0 |
| Li, Xu e Wang (2019) | EE | CNN, GRU | NLPCC2016 | 62.2 |
| Mohtarami et al. (2018) | EE | CNN, LSTM | FNC-1 | 56.8 |
| Lai et al. (2017) | bow | SVM | IberEval2017<br>ConRef<br>STANCE-ita | 49.0<br>48.8 |
| Pamungkas, Basile e Patti (2019) | estruturais<br>pragmática | SVN | RumourEval-17 | 47.0 |
| Bhatt et al. (2018) | EC, estatísticas<br>estruturais, LSE | MLP, LSTM<br>GRU | FNC-1 | 83.0<br>ACC |
| Borges, Martins e Calado (2019) | EC, EE, LSE<br>estatísticas<br>pragmmáticas | BiLSTM<br>max-pooling<br>AT | FNC-1 | 82.2<br>ACC |

Tabela 3.2: Métodos Não-Supervisionados (Ordenado por Pontuação e Datasets)

| Estudo | Features | Modelos | Datasets | Pontuação |
|---|---|---|---|---|
| Samih e Darwish (2021) | EE, EC<br>usuário | BERT | Tópicos<br>polarizados nos EUA | **92,1**<br>ACC |
| Darwish et al. (2020) | usuário | UMAP, SVM<br>Mean-shift | Tópicos sobre-<br>eleições nos USA | 90,4<br>ACC |
| Rashed et al. (2021) | EC<br>usuário | SVM, MUSE | Eleições Turcas<br>tweets de usuários | 85,0<br>ACC |

Tabela 3.3: Métodos Fracamente-Supervisionados (Ordenado por Pontuação e Datasets)

| Estudo | Features | Modelos | Datasets | Pontuação |
|---|---|---|---|---|
| Wei, Mao e Chen (2019) | EE | BiGRU, SRNet | SE16-T6 | **60.7** |
| Augenstein et al. (2016) | EE, EC | LSTM | SE16-T6 | 58,0 |
| Dias e Becker (2016) | LSE, n-gram | SVM | SE16-T6 (B) | 56,2 |

Tabela 3.4: Métodos de Transferência de Aprendizado (Ordenado por Pontuação e Datasets)

| Estudo | Features | Modelos | Datasets | Pontuação |
|---|---|---|---|---|
| Zhao e Yang (2021) | EC | RoBERTa | SE16-T6 | **78.4** |
| Sun et al. (2022) | EC, LSE | BERT | SE16-T6 Perspectrum | 68.4 |
| Liu et al. (2022) | LSE EC, EN ideologia | RoBERTa, CNN | SE16-T6 VAST, Basil | 67.6 |
| Allaway, Srikanth e McKeown (2021) | EC | 2-Layer-feedforward network | SE16-T6 | 54.1 |
| Xu et al. (2018) | EE, EC | MLP, AT | SE16-T6 | 46.1 |
| Kawintiranon e Singh (2021) | EC | BERT | Eleições EUA | 77.2 |
| Luo et al. (2022) | EC, LSE | BERT | VAST | 72.6 |
| Liu et al. (2021) | EC | BERT Concept-Net | VAST | 70.2 |
| Giorgioni et al. (2020) | EC | UmBERTo | Sardistance | 68.5 |
| Allaway e McKeown (2020) | EC | BERT | VAST | 66.6 |
| Conforti et al. (2021) | tf-idf | MLP | WT-WT? | 37.6 |

# 4 SSSD: SEMANTIC SEARCH STANCE DETECTION

Este capítulo detalha o método SSSD, uma abordagem inovadora de DP que integra MPTs e Busca Semântica para aprimorar a DP no ambiente ruidoso e dinâmico do Twitter.

## 4.1 Visão Geral

Ao explorar eficientemente as vantagens dos MPTs em capturar o conteúdo semântico e contextual dos *tweets*, o SSSD apresenta uma estratégia inovadora para DP no Twitter. Ao combinar MPTs e Busca Semântica, o SSSD usa uma abordagem de *Few-shot Learning* para rotular automaticamente uma grande quantidade de *tweets* a partir de um conjunto inicial reduzido de amostras previamente rotuladas. Posteriormente, os *tweets* rotulados servem como base para treinar modelos de DP. A grande vantagem deste método é a considerável redução no esforço de rotulagem de um extenso volume de *tweets*, sem comprometer a qualidade da classificação.

Além de mitigar os desafios decorrentes da limitação de dados rotulados, o SSSD atua como um filtro ao focar nas postagens de maior relevância, descartando conteúdos ruidosos ou não relacionados aos alvos em questão. Isso representa uma vantagem importante em relação aos métodos de aprendizado de máquina não supervisionados, que podem ter dificuldades especialmente em conjuntos de dados ruidosos e complexos como o Twitter. O fato dos MPTs serem pré-treinados em dados de diversos idiomas possibilita que o SSSD seja aplicado em diferentes domínios, tornando-o agnóstico em relação à linguagem. No restante desta seção descrevemos os dados bem como todos os processos envolvidos no SSSD.

## 4.2 Entradas

O SSSD requer duas entradas principais: (a) *domain-set*, formado por um conjunto extenso de *tweets* relacionados ao domínio de interesse; e (b) o *query-set*, um conjunto menor de *tweets* rotulados que que contém argumentos característicos para expressar posicionamentos. Além disso, um conjunto de teste (test-set) permite avaliar a qualidade dos modelos de DP resultantes. Estes conjuntos de dados são descritos na

sequência.

**a) Domain-Set**

O *domain-set* é composto por *tweets* não rotulados relacionados ao alvo, que serão posteriormente rotulados através do método proposto. Os *tweets* podem ser coletados através da API[1] oficial do Twitter em um procedimento que consiste basicamente em definir e filtrar um período de tempo de interesse, identificando e utilizando como argumento de busca palavras-chave que sejam significativas e representativas no contexto do alvo. Usar *hashtags* pode ser uma estratégia útil, pois elas tendem a capturar a homofilia e influência social relacionadas ao alvo (DARWISH et al., 2020). As *hashtags* relevantes podem ser encontradas na seção de tendências do Twitter e também servem como ponto de partida em um processo iterativo que identifica outras *hashtags* relacionadas com base na co-ocorrência. É crucial definir um período de busca apropriado para evitar viés. Por exemplo, ao detectar posicionamentos sobre os candidatos de uma eleição, o período de busca deve ser cuidadosamente escolhido para representar os posicionamentos no período ao qual a campanha eleitoral aconteceu.

**b) Query-Set**

O *query-set* consiste em um conjunto de *tweets* rotulados manualmente, indicando um posicionamento específico em relação ao alvo, a saber *favorável*, *contrário* ou *nenhum*. Estes funcionam como referências iniciais para a rotulagem automática dos *tweets* no *domain-set*, gerando assim um *training-set* para treinar modelos de DP. A escolha adequada das instâncias do *query-set* permite ao SSSD focar nas postagens mais relevantes, além de agir como filtro, sendo útil para mitigar problemas de ruídos nos *tweets*.

Um aspecto determinante no método proposto é a escolha precisa dos *tweets* que constituirão o *query-set*. Na ausência de dados com rotulados e quando o conhecimento do domínio for limitado, uma abordagem recomendada envolve o uso de métodos de Modelagem de Tópicos, como o BERTopic (GROOTENDORST, 2022) ou o Top2Vec (ANGELOV, 2020). Estes métodos oferecem um panorama detalhado do contexto subjacente ao *corpus*, facilitando a identificação de *tweets* que abordam diferentes posicionamentos. Uma de suas maiores virtudes é a habilidade de capturar através de MPTs a similaridade semântica ao realizar agrupamentos baseados em densidade, fazendo com que os tópicos identificados sejam aglomerados densos de *tweets* representativos de um

[1] https://developer.twitter.com/en/docs/twitter-api

mesmo posicionamento. Adicionalmente, este métodos estão disponíveis em *frameworks* que fornecem recursos visuais e interpretativos que ajudam a explorar e compreender os tópicos, possibilitando a escolha de documentos representativos de cada agrupamento. Por exemplo, Ebeling et al. (2022) e Sousa e Becker (2022b) empregaram o BERTopic para identificar argumentos representativos e tendências políticas em posturas opostas e favoráveis à vacinação COVID.

**c) Test-Set**

Para a avaliação do desempenho dos modelos de DP, emprega-se um conjunto específico para testes, denominado *test-set*. Esse conjunto é fundamental para conduzir uma análise criteriosa do desempenho dos modelos, utilizando-se de métricas estabelecidas, como acurácia ou a medida F (*F-measure*). O *test-set* pode ser um conjunto separado de dados ou, alternativamente, uma parte do próprio *query-set*, uma estratégia viável desde que se imponha um limiar máximo de similaridade no processo de busca semântica. A inclusão deste limiar é uma precaução essencial para mitigar o risco de viés, garantindo que a avaliação do modelo seja justa e representativa, conforme detalhado na Seção 4.3.1.

## 4.3 Semantic Search Stance Detection

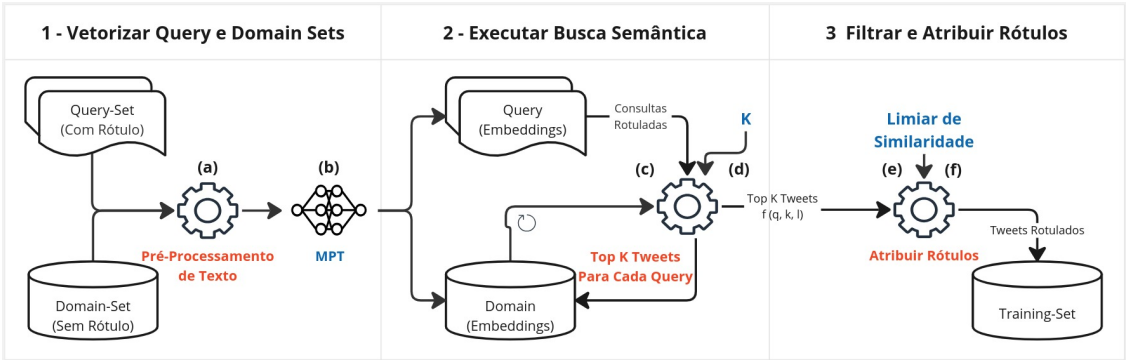
Capturar informações contextuais e nuances na linguagem é fundamental para uma detecção de posicionamento precisa. O SSSD emprega MPTs para converter *tweets* em *embeddings*, visando assimilar o teor semântico do texto e permitir uma comparação e recuperação eficiente de *tweets* similares. O mecanismo central deste processo é definido por uma função que recebe três parâmetros f(q, k, l), na qual $q$ é definido como um par <tweet, posicionamento> do *query-set*, que, quando submetido ao processo de Busca Semântica identifica e rotula automaticamente os $k$ *tweets* do conjunto de domínio obedecendo limiar de similaridade máximo $l$. Esse limiar garante que os *tweets* dos conjuntos de consulta não sejam incluídos nos *tweets* de treinamento rotulados, evitando assim potenciais vieses na avaliação do modelo.

O SSSD é composto por duas fases principais. A primeira, denominada Rotulagem Semântica, emprega MPTs e Busca Semântica para a rotulagem automática de *tweets*. A segunda, Detecção de Posicionamento, foca no treinamento de modelos supervisionados de DP usando os *tweets* rotulados na etapa anterior. Os detalhes de cada fase são explicadas em detalhes no restante desta seção.

### 4.3.1 Rotulagem Semântica

A rotulagem automática dos *tweets* é realizada para criar um *training-set* a partir dos conjuntos *domain-set* e *query-set*. O resultado dessa fase é um conjunto de dados rotulados, pronto para a etapa seguinte de treinamento. Conforme demonstrado na Figura 4.1, essa fase se desdobra em três passos específicos:

Figura 4.1: SSSD - Rotulagem Semântica




1. **Vetorizar Query e Domain-Sets**: Inicialmente os *query-sets* e *domain-sets* são submetidos a um pré-processamento de texto (a), e posteriormente convertidos em *embeddings* mediante um MPT, como BERT (b).

2. **Executar Busca Semântica**: Após obter os *embeddings*, uma função de Busca Semântica (c) é utilizada para comparar cada elemento $q$ do *query-set* com os *tweets* do *domain-set*. Essa comparação é realizada calculando-se as pontuações de similaridade entre os *embeddings* da consulta $q$ e os *embeddings* dos *tweets* do conjunto de domínio. A pontuação de similaridade pode ser calculada usando vários métodos, como a similaridade de cosseno ou produto escalar. Em seguida, usando a entrada $k$, os top-k *tweets* com as pontuações mais altas são selecionados (d).

3. **Filtrar e Atribuir Rótulos**: Os top-k *tweets* selecionados recebem o rótulo de posicionamento associado ao $q$ (e). É possível que um mesmo *tweets* apareça tanto nos *query-sets* quanto nos *domain-sets*. Para prevenir potenciais vieses, especialmente se partes dos *query-sets* são utilizadas para validação, sugere-se estabelecer um limiar de similaridade (f) inferior a 1 (e.g., 0,9) ou diretamente excluir tais *tweets* do *domain-set*. Posteriormente, os *tweets* rotulados são adicionados ao conjunto de treinamento. Também é possível que determinado *tweet* do *domain-set* seja semelhante a diferentes consultas do conjunto de consultas. Se ocorrerem empates,

pode-se selecionar o rótulo associado à pontuação de similaridade mais alta ou mais frequente. Vale ressaltar que quanto maior o valor de *k*, maior a probabilidade de empates. Portanto, é aconselhável escolher um valor adequado para *k* a fim de minimizar os empates e garantir resultados de rotulagem mais consistentes.

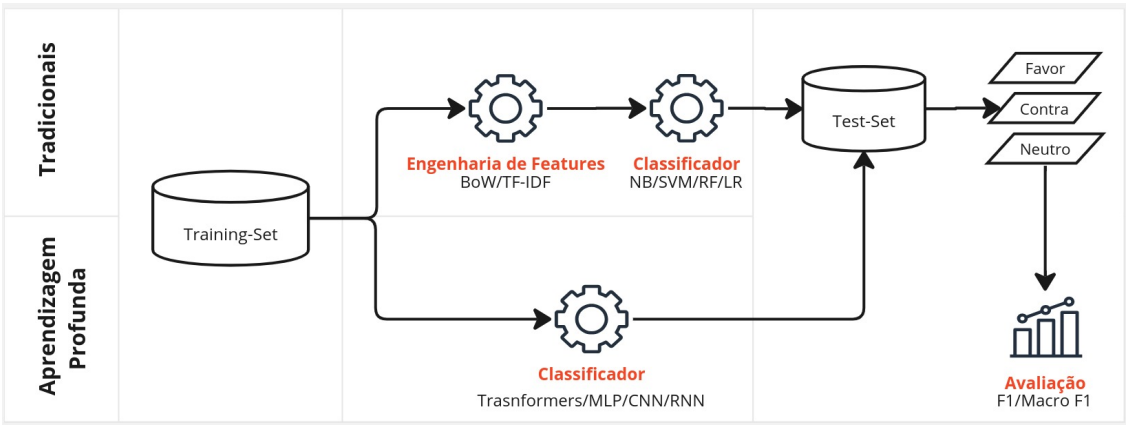
Este método otimiza a rotulagem de *tweets* com posicionamentos alinhados aos do conjunto de consultas, otimizando a DP no Twitter. A quantidade de *tweets* rotulados no conjunto de treinamento é diretamente impactada pelo valor de *k* e pelo tamanho do conjunto de consultas. Um *k* elevado, amplia o número de *tweets* rotulados, mas causa uma degradação nas pontuações de similaridade. Assim, é essencial equilibrar o valor de *k* para manter a robustez das pontuações. Adicionalmente, o tamanhos dos conjuntos de domínio determinam o total de *tweets* que podem ser rotulados. Conjuntos de domínio menores naturalmente resultam em menores quantidades de *tweets* rotulados. Além disso, a seleção e rotulagem criteriosa dos *tweets* para o *query-set* habilita o SSSD a filtrar informações irrelevantes e focar a DP exclusivamente no contexto desejado.

### 4.3.2 Detecção de Posicionamentos

O processo descrito acima é eficaz para DP, mas possui limitações. Aumentar *k* pode ampliar a cobertura dos dados rotulados, mas também aumenta o risco de mais classificações incorretas devido à degradação das pontuações de similaridade. É recomendável treinar modelos de classificação usando dados rotulados gerados na etapa anterior para aprimorar a precisão e a generalização. Em seguida, os *tweets* não rotulados restantes do conjunto de domínio podem receber um rótulo usando este modelo.

Existem vários modelos de AM supervisionados adequados para essa tarefa, incluindo os tradicionais como Regressão Logística, Árvores de Decisão, Máquinas de Vetores de Suporte ou os de aprendizagem profunda como RNNs, CNNs e LSTMs. A seleção do modelo e da estratégia de engenharia de *features* é condicionada pela natureza da tarefa, pela composição do conjunto de dados e pela capacidade computacional disponível. Para abordagens tradicionais, métodos como *tf-idf* e *n-grams* são frequentemente empregados. Por outro lado, em AP, MPTs como BERT podem ser implementados, eliminando a necessidade de extensiva engenharia de *features*. A eficácia do modelo de DP pode ser mensurada com base no conjunto de teste (*test-set*), conforme ilustrado na Figura 4.2.

Figura 4.2: SSSD - Detecção de Posicionamentos



# 5 ANÁLISE QUANTITATIVA

Neste capítulo, é apresentada uma análise quantitativa do SSSD, fundamentada nas métricas e alvos especificados para as Tarefas A e B do SemEval-2016 Tarefa 6.

## 5.1 Questões de Pesquisa

Os experimentos conduzidos nesta análise têm com o objetivo responder às seguintes questões de pesquisa:

- **QP1:** O quão competitivo é o desempenho do SSSD quando comparado a sistemas referência em DP baseados no SemEval-2016 Tarefa 6?
- **QP2:** Qual é o impacto do parâmetro *k* no desempenho, considerando todos os alvos e os diferentes modelos de AM adotados?

## 5.2 SemEval-2016 Tarefa 6: Conjunto de Dados e Métricas

A Tarefa A do SemEval-2016 Tarefa 6 engloba cinco alvos distintos: "Ateísmo (Ate)", "Mudanças climáticas é uma preocupação real (Mdc)", "Feminismo (Fmn)", "Aborto (Abt)" e "Hillary Clinton (Hlr)". O conjunto de treinamento para a Tarefa A reúne 2.914 *tweets*, e o conjunto de teste conta com 1.246 *tweets* ambos rotulados. Já para a Tarefa B, focada no alvo "Donald Trump (Trp)", incentivou-se a adoção de técnicas não supervisionadas , por isso, foram disponibilizados apenas 707 *tweets* para teste e nenhum conjunto rotulado para treino. Um resumo dos dados descritos por Mohammad et al. (2016a) disponibilizados para as duas tarefas pode ser visto na Tabela 5.1.

Tabela 5.1: Distribuição dos dados de Treino e Teste para as Tarefas A e B

| **Alvos** | **% Tweets de Treino** | | | | **% Tweets de Teste** | | | |
|---|---|---|---|---|---|---|---|---|
| | # Tweets | % Favor | % Contra | % Nenhum | # Tweets | % Favor | % Contra | % Nenhum |
| | | | | **Tarefa A** | | | | |
| Ate | 513 | 17,9 | 59,3 | 22,8 | 220 | 14,5 | 72,7 | 12,7 |
| Abt | 653 | 18,5 | 54,4 | 27,1 | 280 | 16,4 | 67,5 | 16,1 |
| Fmn | 664 | 31,6 | 49,4 | 19,0 | 285 | 15,3 | 58,3 | 26,4 |
| Hlr | 689 | 17.1 | 57,0 | 25,8 | 295 | 20,4 | 64,2 | 15.4 |
| Mdc | 395 | 53,7 | 3,8 | 42,5 | 169 | 72,8 | 6,5 | 20,7 |
| **Total** | 2.914 | | | | 1.246 | | | |
| | | | | **Tarefa B** | | | | |
| Trp | - | - | - | - | 707 | 20,9 | 42,2 | 36,7 |

A métrica de avaliação oficial utilizada na competição é baseada no Macro-F1 Médio (*Fmedio*) do desempenho das classes "Favor" (*Ffavor*) e "Contra" (*Fcontra*), conforme mostrado na Fórmula 5.1.

$$F_{\text{medio}} = \frac{F_{\text{favor}} + F_{\text{contra}}}{2} \tag{5.1}$$

Para calcular *Ffavor*, são utilizadas a precisão (*Pfavor*) e a revocação (*Rfavor*) da classe "Favor", conforme a Fórmula 5.2.

$$F_{\text{favor}} = \frac{2 \cdot P_{\text{favor}} \cdot R_{\text{favor}}}{P_{\text{favor}} + R_{\text{favor}}} \tag{5.2}$$

Da mesma forma, *Fcontra* é calculado usando a precisão (*Pcontra*) e a revocação (*Rcontra*) da classe "Contra", conforme a Fórmula 5.3:

$$F_{\text{contra}} = \frac{2 \cdot P_{\text{contra}} \cdot R_{\text{contra}}}{P_{\text{contra}} + R_{\text{contra}}} \tag{5.3}$$

É importante ressaltar que essa métrica de avaliação não desconsidera a classe "Nenhum". No entanto, o foco de comparação está nas classes "Favor" e "Contra", tratando "Nenhum" como classe de não interesse ou classe negativa no contexto da Recuperação de Informação. A classificação incorreta de instâncias da classe negativa pode afetar as pontuações dessa métrica. Além da métrica

Além do *Fmedio*, o SemEval-2016 Tarefa 6 também forneceu a métrica Macro-F1 Geral (*Fgeral*), determinada considerando os resultados de precisão e revocação referentes a todos os alvos combinados, conforme o mostrado nas Fórmulas 5.4 e 5.5.

$$\text{Precisão} = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} (VP_i + FP_i)} \tag{5.4}$$

$$\text{Revocação} = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} (VP_i + FN_i)} \tag{5.5}$$

Nestas fórmulas, *VP*, *FP*, e *FN* representam Verdadeiros Positivos, Falsos Positivos, e Falsos Negativos, respectivamente, para o alvo *i*. *n* é o número de alvos considerados. Isso dá uma visão mais clara de que as pontuações de precisão e revocação de cada alvo são somadas antes de calcular a média. As medidas *Favor* e *Fcontra* individuais para cada classe ("Favor" e "Contra"), são calculadas conforme as Fórmulas 5.2 e 5.3. O *Fgeral* segue a Fórmula 5.1.

## 5.3 Recursos e Configurações

Conforme explicado no Capítulo 4, o SSSD demanda duas entradas de dados principais: um *domain-set*, composto por uma vasta coleção de *tweets* não rotulados, e um *query-set*, formado por um conjunto menor de *tweets* previamente rotulados. Adicionalmente, a etapa de Rotulagem Semântica recebe um MPT para geração de *embeddings*, um parâmetro *k* referente ao número de *tweets* selecionados com base na máxima similaridade de cosseno em relação a uma consulta específica. Há também a definição de um limiar de similaridade para eliminar potenciais vieses. Por fim, algoritmos de classificação específicos são empregados na fase de Detecção de Posicionamentos. A configuração precisa desses elementos nos experimentos, bem como o desenvolvimentos dos modelos de DP segundo o SSSD são detalhadas na sequência.

Tabela 5.2: Resumo dos Tweets por Amostra e Alvo.

| **Amostras** | **Ate** | **Abt** | **Mdc** | **Fmn** | **Hlr** | **Trp** | **Total** |
|---|---|---|---|---|---|---|---|
| Query-Set | 513 | 653 | 395 | 664 | 689 | 707 | **3.621** |
| Test-Set | 220 | 280 | 169 | 285 | 295 | - | **1.249** |
| Domain-Set | 688.854 | 225.889 | 249.656 | 121.049 | 1.481,868 | 598.991 | **3.366,307** |

- **Domain-Set:** os *domain-sets* para cada alvo foram coletados utilizando-se da API do Twitter, conforme as diretrizes recomendadas na Seção 4.2. Foram utilizados os mesmos parâmetros utilizados para a coleta de dados na competição SemEval[1] em termos de período de coleta (1º de janeiro a 31 de dezembro de 2016) e *hashtags*.

- **Query-Set**: para Tarefa A, os conjuntos de consulta e teste foram combinados. Já para a Tarefa B, foram utilizados exclusivamente o conjunto de testes, visto que não foram disponibilizados *tweets* para treinamento. A avaliação do desempenho dos modelos, em relação a todos os alvos, foi feita através dos respectivos conjuntos de testes. Uma visão geral da distribuição dos *tweets* nos diferentes conjuntos pode ser encontrada na Tabela 5.2.

  Todos os *tweets* foram submetidos a uma etapa de pré-processamento padrão, que incluiu a remoção de *tweets* com menos de três termos, *retweets*, menções, entradas duplicadas, pontuações, caracteres especiais, *hashtags* e URLs.

- **MPT**: uma decisão crucial no processo foi a escolha do MPT para transformar os *domain-set* e *query-set* em *embeddings* no processo de Rotulagem Semântica (Se-

[1] www.saifmohammad.com/WebPages/StanceDataset.html

ção 4.3.1). Neste caso, optou-se pelo modelo "all-MiniLM-L6-v2"(WANG et al., 2020a). Esta escolha foi influenciada por sua eficácia em tarefas de Busca Semântica. Embora apresente qualidade comparável a modelos sofisticados, como o Microsoft MPNET (AHMED; CHY; CHOWDHURY, 2020), destaca-se por oferecer um desempenho consideravelmente mais ágil.

- **Limiar de Similaridade**: Adotou-se cosseno como medida de similaridade. Definiu-se um limiar máximo de similaridade de 0.95, descartando-se assim quaisquer *tweets* com cossenos muito próximos ou idênticos entre os *domain-sets* e os conjuntos de consulta/teste. Também foram removidos dos *domain-set* todos os *tweets* em comum entre os conjuntos. Esse cuidado visa garantir a integridade e imparcialidade do processo de treinamento e avaliação dos modelos.

- **Algoritmos de Classificação e Engenharia de Features**: para a etapa de Detecção de Posicionamentos (Seção 4.3.2), foram avaliados algoritmos de Regressão Logística (modelos SSSD-RL), Máquinas de Vetores de Suporte (modelos SSSD-SVM) e Floresta Aleatória (modelos SSSD-RF), a fim de determinar possíveis variações de desempenho entre eles. Em relação a técnicas de engenharia de *features*, foram empregados *tf-idf* e *bigrams*.

- **Valor de $k$ e Modelos de DP**: considerando as configurações acima descritas, foram realizadas 20 iterações de treino, ajustando o valor de $k$ em incrementos de 5 para cada um dos alvos (6) e algoritmos de classificação (6). Isso resultou na geração de 60 modelos distintos por alvo, totalizando 360 modelos ao final dos experimentos. Os melhores resultados obtidos para cada alvo e algoritmo de classificação estão sintetizados nas Tabelas 5.3 e 5.4, juntamente com o respectivo valor de $k$.

Todos os dados e código dos experimentos realizados estão acessíveis em um repositório público[2].

## 5.4 Experimento 1: Desempenho

Neste experimento, o SSSD foi comparado com os métodos propostos por Zhao e Yang (2021) para a Tarefa A e por Lai et al. (2017) para a Tarefa B. Esses métodos são considerados estado da arte no âmbito DP em redes sociais para as tarefas abordadas,

[2]https://github.com/mediote/sssd

alcançando um Macro-F1 Geral de 78,4% e 74,5% nas Tarefas A e B respectivamente, conforme apresentado nas Tabelas 3.4 e 3.1.

Os resultados da pontuação Macro-F1 Geral mostrados na Tabela 5.3 evidenciam que o SSSD, em particular o SSSD-RL, exibe um desempenho notavelmente superior em comparação ao *baseline* em todos os alvos da Tarefa A. Isso se reflete em uma melhoria de 12 pp (pontos percentuais) ao atingir um valor de 90,4%. De maneira semelhante, SSSD-SVM alcançou Fmédio de 89,5%, superando o *baseline* em 11,1 pp. O SSSD-RF, embora com um desempenho ligeiramente inferior em comparação com o SSSD-RL e o SSSD-SVM, ainda conseguiu superar o *baseline* em 7,9 pp. Ao focar em alvos individuais, as diferenças de desempenho também foram notáveis. Por exemplo, o modelo SSSD-LR mostrou melhorias que variam de 1,4 pp em relação o alvo "Hlr" a 36,1 pp para o alvo "Mdc".

Tabela 5.3: Resultados para os Alvos da Tarefa A

| **Sistemas** | **Geral** | | | **Ate** | **Abt** | **Mdc** | **Fmn** | **Hlr** |
|---|---|---|---|---|---|---|---|---|
| | % Ffavor | % Fcontra | % Fgeral | % Fmedio | | | | |
| **Baseline** | | | | | | | | |
| Zhao e Yang (2021) | 75,7 | 81,6 | 78,4 | 77,8 | 71,7 | 53,2 | 74,1 | 73,2 |
| **SSSD** | | | | | | | | |
| SSSD-RL | **87,3** | **93,5** | **90,4** | **89,1**$^{75}$ | **82,0**$^{75}$ | **89,3**$^{55}$ | **78,5**$^{40}$ | 80,1$^{55}$ |
| SSSD-SVM | 86,3 | 92,7 | 89,5 | 88,5$^{80}$ | 80,0$^{20}$ | 88,2$^{80}$ | 77,2$^{20}$ | **81,2**$^{85}$ |
| SSSD-RF | 80,0 | 87,8 | 84,3 | 80,0$^{85}$ | 74,9$^{80}$ | 79,6$^{35}$ | 70,1$^{80}$ | 71,5$^{70}$ |

Tabela 5.4: Resultados para os Alvos da Tarefa B

| **Sistemas** | **Geral** | | | **Trp** |
|---|---|---|---|---|
| | % Ffavor | % Fcontra | % Fmedio | % Favg |
| **Baseline** | | | | |
| Lai et al. (2017) | 79,7 | 69,2 | 74,5 | - |
| **SSSD** | | | | |
| SSSD-RL | 87,4 | 93,2 | 90,3 | 84,7$^{85}$ |
| SSSD-SVM | **88,0** | **93,2** | **90,6** | **85,2**$^{40}$ |
| SSSD-RF | 80,6 | 86,3 | 83,4 | 75,1$^{65}$ |

Nota-se na Tabela 5.3 que o alvo "Clc" no *baseline* tem um desempenho bem inferior aos demais alvos (53,2%), o que parece estar associado ao desbalanceamento de classes, conforme evidenciado na Tabela 5.1. O menor volume de *tweets* na classe "Contra" sugere que o método de DP proposto por Zhao e Yang (2021) tem dificuldades em identificar adequadamente os posicionamentos quando as classes são desbalanceadas. Isso destaca a robustez do SSSD em alvos que apresentam essa característica, já que para

a mesma classe o SSSD alcançou desempenho entre 79,6 e 89,3, e comparável aos demais alvos.

A Tabela 5.4 exibe os resultados dos sistemas voltados à Tarefa B, com foco particular no alvo "Trp". No que diz respeito à pontuação Macro-F1 Geral, o modelo SSSD-SVM destacou-se significativamente, apresentando uma superioridade em relação ao *baseline* por uma margem de 16,1 pp. Esta melhoria pode ser atribuída a avanços tanto na métrica $F favor$ (melhoria de 8,3 pp) quanto na Fcontra (aumento expressivo de 24 pp). Por outro lado, o SSSD-RF, mesmo sendo o modelo com desempenho menos destacado entre as variantes SSSD, ainda assim conseguiu melhorias de 7,9 pp em relação o *baseline*. Esse padrão ressalta a robustez e eficácia global do SSSD em comparação às abordagens tradicionais representadas pelo *baseline*.

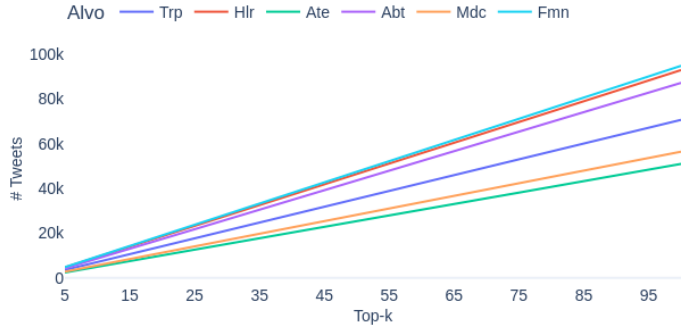
Ao explorar eficientemente as vantagens dos MPTs associados à Busca Semântica, o SSSD demonstrou melhorias promissoras nas Tarefas A e B do SemEval-2016 Tarefa 6, posicionando-se como referência em DP. A melhoria de 12 pp em relação ao *baseline* proposto por Zhao e Yang (2021) evidencia a competência do SSSD em identificar posicionamentos com precisão de maneira equilibrada além de maximizar a utilização dos *tweets*. Da mesma forma, na Tarefa B, os resultados superaram o *baseline* estabelecido por (LAI et al., 2017) por uma margem de aproximadamente 16,1 pp, enfatizando a capacidade do SSSD de detectar posicionamentos de forma eficaz além de possuir um alto grau de generalização.

## 5.5 Experimento 2: Influência do Parâmetro k

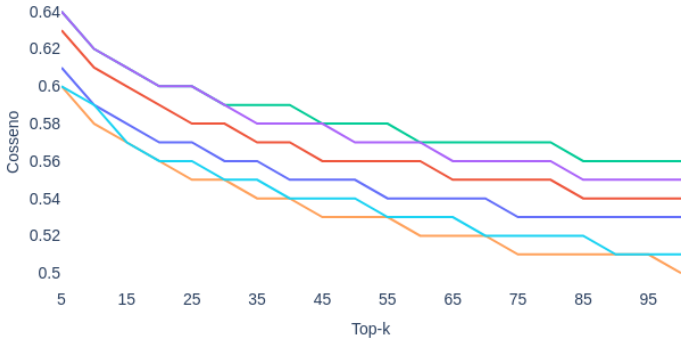
O parâmetro $k$ desempenha um papel fundamental no desempenho do SSSD. Nesse experimento, avaliou-se o impacto de $k$ sob três perspectivas: o volume de *tweets* rotulados, as pontuações de similaridade dos *tweets* escolhidos e o desempenho da classificação.

A Figura 5.1 demonstra a relação entre $k$, o número de *tweets* rotulados e as respectivas pontuações de similaridade. Em 5.1.(a), é evidente um aumento linear no volume de *tweets* rotulados conforme $k$ é incrementado. Nota-se que, para todos os alvos, uma considerável quantia de *tweets* é rotulada mesmo quando $k$ é relativamente pequeno (como por exemplo, aproximadamente 20 mil *tweets* para k = 25). Em contrapartida, a Figura 5.1.(b) ilustra o decréscimo progressivo das pontuações de similaridade com o aumento de $k$.

Figura 5.1: Relação entre $k$, Tweets rotulados e Cosseno.




(a) $k$ e Tweets rotulados




(b) $k$ e Cosseno

A Figura 5.2 evidencia uma correlação consistente entre *Fmedio* e $k$ para todos os alvos e algoritmos de classificação avaliados. Conforme $k$ cresce, o *Fmedio* ascende até alcançar um platô, onde um aglomerado de valores similares de *Fmedio* pode ser identificado. Contudo, quando $k$ se aproxima de valores ao redor de 100, frequentemente observa-se uma regressão nos valores *Fmedio*, sinalizando uma queda na performance. Esta tendência é notadamente marcante nos alvos “Donald Trump", “Ateísmo"e “Hillary Clinton", o que está alinhado com as observações feitas na Figura 5.1.(b).

A obtenção dos melhores desempenhos frequentemente ocorreu com um valor de $k = 60$, conforme evidenciado pelos resultados. No entanto, fixar um valor único de $k$ para todas as circunstâncias não é aconselhável. A análise detalhada dos dados apresentados nas Tabelas 5.3 e 5.4 revela que o $k$ ótimo, que proporciona um equilíbrio ideal entre $k$ e *Fmedio*, varia conforme o alvo e o algoritmo de classificação empregado. Esta nuance é corroborada pela Figura 5.3, que sintetiza as observações chave discutidas. Valores mais elevados de $k$ têm um efeito benéfico sobre a quantidade de *tweets* rotulados, mas prejudicam a similaridade, exercendo uma influência menor sobre o *Fmedio*. Além disso, foi constatada uma correlação negativa entre altas pontuações de similaridade e *Fmedio* e o volume de *tweets* rotulados. Este fenômeno sublinha a necessidade de encontrar um

equilíbrio adequado entre essas variáveis a fim de otimizar o desempenho de classificação.

Figura 5.2: Relação entre *k* e Fmédio

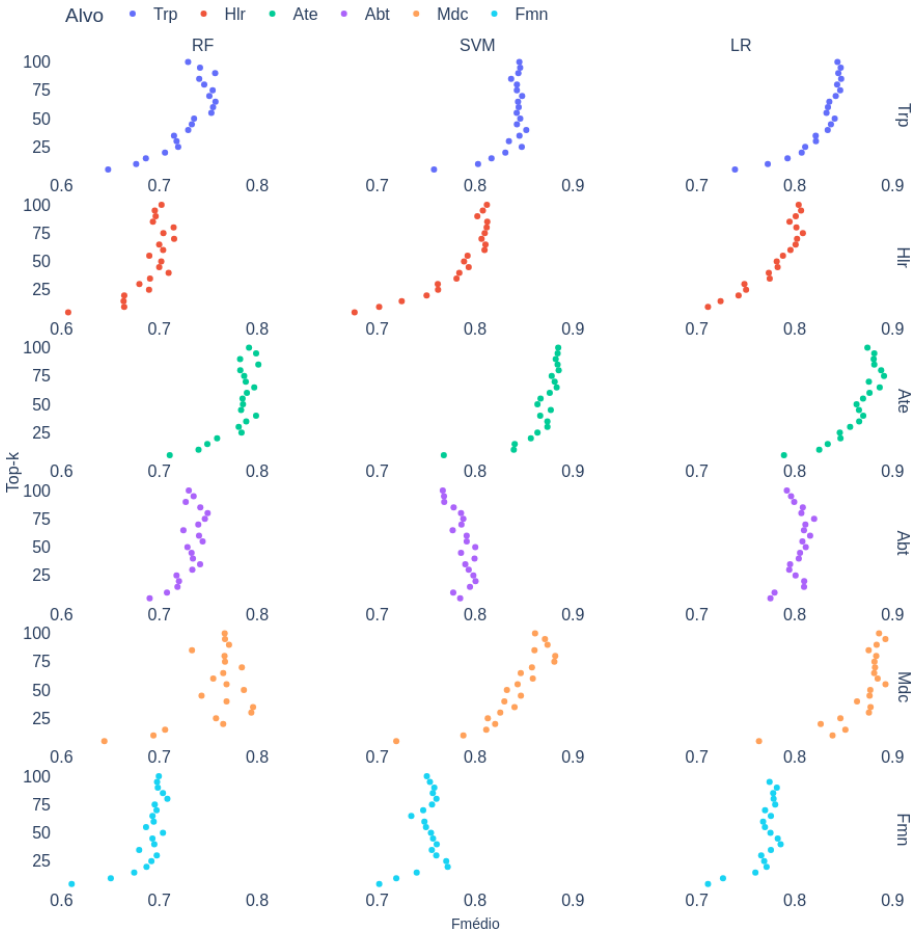
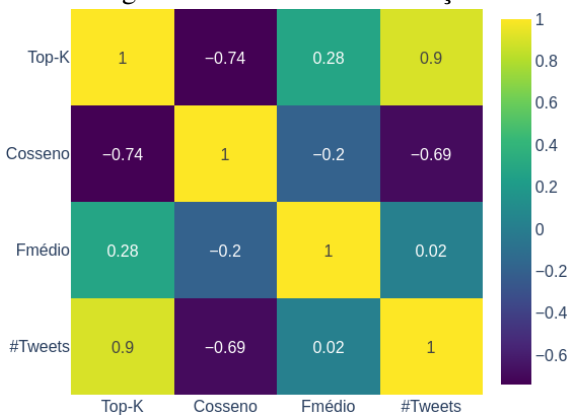



Figura 5.3: Matriz de Correlação



# 6 ANÁLISE QUALITATIVA

Neste capítulo é apresentada uma avaliação qualitativa do SSSD, utilizando como base os conjunto de dados sobre o alvo “Vacinação” no contexto da pandemia de COVID-19 no Brasil. O capítulo relata o estudo de caso usado, a aplicação do SSSD nestes dados, seu desempenho, e sua habilidade de tratar vieses nos dados resultantes do processo de coleta.

## 6.1 Estudo de Caso: Entendendo os Posicionamentos sobre a Vacinação COVID-19 no Brasil

Em trabalhos anteriores do grupo de pesquisa (SOUSA; BECKER, 2021; SOUSA; BECKER, 2022b) realizamos uma análise para entender os argumentos apresentados pelos brasileiros no Twitter, em relação a posicionamentos pró e contra a vacinação da COVID-19. A análise foi baseada em uma amostra de 159.215 *tweets* coletados via API do Twitter de fevereiro de 2020 a maio de 2021. Este intervalo engloba desde o início da pandemia até o momento em que 35 milhões de brasileiros receberam, pelo menos, a primeira dose da vacina. O estudo iniciou com uma análise temporal sobre as evoluções dos posicionamentos (SOUSA; BECKER, 2021), foi posteriormente detalhada em nível de argumentos com o BERTopic (SOUSA; BECKER, 2022b), e finalmente comparada ao posicionamento das pessoas nos Estados Unidos na América (SOUSA; BECKER, 2022a).

O estudo de caso desenvolvido nesta dissertação foca principalmente nos resultados da compreensão dos dados obtidos em nível de argumento detalhada em (SOUSA; BECKER, 2022b).

### 6.1.1 Metodologia de Coleta de Dados

O processo de coleta de dados realizado em (SOUSA; BECKER, 2022b) teve como primeiro passo a identificação de hashtags relevantes para a caracterização dos posicionamentos em relação à vacinação COVID. Inicialmente coletamos através da API do Twitter cerca de 6 milhões de *tweets* contendo os termos "vacina"e/ou "vacinação", considerando o período de 29 de fevereiro de 2020 e 3 de maio de 2021. Para categorizar os *tweets* de acordo com o posicionamento sobre a vacinação, aplicaram-se expressões regu-

lares e inspeção manual das *hashtags*, selecionando aquelas que expressavam claramente opiniões a favor ou contra a vacina. Dentre estas, foram escolhidas as cinco *hashtags* mais frequentes que representavam cada posicionamento. Finalmente, realizou-se uma nova coleta de *tweets* com base nas *hashtags* de posicionamento escolhidas. Para cada posicionamento, a Tabela 6.1 apresenta as hashtags de coleta. As colunas agrupadas sob o rótulo *Bruto* apresentam o número de *tweets* e de usuários únicos coletados do Twitter com as hashtags escolhidas.

Tabela 6.1: Resumo dos Domain-set por Posicionamento

| **Posicionamento** | **Hashtags** | **Bruto** | | **Pre-Processado** | |
|---|---|---|---|---|---|
| | | # Tweets | # Usuários | # Tweets | # Usuários |
| Pro-vaxxers | todospelasvacinas, vacinaja, vemvacina, vacinaparatodos, vacinasim | 216.377 | 76.316 | 139.131 | 55.695 |
| Anti-vaxxers | eunaovoutomarvacina, vacinaorbrigatorianao, vacinanao, naovoutomarvacina, vacinaorbrigatorianunca | 39.073 | 16.675 | 20.084 | 10.244 |
| **Total** | | 255.450 | 92.991 | **159.215** | **65.939** |

Estes *tweets* foram então pré-processados. Excluímos *tweets* com menos de três termos, *retweets*, duplicados, pontuações, caracteres especiais, *hashtags* e URLs. A contagem atualizada de *tweets* e de usuários únicos, após esse pré-processamento, pode ser conferida na Tabela 6.1 (colunas *Pré-Processado*). O conjunto de dados final utilizado na análise foi composto por 159.215 *tweets* de 65.939 usuários distintos.

### 6.1.2 Interpretação de Posicionamentos Baseada em Modelagem de Tópicos

Para identificar os principais tópicos e entender os argumentos comumente usados para defender cada posicionamento, empregou-se o BERTopic sobre os *tweets* pré-processados. Como este método usa um algoritmo de agrupamento baseado em densidade, resulta tipicamente em um grande número de clusters que dificulta a interpretação. Assim, em (SOUSA; BECKER, 2022b) exploramos recursos oferecidos pelo BERTopic para chegar a um conjunto coerente e interpretável de tópicos, tais como: a) agrupamento dos tópicos mais similares; b) visualização gráfica dos tópicos em um plano bidimensional baseado em similaridade; c) análise dos tópicos considerando termos com altas pontuações de c-TF-IDF (unigramas e bigramas) e d) exame dos documentos centrais de cada tópico. Como resultado, chegou-se a um total de 10 tópicos para cada posicionamento.

A Tabela 6.2 sumariza os posicionamentos dos apoiadores da vacinação (“Pro-Vaxxers”) e dos contrários (“Anti-Vaxxers”). Esta tabela foi elaborada para efeitos comparativos com a população dos EUA realizada em (SOUSA; BECKER, 2022a), e resume os macro-temas dos argumentos à favor e contra, junto com o número de tópicos nas quais estas ideias são defendidas. No Anexo A detalhamos nas tabelas A.1 e A.2 todos os 10 tópicos e os argumentos representativos dos Pro-Vaxxers/Anti-Vaxxers, respectivamente. Maiores detalhes sobre o estudo de caso e seus achados podem ser encontrados em (SOUSA; BECKER, 2022b).

Tabela 6.2: Resumo dos Posicionamentos Pro/Anti-vax no Brasil

| Posicionamento | # Tópicos | Argumentos |
|---|---|---|
| Pro-Vaxxers | 4 | Expectativa e celebração pela aprovação da vacina, alegria por ser vacinado, defesa e celebração da ciência e do SUS |
| | 2 | Apoio à vacinação COVID (com **engajamento artificial**). |
| | 2 | Criticas ao Presidente e ao governo pela indisponibilidade de vacinas. |
| | 2 | Discussões sobre vacinação para retorno às aulas. |
| Anti-Vaxxers | 3 | Contra a vacinação ou sua obrigatoriedade (com **engajamento artificial**). |
| | 4 | Contra a “vacina chinesa”; criticas a João Dória, aos políticos em geral, e ao STF. |
| | 3 | Ironia aos “anti-vaxxers” e criticas ao uso de *hashtags* para promover movimento “Anti-Vaxxers” (com **falso-negativos**). |

Em resumo, os “Pro-Vaxxers” manifestam entusiasmo e expectativa pela chance de vacinação, louvam os avanços da ciência e o papel do SUS, ao mesmo tempo em que fazem severas críticas ao Presidente e seu governo pela insuficiência de vacinas. A vacinação como pré-requisito para o retorno às aulas presenciais também é um ponto de destaque. Contudo, em dois dos tópicos foram identificados problemas ligados ao uso de engajamento artificial. Neste tópico, um número significativo de *tweets* usava músicas ou personalidades para alavancar manifestar o posicionamento pró-vacinação, sem um conteúdo específico.

Por outro lado, os “Anti-Vaxxers” propagaram ideias contrárias à vacinação. Há uma expressiva insatisfação com a potencial obrigatoriedade da vacina. Eles também dirigem críticas e apelos a governadores e políticos, opondo-se à decisão do STF que declarou legítima tal obrigatoriedade. Ressalta-se ainda a oposição específica à chamada ”vacina chinesa”, frequentemente associada ao governador João Dória. É importante destacar que neste grupo, foram identificados seis tópicos com problemas. Em três dos tópicos boa parte dos *tweets* são na verdade falso-positivos, onde os “Pro-Vaxxers” zombam e repreendem aqueles resistentes à vacinação, usando menções em postagens anti-vacinas que continham originalmente *hashtags* de coleta relacionadas ao movimento “Anti-Vaxxers”.

Além disso, observa-se em outros três tópicos a ampla utilização de engajamento artificial como meio de difundir seus posicionamentos.

Esta análise nos permitiu identificar vieses nos dados coletados. Ambos os grupos empregaram estratégias de engajamento artificial usando nomes de celebridades, jogadores de futebol, times, músicas ou eventos para promover suas respectivas perspectivas, sem agregar ideias próprias. Os falsos negativos, presentes nos "Anti-Vaxxers", são carregados de ironia e sarcasmo, tornando desafiadora a tarefa de DP neste cenário. Ainda, existem diferenças sutis na expressão de posicionamentos distintos, como por exemplo, os *tweets* "Não tome vacina, é bom que sobra!" e "Não vou tomar vacina, assim sobra mais para vocês" que apresentam semelhanças sutis. Exemplos dos vieses mencionados são listados na Tabela 6.3.

Tabela 6.3: Exemplos de Vieses por Posicionamento

| **Viés** | **Tweet** |
|---|---|
| Levantamento de Hashtags (Pro-Vaxxers) | Arão Ceni Liverpool #TodosPelasVacinas Dembele Origi Burnley VAMOS FLAMENGO.<br>Poderia tocar LIFE GOES ON do BTS_twt na programação? #TodosPelasVacinas.<br>#AprendiNoEnem #FantasticoLomba Andressa Messi Barcelona Yuri Alberto #VemVacina. |
| Falsos Positivos (Anti-Vaxxers) | Ao povo da #VacinaNao: Não tome vacina é bom que sobra!<br>Que os bolsominion não tomem vacina porque sobra menos deles #VacinaNao.<br>Gente que tag ridícula é essa: #NaoVoutomarVacina? |
| Levantamento de Hashtags (Anti-Vaxxers) | #AFazenda12 #VacinaObrigatoriaNao #HappyBirthdayJimin Conan MIRELLA MERECE RESPEITO chan POSITIONS IS COMING All About Luv #followtrick ariana Josh VERBO<br>#EuNaoVouTomarVacina Gandalf Sweet Melody Yuri CONTRA TUDO E TODOS Peixe #LOUIES 5SOS IS COMING Luiz Fernando Diego Souza Atlético Copa do Brasil Igor Gomes Rabello Everson Thaciano<br>#VacinaObrigatoriaNao Liverpool Adrian Everton Pickford Van Dijk André Gomes Bwipo Rodrigo Hilbert James Rodrigues. |

Nota-se neste estudo de caso que o método de coleta baseado em *hashtags*, bastante popular neste tipo de problema, resulta em um conjuntos de dados com ruídos, e que não pode ser usado de forma direta como conjunto de treino para um classificador. Em (SÁENZ; BECKER, 2021) foram desenvolvidos experimentos de classificação deste conjunto de dados com diferentes algoritmos, sendo que o desempenho máximo alcançado não ultrapassou 70% de F1 ponderada (*weighted*). A classe dos "Anti-vaxxers" apresentou desempenho inferior a 60% na métrica F1.

Vale destacar que essa metodologia de coleta foi a mesma empregada na elaboração dos datasets para o SemEval-2016 Tarefa 6, e que os mesmos também apresentam

problemas de qualidade. O processo para remoção destes ruídos é bastante trabalhoso, requer um amplo conhecimento do conjunto de dados, e as ações de limpeza podem remover apenas parcialmente os problemas ou resultar na perda de informação importante (SOUSA; BECKER, 2022b).

## 6.2 Objetivos

A avaliação quantitativa desenvolvida nesta dissertação com base no estudo de caso descrito na Seção 6.1 abordou dois aspectos:

**Aplicabilidade**: O objetivo foi analisar a habilidade de aplicar o SSSD em um conjunto de dados novo, mantendo uma performance consistente com a apresentada nos dados da competição SemEval-2016 Tarefa 6 Tarefa 6 quando considerados domínios e alvos distintos daqueles. Além disso, os dados do estudo de caso estão expressos na língua Portuguesa. Em particular, procurou-se entender como aplicar o SSSD em cenários onde não há disponibilidade de dados rotulados.

**Sanidade**: O objetivo foi medir a robustez do SSSD face ao ruídos inerentes à coleta automatizada de *tweets* mediante palavras-chave ou *hashtags*. Neste método de coleta, a classificação dos posicionamentos é feita com auxílio de *hashtags* que os endossam ou contestam segundo o princípio da homofilia. Embora amplamente adotada, essa técnica pode trazer consigo variados vieses, como já discutido na Seção 6.1.2.

Especificamente, tratamos nesta avaliação dois tipos de problemas resultantes da coleta, exemplificados em nosso estudo de caso:

- **Falso-Positivos/Negativos:** *Tweets* são falsamente inseridos no contexto de uma *hashtag* porque um autor refuta uma ideia representada pela *hashtag*. Em outras palavras, alguns autores ao criticarem determinado posicionamento no Twitter, acabam replicando justamente a *hashtag* que endossa esse posicionamento.
- **Engajamento Artificial:** Neste fenômeno, eventos importantes, personalidades ou notícias são utilizados para engajar um posicionamento representado por uma *hashtag*, mesmo que não estejam diretamente relacionados ao contexto em questão. Neste caso, não há nenhuma ideia sendo expressa sobre um posicionamento.

## 6.3 Construção de um Query-set para o Domínio da Vacinação COVID-19

Para esta análise qualitativa, adotou-se o *dataset* do estudo de caso detalhado na Tabela 6.2 (*Pré-processado*). Seguindo as diretrizes da Seção 4.2, exploramos a Modelagem de Tópicos para criação de um *query-set*. Selecionamos um conjunto tweets para o *query-set* combinando três passos: (1) Modelagem de Tópicos para extração tweets sobre temas variados; (2) anotação manual destes tweets para confirmação de rótulos e identificação de tweets que constituissem ruídos; (3) divisão dos tweets manualmente rotulados em conjuntos de consulta (*query-set*) e de teste (*test-set*) usando classificação com validação cruzada. Estes passos são detalhados no restante desta seção.

### 6.3.1 Seleção de Tweets Baseada em Modelagem de Tópicos

Para assegurar a eficácia da DP e, por extensão, permitir que o SSSD classifique com bom desempenho o maior espectro de casos possíveis, é fundamental obter amostras de *tweets* que representem a diversidade das manifestações dos autores em relação ao alvo em questão (ALLAWAY; MCKEOWN, 2020). Nesse cenário, a Modelagem de Tópicos se destaca por sua habilidade em capturar de forma abrangente o contexto inerente a um *corpus* específico, simplificando a identificação de *tweets* que expressam variados posicionamentos. A implementação da Modelagem de Tópicos para os fins da presente avaliação qualitativa foi estruturada em duas fases principais:

- **Extração de Tópicos:** aplicou-se o BERTopic ao *domain-set* pré-processado, seguindo uma abordagem semelhante à descrita no estudo de caso (Seção 6.1.2), porém com propósito distinto. O foco aqui foi na seleção de *tweets* representativos de posicionamentos, e não na interpretação de tópicos. Exploramos as ferramentas de visualização e análise do *framework* BERTopic para determinar um número ótimo de tópicos de forma a garantir a representatividade e a diversidade desejada. O MPT escolhido foi o MPNET, por possuir uma boa qualidade para Modelagem de Tópicos. Como resultado, foram identificados 26 tópicos para cada posicionamento.
- **Seleção Aleatória de Tweets:** Para cada tópico extraído, foram selecionados aleatoriamente amostras de *tweets* obedecendo a proporção de 0,01% para os "Pro-Vaxxers"e 0,1% para os "Anti-Vaxxers". Esta proporção visou equilibrar o número de sementes para os dois posicionamentos. Após um pré-processamento que envolveu a remoção de

*tweets* duplicados, links, *hashtags* e menções, restou um total de 1445 *tweets*, dos quais 496 pertenciam ao grupo "Pro-vaxxers" e 949 ao grupo "Anti-vaxxers". Vale destacar que este pré-processamento não foi tão rigoroso quanto o adotado para a modelagem de tópicos, nos *domain-sets* ou em experimentos anteriores, pois precisávamos facilitar que humanos lessem os *tweets* para fins de anotação.

### 6.3.2 Anotação Manual de Tweets

Os *tweets* aleatoriamente selecionados no passo anterior foram distribuídos entre dois anotadores com conhecimento sobre a vacinação no contexto do pandemia de COVID-19. Os *tweets* foram organizados em dois arquivos distintos: "Pro-Vaxxers" (495 *tweets*) e "Anti-Vaxxers" (949 *tweets*). Cada arquivo é formado por duas colunas: "texto", que apresenta o *tweet* em formato texto, e "rótulo", que está em branco, destinada a indicar o posicionamento.

O objetivo deste formato de anotação foi identificar os vieses que foram introduzidos pelo método de coleta, conforme discutido anteriormente. Neste sentido, assumiu-se que o rótulo dos *tweets* era *a priori* aquele resultante do método de coleta, inferido pela respectivas *hashtag* a favor/contra. A tarefa dos anotadores foi rotular apenas quando discordassem do rótulo *a priori*.

Previamente à anotação, os anotadores foram orientados a ler (EBELING et al., 2022) para ter em mente o contexto político e a influência do mesmo sobre a expressão dos posicionamentos. As instruções de anotação encontram-se no Anexo B. Em resumo, os anotadores receberam os dois arquivos com o texto dos *tweets* sem *hashtags*, e foram orientados a considerar explicitamente o que estava escrito, sendo que o texto poderia ser interpretado no contexto político (e.g., manifestação contra o STF pela obrigatoriedade da vacina poderia ser interpretado como um *tweet* contra a vacinação). Foram fornecidos exemplos representativos de *tweets* a favor, contra ou não relacionados (nenhum) que cobrisse uma variedade de casos. Os exemplos foram discutidos, e os anotadores foram orientados a avaliar cuidadosamente cada *tweet* e, quando discordassem do posicionamento sugerido pelo nome do arquivo ("Pro-Vaxxers" ou "Anti-Vaxxer"), registrar o rótulo correto observando as seguintes regras:

- **Favor:** O texto do *tweet* permite compreender que a pessoa é a favor da vacinação.
- **Contra:** O texto do *tweet* permite compreender que a pessoa é contra a vacinação.

- **Nenhum:** O texto do *tweet* não expressa claramente posicionamento contra/a favor, ou não tem relação com o alvo.

Ao final desse processo, registrou-se uma convergência média em 79% nas anotações. Especificamente, houve 80% de concordância nas anotações dos *tweets* de "Pro-Vaxxers" e 78% nas dos "Anti-Vaxxers", resultando em um total de 1.171 *tweets* anotados com concordância. Em contrapartida, os anotadores apresentaram divergências na anotação de 308 *tweets*, os quais foram desprezados. A Tabela 6.4 sintetiza os resultados do processo de rotulagem por pares.

Tabela 6.4: Composição dos posicionamentos resultantes da rotulagem por pares.

| **Posicionamentos** | **# Tweets** | **% Favor** | **% Contra** | **% Nenhum** |
|---|---|---|---|---|
| Pro-Vaxxers | 385 | 75,5 | 0 | 24,5 |
| Anti-Vaxxers | 752 | 34,9 | 50,7 | 14,4 |
| **Total** | 1.171 | | | |

Na avaliação da distribuição de classes para os "Pro-Vaxxers", constatou-se que 75,5% foram classificados como "Favor", enquanto os restantes 24,5% enquadram-se como "Nenhum". Neste grupo não houve *tweets* classificados como "Contra", indicando ausência ou presença ínfima de viés de falsos-positivos. Em contrapartida, no conjunto "Anti-Vaxxers", observa-se que apenas 50,9% dos *tweets* pertencem efetivamente à classe "Contra" e que 14,4% foram rotulados na classe "Nenhum". A presença de 34,9% na categoria "Favor" sinaliza a alta ocorrência de falso-negativos. Estes resultados estão alinhados com as descobertas do estudo de caso, solidificando a confiança e robustez do método de seleção e rotulagem adotado.

### 6.3.3 Divisão dos Tweets em Amostras para Treinamento/Consulta

Inicialmente, os *tweets* anotados, tanto dos "Pro-Vaxxers" quanto dos "Anti-Vaxxers", foram consolidados em uma única amostra. Em seguida, empregou-se validação cruzada para treinar um modelo de Regressão Logística, segmentando a amostra em cinco partições, cada uma contendo 20% do total de *tweets*. Para o treinamento, os parâmetros do modelo foram configurados para *class_weight* = *balanced* e *solver* = *liblinear*.

Após o treinamento do modelo através das partições previamente determinadas, foi observado um pico no Macro-F1 de 76,3%, com uma média global de 73,0%. Adicionalmente, registrou-se um desvio padrão de 2,3 pp nos resultados, indicando uma variação

moderada entre as diferentes partições. Tais resultados evidenciam a eficácia do modelo treinado a partir da validação cruzada em capturar a diversidade e representatividade dos argumentos e posicionamentos veiculados nos *tweets*.

Estabeleceu-se o *query-set* para corresponder ao *subset* de treino e o *test-set* ao *subset* de teste resultantes da validação cruzada. Tal escolha baseou-se na combinação de partições que alcançou o valor mais elevado de Macro-F1, conforme detalhado na Tabela 6.5.

Tabela 6.5: Composição do conjunto de dados para consulta e treino.

| **Amostra** | | | **Macro-F1** | | |
|---|---|---|---|---|---|
| | # Tweets | % Favor | % Contra | % Nenhum | % Total |
| Query-Set | 910 | 48,6 | 33,6 | 17,7 | 80 |
| Test-Set | 227 | 48,6 | 33,3 | 18,0 | 20 |
| **Total** | **1.117** | | | | |

## 6.4 Experimentos

Os experimentos foram estruturados para avaliar o SSSD considerando as perspectivas enfatizadas na Seção 6.2. Com isso, buscamos responder às seguintes questões de pesquisa:

- **QP1:** Qual é o diferencial do SSSD em relação a um modelo tradicional de DP, levando em consideração diferentes quantidade de dados rotulados disponíveis para treinamento?
- **QP2:** Ao fornecer exemplos, é possível mitigar os casos associados a falsos negativos?
- **QP3:** Ao fornecer exemplos, é possível mitigar os casos associados ao engajamento artificial?

A métrica de avaliação utilizada foi a F1 sobre as diferentes classes, ae Macro F1 para a avaliação do conjunto, e a SemEval Macro-F1 para utilização dos mesmos critérios da competição SemEval. A métrica F1 calcula a média harmônica entre precisão e revocação, fornecendo uma boa perspectiva do desempenho do modelo, sobretudo quando há classes desbalanceadas. A Macro-F1 considera todas as classes (“Favor”, “Contra” e “Nenhum”). Já a métrica SemEval Macro-F1 considera apenas as classes “Favor” e “Contra”, como detalhado na Equação 5.1.

Todos os dados e código dos experimentos realizados estão acessíveis em um

repositório público[1].

### 6.4.1 QP1: Desempenho e Dados de Treino

**a) Método**

Este experimento teve como objetivo avaliar o desempenho do SSSD considerando quantidades diferentes de *tweets* para treinamento. Para este objetivo, o estudo envolveu uma análise comparativa entre o modelo de DP que emprega a o SSSD e outro modelo de DP que utiliza dados anotados manualmente. Por ter apresentado o melhor desempenho nas avaliações quantitativas (Seção 5.4), SSSD foi combinado com uma Regressão Logística (SSSD-RL). Como *baseline*, foi adotada a Regressão Logística tradicional (RL). É importante ressaltar que ambos os modelos, SSSD-RL e RL, foram parametrizados de forma idêntica, com *class_weight = balanced* e *solver = liblinear*. Para a extração de *features* dos *tweets* rotulados, foram aplicadas as técnicas de *tf-idf* e *bigrams*.

Para criar o modelo SSSD-RL, utilizou-se como *domain-set* o conjunto original de dados (Tabela 6.1, Tweets Pré-processados), do qual foram excluídos todos os *tweets* incluídos no *query-set* e *test-set*. Definiu-se um limiar de similaridade máximo de 0.90 para a Rotulagem Semântica.

Os modelos RL e SSSD-RL passaram por um processo de treinamento que envolveu a variação na quantidade de *tweets* rotulados utilizados. A partir dos 910 *tweets* disponíveis, o treinamento foi iniciado com uma amostra de 10% (91 *tweets*), incrementando-se em segmentos de 10 pp até abranger a totalidade dos dados. No caso dos modelos RL, cada segmento correspondia a uma quantidade de dados anotados usados como conjunto de treino. No caso dos modelos SSSD-RL, cada segmento correspondia a um *query-set* usado para rotular dados de treinos oriundos do *domain-set*. Os 227 *tweets* designados como *test-set* permaneceram inalterados ao longo das 10 iterações e serviram para avaliar todos os modelos.

Para definir o valor de $k$ dos modelos SSSD-RL, para cada segmento foram realizadas 20 iterações, ajustando-se o valor de $k$ em intervalos de 5. Este procedimento gerou 200 modelos ao longo do experimento, sendo adotado o de melhor desempenho em cada segmento, o que resultou em 10 modelos SSSD-RL.

**b) Resultados**

[1] https://github.com/mediote/sssd

Os resultados são apresentados na Figura 6.1, que mostra os resultados em termos de Macro-F1, e na Tabela 6.6, que detalha também as métricas usadas no SEMEval e o valor de *k* correspondente a cada modelo.

Figura 6.1: SSSD-RL x RL em função do % Tweets (Macro-F1).

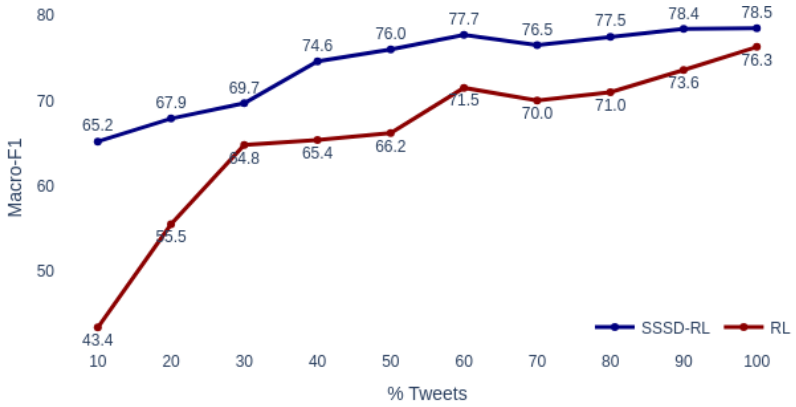



Tabela 6.6: Resultados do treinamento com diferentes quantidades de tweets.

| **Treino/Query-Set (910 tweets)** | | **Macro-F1 (favor, contra, nenhum)** | | | **SemEval Macro-F1 (favor, contra)** | | |
|---|---|---|---|---|---|---|---|
| % Tweets | # Tweets | % RL | % SSSD-RL | pp Melhora | % RL | % SSSD-RL | pp Melhora |
| 10 | 91 | 43,4 | $65,2^{65}$ | 21,8 | 54,5 | $65,6^{65}$ | 11,1 |
| 20 | 182 | 55,5 | $67,9^{100}$ | 12,4 | 63,2 | $69,6^{100}$ | 6,4 |
| 30 | 273 | 64,8 | $69,7^{80}$ | 4,9 | 72,7 | $79,1^{80}$ | 6,4 |
| 40 | 364 | 65,4 | $74,6^{65}$ | 9,2 | 72,8 | $77,1^{65}$ | 4,3 |
| 50 | 455 | 66,2 | $76,0^{85}$ | 9,8 | 74,4 | $78,4^{85}$ | 4,0 |
| 60 | 546 | 71,5 | $77,7^{55}$ | 6,2 | 77,9 | $80,0^{55}$ | 2,1 |
| 70 | 637 | 70 | $76,5^{30}$ | 6,5 | 76,2 | $79,9^{30}$ | 3,7 |
| 80 | 728 | 71 | $77,5^{15}$ | 6,5 | 77,9 | $80,6^{15}$ | 2,7 |
| 90 | 819 | 73,6 | $78,4^{30}$ | 4,8 | 78,0 | $80,5^{30}$ | 2,5 |
| 100 | 910 | 76,3 | $78,5^{65}$ | 2,2 | 79,8 | $81,9^{65}$ | 2,1 |

Ao analisar os resultados da Figura 6.1, que destaca a métrica Macro-F1, percebe-se uma diferença notável no desempenho entre os modelos SSSD-RL e RL considerando os diferentes segmentos de *tweets*. Com o uso de apenas 10% de *tweets* rotulados, o SSSD-RL apresenta um desempenho de 65,2%, enquanto o RL tem um desempenho de apenas 43,4%. Esta diferença inicial de 21,8 pp destaca a capacidade superior do SSSD-RL em adaptar-se com menos dados rotulados disponíveis. É importante enfatizar que, para todas as iterações do SSSD-RL, a quantidade de *tweets* empregados no treinamento do modelo é resultado da multiplicação do tamanho do segmento pelo valor de $K$, como indicado pelo expoente nos resultados apresentados. Neste contexto, quando o RL foi treinado com 91 *tweets*, usando a mesma quantidade de tweets como *query-set* o SSSD-RL utilizou na verdade um conjunto expandido de 5.915 *tweets* (considerando $K$=65).

Esta abordagem não apenas maximiza o uso de *tweets* disponíveis, mas também otimiza significativamente os resultados de classificação em comparação ao RL.

A vantagem do SSSD-RL se mantém à medida que mais dados são incorporados ao treinamento/consulta. Por exemplo, com 50% dos dados, o SSSD-RL registra um desempenho de 76%, enquanto o RL tradicional alcança 66,2%. Neste ponto, a vantagem é de 9,8 pp em favor do SSSD-RL. Mesmo utilizando apenas metade dos dados disponíveis, o SSSD-RL tem o mesmo desempenho que o RL alcança ao ser treinado com a totalidade dos *tweets*. Isso ressalta não apenas a eficácia do SSSD-RL em cenários de dados limitados, mas também seu potencial de economia de recursos, uma vez que pode alcançar desempenhos comparáveis com conjuntos de dados rotulados significativamente menores. Quando o uso de *tweets* para treinamento/consulta se aproxima de 100%, a diferença entre os algoritmos se estreita, mas o SSSD-RL ainda mantém uma ligeira vantagem. Aos 100%, o SSSD-RL alcança 78,5% e o DP tradicional 76,3%, sendo a vantagem de 2,2 pp para o SSSD-RL.

Em relação aos resultados obtidos através da métrica SemEval Macro-F1, detalhados na Figura 6.2, o SSSD-RL supera consistentemente o RL em todas as faixas de dados. Com 10% dos dados, o SSSD-RL tem desempenho de 65,6%, enquanto o RL registra 54,5%. Aos 30%, o SSSD-RL mostra um avanço significativo, atingindo 79,1%, superando o RL em 6,4 pp. Mesmo com 100% dos dados, o SSSD-RL continua liderando com 81,9%, com uma diferença de 2,1 pp em relação ao RL, evidenciando sua eficiência tanto em pequenos quanto em grandes conjuntos de dados. Traçando um paralelo com os resultados da avaliação quantitativa discutida no Capítulo 5, o SSSD-RL exibiu performances muito parecidas considerando os alvos “Legalização do Aborto”, “Movimento Feminista”, “Hillary Clinton” e “Trump Trump”, que tiveram pontuações de 82,0%, 78,5%, 81,2% e 85,2%, respectivamente. Essa constatação realça a versatilidade e solidez do SSSD quando aplicado em contextos variados.

Figura 6.2: SSSD-RL x RL em função do % Tweets (SemEval Macro-F1).

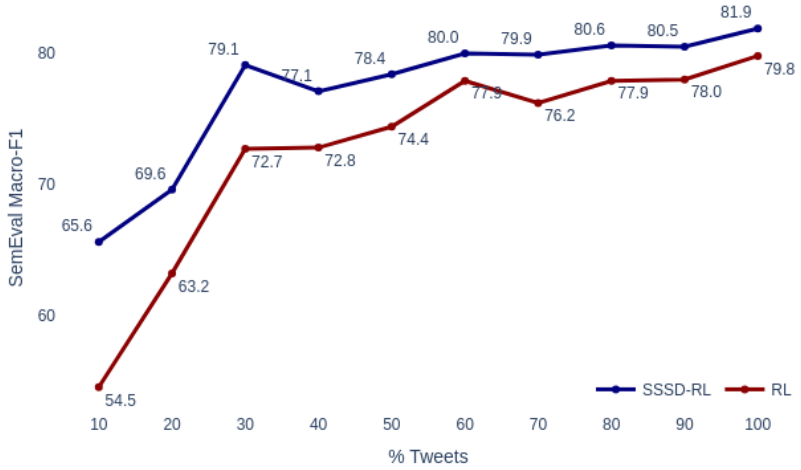
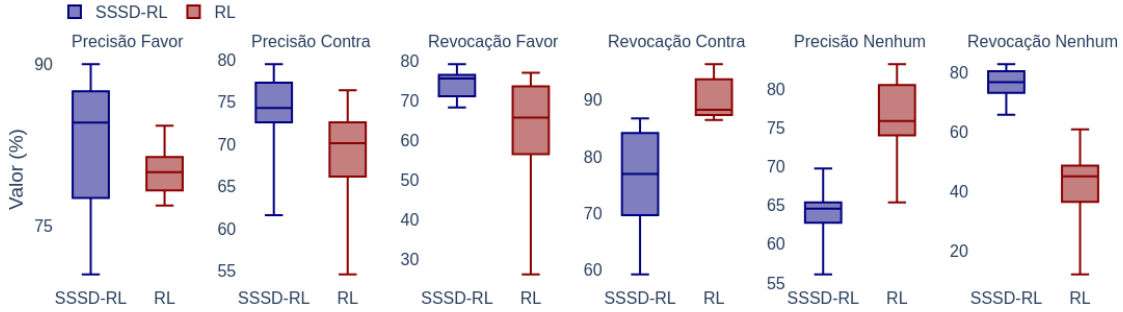



Figura 6.3: Distribuição da Precisão e Revocação entre as classes (Macro-F1).




A Figura 6.3 detalha o comportamento da precisão e revocação nos dois métodos usando *boxplots* para cada uma das classes. Verificam-se comportamentos distintos para as classes "Favor", "Contra" e "Nenhum" nos algoritmos SSSD-RL e RL. O SSSD-RL apresenta uma superioridade consistente em termos de precisão e revocação para as classes "Favor". Já para a classe "Contra", o SSSD-RL tem desempenho superior para precisão, e inferior para a revocação. Em contraste, o RL se destaca com uma revocação relativamente alta com mediana 88,3% para a classe "Contra", mas sua precisão para a mesma classe é consideravelmente menor (70,1%), resultando em um Macro-F1 mediano de 68,1%. O SSSD-RL exibe um maior equilíbrio, com medianas de 76,9% em revocação e 74,3% em precisão, o que resulta em um Macro-F1 mediano de 77,7%.

Já na classe "Nenhum" observam-se diferenças mais significativas. Para esta classe a mediana de precisão no SSSD-RL é de 64,6%, e de revocação, 76,8%. Esses números, apesar de modestos, refletem um equilíbrio considerável entre as duas métricas, indicando sua habilidade em identificar corretamente *tweets* neutros, minimizando ao mesmo tempo as chances de classificações incorretas nas demais classes. Em contraste, o RL apresenta uma precisão mais elevada, de aproximadamente 75,9%, mas com uma revocação consideravelmente mais baixa, em torno de 45,1%. No cenário de DP, isso

pode ser problemático, pois aumenta a chance de classificações errôneas. Deste modo, quando se busca um desempenho equilibrado entre precisão e revocação, sobretudo em aplicações de DP, o SSSD-RL emerge como a alternativa mais apropriada.

Diante das análises, fica evidente a superioridade em termos de qualidade e desempenho do SSSD-RL em relação ao RL, especialmente ao considerarmos a economia e eficácia na utilização de dados rotulados. Mesmo quando operando com quantidades limitadas de dados, o SSSD-RL demonstra capacidade notável em otimizar a classificação. Além disso, essa vantagem não se restringe a pequenos conjuntos de dados, pois mesmo ao utilizar todo o conjunto disponível, o SSSD-RL alcança resultados superiores. A sua habilidade em equilibrar precisão e revocação, especialmente para todas as classes, ressalta a sua relevância prática neste domínio de aplicação. Ao considerar diferentes métricas e contextos, a consistência do SSSD-RL em superar o RL destaca sua robustez e versatilidade, tornando-o uma ferramenta promissora para futuras aplicações e pesquisas na área de DP.

### 6.4.2 QP2: Falsos Negativos

**a) Método**

O objetivo deste experimento foi verificar se os *tweets* inicialmente associados ao movimento "Anti-Vaxxers", mas anotados como "Favor" durante o processo de rotulagem manual detalhado na Seção 6.3.2, preservam esse rótulo ao serem submetidos ao SSSD. Esta avaliação é importante para mostrar a capacidade do SSSD em lidar com os vieses de falso-negativos.

Para conduzir esta análise, optou-se pelos modelos SSSD-RL e RL treinados utilizando 50% dos 910 *tweets* acessíveis, o que corresponde a 455 *tweets*. Do conjunto residual de 455 *tweets* anotados, utilizamos como conjunto de teste apenas 303 *tweets* que foram coletados por meio de *hashtags* ligadas ao movimento "Anti-Vaxxers". Dentro desse subconjunto, todos aqueles pertencentes a classe "Favor" (108) são referentes a falso-negativos detectados durante o processo de rotulagem manual.

**b) Resultados**

Ao analisar os resultados dos experimentos sintetizados na Tabela 6.7, percebe-se diferenças claras nos resultados dos algoritmos SSSD-RL e RL, com variações significativas conforme a métrica e a classe em foco. De maneira global, o SSSD-RL demonstrou

superioridade em relação ao RL, especialmente sob a ótica da métrica Macro-F1, onde alcançou uma pontuação de 71,5%, em comparação aos 62,0% obtidos pelo RL.

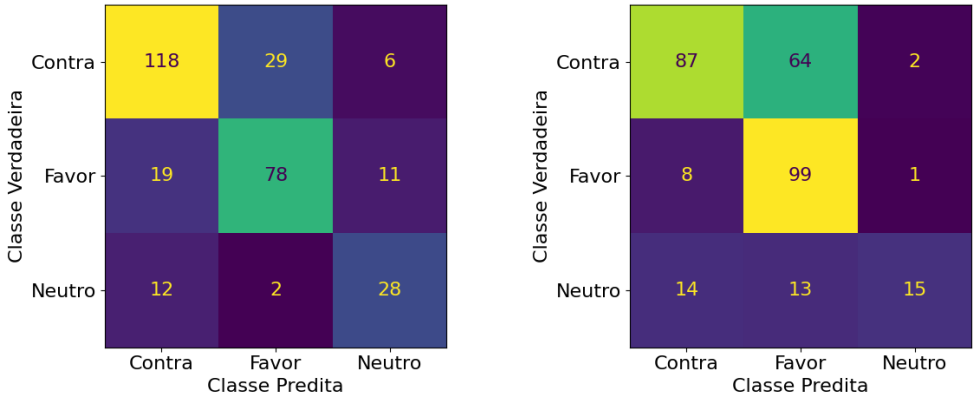
Tabela 6.7: Resultados de classificação nos casos de falsos positivos.

| **Classe** | **SSSD-RL** (50% Tweets) | | | **RL** (50% Tweets) | | | |
|---|---|---|---|---|---|---|---|
| | Precisão | Revocação | F1 | Precisão | Revocação | F1 | Suporte |
| Favor | 71,6 | 72,2 | 71,9 | 56,2 | 91,7 | 69,7 | 108 |
| Contra | 79,2 | 77,1 | 78,1 | 79,8 | 56,9 | 66,4 | 153 |
| Nenhum | 62,2 | 66,7 | 64,4 | 83,3 | 35,7 | 50,0 | 42 |
| **Macro-F1** | 71,0 | 72,0 | 71,5 | 73,1 | 61,4 | 62,0 | 303 |

Ao focar na classe “Favor”, que representa os falso-negativos sob o ponto de vista dos vieses introduzidos, notamos que o SSSD-RL detém uma vantagem considerável em precisão, registrando 71,6% contra os 56,2% do RL. Paralelamente, o SSSD-RL apresenta uma revocação de 72,2% e uma pontuação F1 de 71,9%, enfatizando sua competência em reconhecer de forma equilibrada uma parcela considerável dos *tweets* que genuinamente pertencem a classe "Favor", sem comprometer outras classes. Os erros/acertos do SSSD-RL são mostrados na matriz de confusão da Figura 6.4.(a), que denota um relativo equilíbrio nos erros cometidos ao longo das demais classes.

Em contrapartida, o RL registou uma revocação excelente de 91,7%, ultrapassando largamente os 72,2% alcançados pelo SSSD-RL. No entanto, esta precisão ocorre ao custo da revocação, de apenas 56,2%, a qual resulta em uma quantidade considerável de classificações incorretas já que tenta rotular instância de outras classes com “Favor”. Este comprometimento na precisão é evidenciado pelo incremento substancial de 120,6% nos erros de classificação entre as classes “Contra” e “Favor” em relação ao SSSD-RL, conforme demonstrado na Figura 6.4.(b).

Figura 6.4: Matriz de Confusão.




(a) SSSD-RL (b) RL

Os erros de classificação entre as classes “Contra” e “Favor” ressaltam a complexidade inerente ao enfrentamento desse viés em específico. Esta complexidade é exacerbada pelas sutilezas na expressão de posicionamentos, muitas vezes carregadas de ironia e sarcasmo, especialmente em contextos politicamente carregados, como destacado por Ebeling et al. (2022). Tais nuances contribuem significativamente para a ambiguidade, tornando a distinção entre posturas favoráveis e contrárias mais desafiadoras para os algoritmos de classificação. Manifestações falso-negativas como “Não toma vacina, otário, o problema é todo seu” ou “Não tomem mesmo, bom que sobra” confundem-se com expressões verdadeiro-negativas como “Também não vou tomar vacina” ou “Não tomo, pode dar pra alguém da esquerdalha”, ilustrando a dificuldade em discernir os posicionamentos corretamente nesta situação.

No que se refere a erros de classificação associados à classe “Nenhum”, embora ocorram em menor número neste experimento, observa-se uma propensão distinta entre os modelos analisados, conforme mostrado na Figura 6.4. O SSSD-RL classifica corretamente mais instâncias da classe “Nenhum” em relação ao RL. Contudo, observa-se que o SSSD-RL possui uma inclinação a classificar equivocadamente *tweets* das classes “Favor” ou “Contra” como “Nenhum”. Em contraste, o RL manifesta uma propensão inversa, tendendo a classificar incorretamente a classe “Nenhum” como “Favor” ou “Contra”. Essa dinâmica ressalta as diferenças sutis nas abordagens de classificação adotadas por cada sistema, influenciando sua capacidade de identificar corretamente a neutralidade ou a polarização nas postagens analisadas.

Portanto, diante de todos os desafios apresentados e com o objetivo de atenuar os vieses de falso-negativos sem acarretar um aumento significativo de erros de classificação em outras classes, o SSSD-RL apresenta-se como a solução mais adequada. Essa assertividade se mantém mesmo quando ambos os modelos operam com volumes idênticos de *tweets* para treinamento e consulta. O SSSD-RL estabelece um maior equilíbrio entre precisão e revocação, assegurando que a maioria dos *tweets* categorizados como “Favor” seja efetivamente representativa dessa classe, simultaneamente reconhecendo uma parcela significativa de *tweets* autênticos da classe “Contra”.

### 6.4.3 QP3: Engajamento Artificial

**a) Método**

O objetivo desse experimento foi verificar se os *tweets* associados engajamento

artificial, anotados como “Nenhum” durante o processo de rotulagem manual, preservam esse rótulo ao serem submetidos ao SSSD. Novamente, utilizou-se ambos os modelos SSSD-Rl e RL treinados com 50% dos *tweets* disponíveis.

Para os testes, foram empregados 391 *tweets* anotados manualmente, dos quais 221 pertencem a classe “Favor”, 153 a classe “Contra” e 17 a classe “Nenhum”. Adicionalmente, 80 *tweets* associados a engajamento artificial e categorizados como “Nenhum” foram cuidadosamente selecionados e incluídos no conjunto de teste pelo autor, baseando-se em seu conhecimento prévio sobre tópicos que apresentavam esse viés, conforme descrito em trabalhos anteriores (SOUSA; BECKER, 2022b), e utilizando exemplos conhecidos relacionados a celebridades ou eventos. Essa inclusão foi necessária porque apenas 17 *tweets* manualmente rotulados como “Nenhum” estavam especificamente associados ao engajamento artificial. Todos os 97 *tweets* pertencentes a classe “Nenhum” no conjunto de teste, correspondem a exemplos de engajamento artificial.

**b) Resultados**

Os resultados decorrentes do treinamento dos modelos SSSD-RL e RL estão delineados na Tabela 6.8. O SSSD-RL reafirmou sua superioridade, registrando uma pontuação Macro-F1 de 84,2%, registrando uma melhora de 4pp em relação aos 80,3% alcançados pelo RL.

Tabela 6.8: Resultados de classificação nos casos de engajamento artificial.

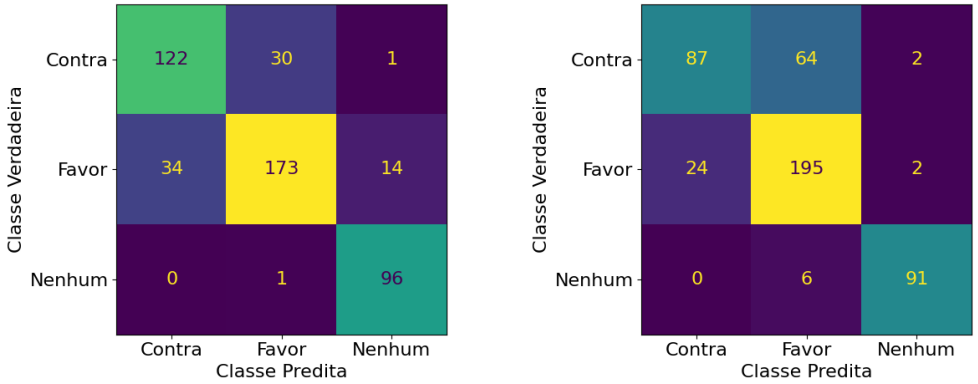
| **Classe** | **SSSD-RL** (50% Tweets) | | | **RL** (50% Tweets) | | | |
|---|---|---|---|---|---|---|---|
| | Precisão | Revocação | F1 | Precisão | Revocação | F1 | Suporte |
| Favor | 84,8 | 78,3 | 81,4 | 73,6 | 88,2 | 80,2 | 221 |
| Contra | 78,2 | 79,7 | 79,0 | 78,4 | 56,9 | 65,9 | 153 |
| Nenhum | 86,5 | 99,0 | 92,3 | 95,8 | 93,8 | 94,8 | 97 |
| **Macro-F1** | 83,2 | 85,7 | 84,2 | 82,6 | 79,6 | 80,3 | 471 |

Ao concentrar-se na classe “Nenhum”, o RL supera o SSSD-RL por uma pequena margem de 2,5 pp na classificação correta das instâncias desta classe, atingindo uma pontuação F1 de 94,8% contra 92,3%. Entretanto, uma inspeção minuciosa das matrizes de confusão, apresentadas na Figura 6.5, mostra que, dos 97 *tweets* identificados como engajamento artificial, o SSSD-RL classificou corretamente 96, enquanto o RL acertou em 91. Considerando que todos os erros de classificação feitos por ambos os modelos nesse subconjunto específico categorizaram equivocadamente os *tweets* como “Favor”, a habilidade do SSSD-RL de equilibrar uma alta taxa de revocação de 99,0% para a classe “Nenhum” com uma precisão de 84,8% para a classe “Favor” levou a apenas 1 classificação incorreta. Em contraste, o RL, que possui uma precisão de apenas 73,6% para a

classe “Favor”, cometeu 6 erros de classificação.

Em termos absolutos, ambos os modelos exibiram desempenhos semelhantes ao considerar especificamente esse tipo de viés. Isso sugere que, nesse contexto, o processo de rotulagem assume uma importância crítica, possivelmente sendo até mais influente do que o modelo de classificação adotado. Assim, a acurácia na rotulagem inicial revela-se essencial para a eficácia do sistema de classificação, independentemente da complexidade do modelo de DP utilizado.

Figura 6.5: Matriz de Confusão.




(a) SSSD-RL (b) RL

Contudo, ao avaliar a classe “Contra”, conforme apresentado na Tabela 6.8, revela que o RL apresenta uma baixa revocação para o RL (56,9%), culminando em uma pontuação F1 de apenas 65,9%. Isso se traduz em um volume significativo de classificações errôneas. Tal tendência, evidenciada na Figura 6.5, alinha-se com os padrões observados no experimento anterior. Comparado ao SSSD-RL, o RL introduz uma considerável quantidade de erros na classe “Contra” ao tentar classificar a maioria dos *tweets* da categoria “Favor”. Este aspecto enfatiza a superioridade do SSSD-RL na gestão de vieses, indicando que, embora os casos de engajamento artificial estejam mais intrinsecamente ligados à qualidade da rotulagem do que ao modelo de classificação, o SSSD-RL consegue um resultado melhor ao mesmo tempo que introduz menos erros nas outras classes e economiza recursos.

# 7 CONCLUSÕES E TRABALHOS FUTUROS

A DP em redes sociais é uma tarefa complexa e repleta de desafios. Um dos mais importantes é a escassez crônica de dados rotulados, um problema que impede muitas abordagens de AM de alcançarem seu potencial completo comprometendo sua aplicabilidade. O SSSD emerge como uma solução viável e eficiente para este impasse, explorando a eficácia dos MPTs em capturar o conteúdo semântico e contextual dos *tweets* através da Busca Semântica, permitindo a rotulagem precisa de grandes volumes de dados com menor custo humano para treinamento de modelos de DP. O SSSD é agnóstico quanto às técnicas específicas de classificação, e pode ser aplicado a diferentes domínios, alvos e linguagens.

Os resultados da análise quantitativa mostram que o SSSD superou os trabalhos estado da arte para todos os alvos, testados diferentes algoritmos tradicionais de classificação. As melhorias substanciais observadas nas Tarefas A e B do SemEval-2016 Tarefa 6 ilustram a capacidade do método de adaptar-se e detectar posicionamentos em relação a diferentes alvos com precisão e confiabilidade. Um ponto chave para essa eficácia reside na manipulação do valor de $k$, de maneira a encontrar um equilíbrio ideal entre o número de *tweets* rotulados e as pontuações de similaridade. Os experimentos mostraram que a ausência de um padrão no valor de $k$, mas a abordagem utilizada nos experimentos permite encontrar os modelos com melhor desempenho ao variar o valor de $k$.

Do ponto de vista qualitativo, os experimentos demonstraram que SSSD é eficiente quando aplicado a novos domínios mostrando um desempenho sólido para o alvo "Vacinação" no contexto da pandemia de COVID-19 no Brasil evidenciando sua reprodutibilidade. Adicionalmente, o SSSD é um método relativamente simples de se aplicar, pois as técnicas envolvidas são amplamente conhecidas e disponibilizadas através de *frameworks* como o SentenceTranformers[1] por exemplo. A sua capacidade em alcançar resultados competitivos, mesmo com menos dados, enfatiza sua eficiência, sendo uma opção viável em contextos com escassez de dados rotulados. Um ponto forte do SSSD, é sua habilidade de filtrar informações irrelevantes e identificar representações genuínas de posicionamentos, crucial para combater problemas como engajamento artificial e falso-positivos, comuns em métodos automatizados de coleta de *tweets*.

O SSSD demonstrou sua capacidade de ser agnóstico em relação à linguagem, alcançando bons resultados tanto em conjuntos de dados relacionados à vacinação na

[1] https://www.sbert.net/examples/applications/semantic-search/README.html

língua portuguesa quanto em conjuntos de dados do SemEval-2016 Tarefa 6 em inglês.

Este trabalho resultou nas seguintes publicações como primeiro autor, listadas abaixo em ordem cronológica, junto à classificação QUALIS do forum:

- KDMille (B5): Pro/anti-vaxxers in Brazil: a temporal analysis of COVID vaccination stance in Twitter. Anais do IX Symposium on Knowledge Discovery, Mining and Learning, 2021. Este artigo foi agraciado com o prêmio de Melhor Artigo do evento.

- JIDM (B4): Understanding the COVID vaccination stances in Brazil: a temporal analysis using Twitter data. Journal of Information and Data Management, v. 13, n. 6, 2022.

- SBBD (A4): Comparando os posicionamentos a favor/contra a vacinação COVID nos Estados Unidos da América e no Brasil. Anais do XXXVII Simpósio Brasileiro de Bancos de Dados. 2022.

- RANLP (A3): SSSD: Leveraging pre-trained models and semantic search for semi-supervised stance detection. Proceedings of Recent Advances in Natural Language Processing (RANLP). 2023.

A principal limitação do SSSD reside na seleção dos *tweets* que compõem o *query-set*. É fundamental que os exemplos escolhidos sejam representativos e possuam o mínimo de ambiguidade, pois, caso contrário, a rotulagem automática pode se mostrar imprecisa. Ademais, o método pode ser sensível às variações temporais durante a coleta de dados, uma vez que mudanças no sentimento público ou no contexto podem impactar os resultados. Outro aspecto a ser considerado é que o desempenho dos modelos pode ser restrito caso o *test-set* não seja suficientemente representativo em relação ao *domain-set* e ao *query-set*.

Em trabalhos futuros, pretende-se testar o SSSD em conjunto com outras estratégias de AP, já que estas se beneficiariam enormemente da habilidade do SSSD em rotular grandes volumes de dados automaticamente. Esse teste poderia revelar o real potencial do SSSD, pois os métodos convencionais de AM usados nos experimentos podem estar agindo como um fator limitador em frente às complexidades da DP no contexto de redes sociais como o Twitter. Ademais, o SSSD ainda possui amplo espaço para otimizações. Há várias oportunidades de melhoria, como o *fine-tunninng* dos MPTs usados na Rotulagem Semântica ou a adoção de modelos como o BERT também na fase de Detecção de Posicionamentos.

# REFERÊNCIAS

AHMED, M.; CHY, A. N.; CHOWDHURY, N. K. Incorporating hand-crafted features in a neural network model for stance detection on microblog. In: **Proceedings of the 6th International Conference on Communication and Information Processing**. [S.l.: s.n.], 2020. p. 57–64.

AL-GHADIR, A. I.; AZMI, A. M.; HUSSAIN, A. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. **Information Fusion**, Elsevier, v. 67, p. 29–40, 2021.

ALDAYEL, A.; MAGDY, W. Your stance is exposed! analysing possible factors for stance detection on social media. **Proceedings of the ACM on human-computer interaction**, ACM New York, NY, USA, v. 3, n. CSCW, p. 1–20, 2019.

ALDAYEL, A.; MAGDY, W. Stance detection on social media: State of the art and trends. **Information Processing & Management**, Elsevier, v. 58, n. 4, p. 102597, 2021.

ALLAWAY, E.; MCKEOWN, K. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In: WEBBER, B. et al. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 8913–8931. Available from Internet: <https://aclanthology.org/2020.emnlp-main.717>.

ALLAWAY, E.; SRIKANTH, M.; MCKEOWN, K. Adversarial learning for zero-shot stance detection on social media. In: TOUTANOVA, K. et al. (Ed.). **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Online: Association for Computational Linguistics, 2021. p. 4756–4767. Available from Internet: <https://aclanthology.org/2021.naacl-main.379>.

ALTURAYEIF, N.; LUQMAN, H.; AHMED, M. A systematic review of machine learning techniques for stance detection and its applications. **Neural Computing and Applications**, Springer, v. 35, n. 7, p. 5113–5144, 2023.

ANGELOV, D. Top2vec: Distributed representations of topics. **arXiv preprint arXiv:2008.09470**, 2020.

AUGENSTEIN, I. et al. Stance detection with bidirectional conditional encoding. In: SU, J.; DUH, K.; CARRERAS, X. (Ed.). **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. Austin, Texas: Association for Computational Linguistics, 2016. p. 876–885. Available from Internet: <https://aclanthology.org/D16-1084>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. [S.l.]: ACM press New York, 1999.

BESSI, A. et al. Homophily and polarization in the age of misinformation. **The European Physical Journal Special Topics**, Springer, v. 225, p. 2047–2059, 2016.

BHATT, G. et al. Combining neural, statistical and external features for fake news stance identification. In: **Companion Proceedings of the The Web Conference 2018**. [S.l.: s.n.], 2018. p. 1353–1357.

BORGES, L.; MARTINS, B.; CALADO, P. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. **Journal of Data and Information Quality (JDIQ)**, ACM New York, NY, USA, v. 11, n. 3, p. 1–26, 2019.

CAVNAR, W. B.; TRENKLE, J. M. et al. N-gram-based text categorization. In: LAS VEGAS, NV. **Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval**. [S.l.], 1994. v. 161175, p. 14.

CHEN, H. et al. A tutorial on network embeddings. **arXiv preprint arXiv:1808.02590**, 2018.

CHEN, P.; YE, K.; CUI, X. Integrating n-gram features into pre-trained model: a novel ensemble model for multi-target stance detection. In: SPRINGER. **International conference on artificial neural networks**. [S.l.], 2021. p. 269–279.

CHEN, W.-F.; KU, L.-W. UTCNN: a deep learning model of stance classification on social media text. In: MATSUMOTO, Y.; PRASAD, R. (Ed.). **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 1635–1645. Available from Internet: <https://aclanthology.org/C16-1154>.

CIGNARELLA, A. T. et al. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. In: CEUR. **Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)**. [S.l.], 2020. p. 1–10.

CONFORTI, C. et al. Will-they-won't-they: A very large dataset for stance detection on Twitter. 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. 1715–1724. Available from Internet: <https://aclanthology.org/2020.acl-main.157>.

CONFORTI, C. et al. Synthetic examples improve cross-target generalization: A study on stance detection on a twitter corpus. In: **Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. [S.l.: s.n.], 2021. p. 181–187.

CONOVER, M. et al. Political polarization on twitter. In: **Proceedings of the international aaai conference on web and social media**. [S.l.: s.n.], 2011. v. 5, n. 1, p. 89–96.

DARWISH, K. et al. Predicting online islamophobic behavior after# parisattacks. **The Journal of Web Science**, Now Publishers, Inc., v. 4, n. 3, p. 34–52, 2018.

DARWISH, K. et al. Unsupervised user stance detection on twitter. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2020. v. 14, p. 141–152.

DERCZYNSKI, L. et al. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In: BETHARD, S. et al. (Ed.). **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 69–76. Available from Internet: <https://aclanthology.org/S17-2006>.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: **Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)**. [S.l.: s.n.], 2019. p. 4171–4186.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DEY, K.; SHRIVASTAVA, R.; KAUSHIK, S. Twitter stance detection—a subjectivity and sentiment polarity inspired two-phase approach. In: IEEE. **2017 IEEE international conference on data mining workshops (ICDMW)**. [S.l.], 2017. p. 365–372.

DEY, K.; SHRIVASTAVA, R.; KAUSHIK, S. Topical stance detection for twitter: A two-phase lstm model using attention. In: SPRINGER. **Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40**. [S.l.], 2018. p. 529–536.

DIAS, M.; BECKER, K. Inf-ufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. [S.l.: s.n.], 2016. p. 378–383.

EBELING, R. et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. In: **Proc. of the 15th Intl. Conference on Web and Social Media (ICWSM) - To appear**. [s.n.], 2022. Available from Internet: <arXiv:2110.03382>.

EL-ALFY, E.-S. M.; LUQMAN, H. A comprehensive survey and taxonomy of sign language research. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 114, p. 105198, 2022.

GHOSH, S. et al. Stance detection in web and social media: a comparative study. In: SPRINGER. **Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10**. [S.l.], 2019. p. 75–87.

GIORGIONI, S. et al. Unitor@ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In: **EVALITA**. [S.l.: s.n.], 2020.

GÓMEZ-SUTA, M.; ECHEVERRY-CORREA, J.; SOTO-MEJÍA, J. A. Stance detection in tweets: A topic modeling approach supporting explainability. **Expert Systems with Applications**, Elsevier, v. 214, p. 119046, 2023.

GORRELL, G. et al. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: MAY, J. et al. (Ed.). **Proceedings of the 13th International Workshop on Semantic Evaluation**. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 845–854. Available from Internet: <https://aclanthology.org/S19-2147>.

GRIMMINGER, L.; KLINGER, R. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In: CLERCQ, O. D. et al. (Ed.). **Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. Online: Association for Computational Linguistics, 2021. p. 171–180. Available from Internet: <https://aclanthology.org/2021.wassa-1.18>.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

HACOHEN-KERNER, Y.; IDO, Z.; YA'AKOBOV, R. Stance classification of tweets using skip char ngrams. In: SPRINGER. **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10**. [S.l.], 2017. p. 266–278.

HAN, X. et al. Pre-trained models: Past, present and future. **AI Open**, Elsevier, v. 2, p. 225–250, 2021.

HOSSEINIA, M.; DRAGUT, E.; MUKHERJEE, A. Stance prediction for contemporary issues: Data and experiments. In: KU, L.-W.; LI, C.-T. (Ed.). **Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media**. Online: Association for Computational Linguistics, 2020. p. 32–40. Available from Internet: <https://aclanthology.org/2020.socialnlp-1.5>.

ILIĆ, S. et al. Deep contextualized word representations for detecting sarcasm and irony. In: BALAHUR, A. et al. (Ed.). **Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 2–7. Available from Internet: <https://aclanthology.org/W18-6202>.

JAFFE, A. **Stance: sociolinguistic perspectives**. [S.l.]: Oup Usa, 2009.

KAWINTIRANON, K.; SINGH, L. Knowledge enhanced masked language model for stance detection. In: **Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies**. [S.l.: s.n.], 2021. p. 4725–4735.

KOCHKINA, E.; LIAKATA, M.; AUGENSTEIN, I. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. **arXiv preprint arXiv:1704.07221**, 2017.

KÜÇÜK, D.; CAN, F. Stance detection: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 53, n. 1, p. 1–37, 2020.

LAI, M. et al. itacos at ibereval2017: Detecting stance in catalan and spanish tweets. In: CEUR-WS. ORG. **CEUR WORKSHOP PROCEEDINGS**. [S.l.], 2017. v. 1881, p. 185–192.

LAI, M. et al. Multilingual stance detection in social media political debates. **Computer Speech & Language**, Elsevier, v. 63, p. 101075, 2020.

LAI, M. et al. Friends and enemies of clinton and trump: using context for detecting stance in political tweets. In: SPRINGER. **Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I 15**. [S.l.], 2017. p. 155–168.

LAI, M. et al. # brexit: Leave or remain? the role of user's community and diachronic evolution on stance detection. **Journal of Intelligent & Fuzzy Systems**, IOS Press, v. 39, n. 2, p. 2341–2352, 2020.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: PMLR. **International conference on machine learning**. [S.l.], 2014. p. 1188–1196.

LI, Q. et al. A survey on text classification: From traditional to deep learning. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, v. 13, n. 2, p. 1–41, 2022.

LI, W.; XU, Y.; WANG, G. Stance detection of microblog text based on two-channel cnn-gru fusion network. **IEEE Access**, IEEE, v. 7, p. 145944–145952, 2019.

LIANG, B. et al. Target-adaptive graph for cross-target stance detection. In: **Proceedings of the Web Conference 2021**. [S.l.: s.n.], 2021. p. 3453–3464.

LIU, R. et al. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In: **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**. [S.l.: s.n.], 2021. p. 3152–3157.

LIU, Y. et al. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In: CARPUAT, M.; MARNEFFE, M.-C. de; RUIZ, I. V. M. (Ed.). **Findings of the Association for Computational Linguistics: NAACL 2022**. Seattle, United States: Association for Computational Linguistics, 2022. p. 1354–1374. Available from Internet: <https://aclanthology.org/2022.findings-naacl.101>.

LUO, Y. et al. Exploiting sentiment and common sense for zero-shot stance detection. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the 29th International Conference on Computational Linguistics**. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. p. 7112–7123. Available from Internet: <https://aclanthology.org/2022.coling-1.621>.

LYNN, V. et al. Tweet classification without the tweet: An empirical examination of user versus document attributes. In: **Proceedings of the third workshop on natural language processing and computational social science**. [S.l.: s.n.], 2019. p. 18–28.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MOHAMMAD, S. et al. A dataset for detecting stance in tweets. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. [S.l.: s.n.], 2016. p. 3945–3952.

MOHAMMAD, S. et al. Semeval-2016 task 6: Detecting stance in tweets. In: **Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)**. [S.l.: s.n.], 2016. p. 31–41.

MOHAMMAD, S. M.; SOBHANI, P.; KIRITCHENKO, S. Stance and sentiment in tweets. **ACM Transactions on Internet Technology (TOIT)**, ACM New York, NY, USA, v. 17, n. 3, p. 1–23, 2017.

MOHTARAMI, M. et al. Automatic stance detection using end-to-end memory networks. **arXiv preprint arXiv:1804.07581**, 2018.

MUENNIGHOFF, N. Sgpt: Gpt sentence embeddings for semantic search. **arXiv preprint arXiv:2202.08904**, 2022.

NGUYEN, D. et al. Computational sociolinguistics: A survey. **Computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 42, n. 3, p. 537–593, 2016.

PAMUNGKAS, E. W.; BASILE, V.; PATTI, V. Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. **arXiv preprint arXiv:1901.01911**, 2019.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.

PODDAR, L. et al. Predicting stances in twitter conversations for detecting veracity of rumors: A neural approach. In: IEEE. **2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)**. [S.l.], 2018. p. 65–72.

RADFORD, A. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RASHED, A. et al. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2021. v. 15, p. 537–548.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

ROBERTSON, S.; ZARAGOZA, H. et al. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends® in Information Retrieval**, Now Publishers, Inc., v. 3, n. 4, p. 333–389, 2009.

RUDER, S. **Neural transfer learning for natural language processing**. Thesis (PhD) — NUI Galway, 2019.

SÁENZ, C. A. C.; BECKER, K. Interpreting bert-based stance classification: a case study about the brazilian COVID vaccination. In: **Anais do 36th Brazilian Symposium on Databases, SBBD**. [S.l.]: SBC, 2021. p. 73–84.

SAMIH, Y.; DARWISH, K. A few topical tweets are enough for effective user stance detection. In: **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**. [S.l.: s.n.], 2021. p. 2637–2646.

SIDDIQUA, U. A.; CHY, A. N.; AONO, M. Stance detection on microblog focusing on syntactic tree representation. In: SPRINGER. **Data Mining and Big Data: Third International Conference, DMBD 2018, Shanghai, China, June 17–22, 2018, Proceedings 3**. [S.l.], 2018. p. 478–490.

SOBHANI, P.; MOHAMMAD, S.; KIRITCHENKO, S. Detecting stance in tweets and analyzing its interaction with sentiment. In: **Proceedings of the fifth joint conference on lexical and computational semantics**. [S.l.: s.n.], 2016. p. 159–169.

SONG, K. et al. Mpnet: Masked and permuted pre-training for language understanding. **Advances in Neural Information Processing Systems**, v. 33, p. 16857–16867, 2020.

SOUSA, A. M. de; BECKER, K. Pro/anti-vaxxers in Brazil: a temporal analysis of covid vaccination stance in Twitter. In: SBC. **Anais do IX Symposium on Knowledge Discovery, Mining and Learning**. [S.l.], 2021. p. 105–112.

SOUSA, A. M. de; BECKER, K. Comparando os posicionamentos a favor/contra a vacinação covid nos estados unidos da américa e no brasil. In: SBC. **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados**. [S.l.], 2022. p. 65–77.

SOUSA, A. M. de; BECKER, K. Understanding the covid vaccination stances in brazil: a temporal analysis using twitter data. **Journal of Information and Data Management**, v. 13, n. 6, 2022.

SOUSA, A. M. de; BECKER, K. Sssd: Leveraging pre-trained models and semantic search for semi-supervised stance detection. In: **Proceedings of Recent Advances in Natural Language Processing (RANLP)**. [S.l.: s.n.], 2023. p. 264–273.

SUN, L. et al. Learning stance classification with recurrent neural capsule network. In: SPRINGER. **Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8**. [S.l.], 2019. p. 277–289.

SUN, Q. et al. Stance detection via sentiment information and neural network model. **Frontiers of Computer Science**, Springer, v. 13, p. 127–138, 2019.

SUN, Q. et al. Stance detection with hierarchical attention network. In: **Proceedings of the 27th international conference on computational linguistics**. [S.l.: s.n.], 2018. p. 2399–2409.

SUN, Q. et al. Stance detection with a multi-target adversarial attention network. **ACM Transactions on Asian and Low-Resource Language Information Processing**, ACM New York, NY, v. 22, n. 2, p. 1–21, 2022.

TAULÉ, M. et al. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In: CEUR-WS. **CEUR Workshop Proceedings**. [S.l.], 2017. v. 1881, p. 157–177.

VASWANI, A. et al. **Attention Is All You Need**. 2017.

WANG, W. et al. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. **Advances in Neural Information Processing Systems**, v. 33, p. 5776–5788, 2020.

WANG, Y. et al. Generalizing from a few examples: A survey on few-shot learning. **ACM computing surveys (csur)**, ACM New York, NY, USA, v. 53, n. 3, p. 1–34, 2020.

WEI, P.; MAO, W.; CHEN, G. A topic-aware reinforced model for weakly supervised stance detection. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, n. 01, p. 7249–7256.

WEI, P.; MAO, W.; ZENG, D. A target-guided neural memory model for stance detection in twitter. In: IEEE. **2018 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2018. p. 1–8.

WEI, P.; XU, N.; MAO, W. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. **arXiv preprint arXiv:1909.08211**, 2019.

XU, C. et al. Cross-target stance classification with self-attention networks. In: GUREVYCH, I.; MIYAO, Y. (Ed.). **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 778–783. Available from Internet: <https://aclanthology.org/P18-2123>.

XU, R. et al. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In: SPRINGER. **Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24**. [S.l.], 2016. p. 907–916.

YANG, Y. et al. Tweet stance detection: A two-stage dc-bilstm model based on semantic attention. In: IEEE. **2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)**. [S.l.], 2020. p. 22–29.

ZHANG, B. et al. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2020. p. 3188–3197.

ZHANG, Y.; JIN, R.; ZHOU, Z.-H. Understanding bag-of-words model: a statistical framework. **International journal of machine learning and cybernetics**, Springer, v. 1, p. 43–52, 2010.

ZHAO, G.; YANG, P. Pretrained embeddings for stance detection with hierarchical capsule network on social media. **ACM Transactions on Information Systems (TOIS)**, ACM New York, NY, USA, v. 39, n. 1, p. 1–32, 2021.

ZHOU, Y.; CRISTEA, A. I.; SHI, L. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In: SPRINGER. **Web Information Systems Engineering–WISE 2017: 18th International Conference, Puschino, Russia, October 7-11, 2017, Proceedings, Part I 18**. [S.l.], 2017. p. 18–32.

ZHU, L.; HE, Y.; ZHOU, D. Neural opinion dynamics model for the prediction of user-level stance dynamics. **Information Processing & Management**, Elsevier, v. 57, n. 2, p. 102031, 2020.

ZHUANG, L. et al. A robustly optimized BERT pre-training approach with post-training. In: LI, S. et al. (Ed.). **Proceedings of the 20th Chinese National Conference on Computational Linguistics**. Huhhot, China: Chinese Information Processing Society of China, 2021. p. 1218–1227. Available from Internet: <https://aclanthology.org/2021.ccl-1.108>.

# APÊNDICE A — INTERPRETAÇÃO GLOBAL DOS TÓPICOS CASO DE USO VACINAÇÃO

As Tabelas A.1 e A.2 apresentam uma síntese dos tópicos identificados conforme a metodologia estabelecida na Seção 6.1, ordenados em função do volume de *tweets*. A caracterização de cada tópico abrange o número específico de *tweets*, os cinco termos que mais se destacam (sejam eles unigramas ou bigramas), com base na métrica c-TF-IDF, além de três argumentos centrais. Devido à natureza do BERTopic, que se fundamenta em técnicas de agrupamento por densidade, *tweets* que se situam em regiões menos densas são categorizados como ruído, não sendo assim associados diretamente aos tópicos em evidência.

Tabela A.1: Pro-Vaxxers: Representação global dos Tópicos

| Tópico | # Tweets | Termos Pro-Vaxxers | Argumentos Representativos |
|---|---|---|---|
| 0 | 9185 | brasil, brasileiro, brasileiros, não, mortes | "O Brasil merece isso! Pobres brasileiros abandonados, precisam ser salvos desses alienados!"; "Só as vacinas salvam. Quantas vidas teriam sido salvas se as vacinas tivessem sido compradas mais cedo?"; "Que vergonha, mil mortes até o final da semana." |
| 1 | 7486 | brasil, vacina, brasileiros, vacinação, não | "Tão emocionante o discurso da enfermeira Monica."; "Tenho tanto orgulho de nossos enfermeiros, viva o SUS!"; "O primeiro brasileiro vacinado!" |
| 2 | 6338 | on, life goes, goes on, goes, defenda vacinas | "Eu amo ouvir Life Goes On no rádio! Toca de novo! Obrigado, defendam a vacinação."; "Kim Taehyung defende as vacinas, nos mantenha informados."; "Caramba, essa vacina é tão boa que silenciou o Bolsonaro." |
| 3 | 4829 | defenda vacinas, defenda, vacinas, vacina, mantenha informado | "Defenda a vacinação, espalhe informação."; "Defenda a vacinação kim taehyung."; "Parem de nos impedir de conseguir a vacinação!" |
| 4 | 4554 | coronavac, coronavírus, emergência, vacina, dose | "Que dia para o Brasil com a aprovação da Coronavac pela Anvisa."; "Aprovem a Oxford também. Vamos lá, Brasil!"; "Completamente vacinado com a segunda dose, gratidão."; "A covid recua no mundo, mas não no Brasil. Tragam mais vacinas governo incompetente e criminoso!" |
| 5 | 3901 | viva, ciência, viva a ciência, sus, viva o sus | "Viva o sus, viva a ciência"; "Que dia! Viva o sus, viva a ciência, f**** o Bolsonaro"; "Oficialmente um alien: vacinado! Abençoado seja o sus, viva a ciência!" |
| 6 | 3288 | dose, primeira dose, dose da vacina, segunda, primeira | "Minha querida mãe foi vacinada."; "Mãe, eu recebi minha primeira dose hoje!"; "Meu pai recebeu a primeira dose hoje, e a mãe vai recebê-la na segunda-feira." |
| 7 | 2410 | professores, aulas, educação, escolas, não, | "Todos tão animados! Em breve nossas crianças também estarão felizes e de volta à escola."; "De volta às aulas presenciais, nossas crianças merecem isso!"; "Todo o meu apoio aos professores! Eu elogio São Paulo por priorizar a vacinação desses anjos guardiães." |

Continua na próxima página.

Tabela A.1: Pro-vaxxers: Representação global dos Tópicos

| Tópico | # Tweets | Termos Pro-Vaxxers | Argumentos Representativos |
|---|---|---|---|
| 8 | 2361 | pandemia, descontrolado, responsabilidade social, crítica responsável, pandemia descontrolada | "Como reabrir as escolas com essa pandemia controlada?"; "Compartilhem, todos contra esse projeto, responsabilidade social!"; "Precisamos controlar a pandemia antes de podermos retornar com segurança às escolas." |
| 9 | 1961 | presidente, não, vacina, presidência, já | "Pazuello, precisamos de esperança. Presidente de m****"; "Idolatrar o presidente e ansiar pela vacina: incompatíveis!"; "Precisamos de um Presidente, não de um youtuber patético!" |

Os "Pro-Vaxxers" estão empenhados em promover a campanha de vacinação e criticar as políticas adotadas pelo Governo Federal para gerenciar a pandemia. No Tópico 0, fica claro que os brasileiros estão muito ansiosos para se vacinare e culpam o Governo Federal pela falta de vacinas. Os Tópicos 1 e 4 revelam apoio à vacinação, alegria com a aprovação das vacinas pela Agência Nacional de Vigilância Sanitária (Anvisa) e felicidade com a vacinação dos brasileiros. Elogios à ciência e ao Sistema Único de Saúde (SUS), alegria e gratidão por se vacinar ou ter seus entes queridos protegidos, são os argumentos centrais nos Tópicos 5 e 6.

Para compreender os Tópicos 7 e 8, foram analisadas amplas amostras de *tweets*, identificando-se que representam argumentações distintas acerca da educação. O Tópico 7 ressalta a necessidade de priorizar os professores na vacinação, possibilitando assim a retomada segura das atividades educacionais presenciais. Em contraste, o Tópico 8 manifesta resistência ao Projeto de Lei 5595/20[1], que propõe reconhecer os setores de educação básica e superior como serviços essenciais, pressionando pelo retorno das aulas presenciais independentemente da vacinação completa dos profissionais da educação.

O argumento central do Tópico 9 são críticas e ofensas ao Presidente, ao Ministro da Saúde, ao governo federal e aos apoiadores do Presidente. Esse argumento é encontrado em outros tópicos também. Um número significativo de *tweets* envolve engajamento artificial para manter o movimento pró-vacinação em evidência. No Tópico 2, o termo "Life Goes On" corresponde a pedidos pela música do artista Oliver Tree. Na mesma direção, no Tópico 3, existem termos associados ao cantor e compositor sul-coreano Kim Tae-Hyung, mais conhecido em sua carreira musical pelo nome artístico V.

[1] https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2267745

Tabela A.2: Anti-vaxxers: Representação global dos Tópicos

| Tópico | # Tweets | Termos Anti-Vaxxers | Argumentos Representativos |
|---|---|---|---|
| 0 | 5361 | vacina, tomar vacina, tomar, não, ir | “Tão bom que você não vai se vacinar! mais uma dose para quem merece!”; “Estou feliz que você não vai tomar sua dose, eu vou pegar a minha mais cedo!”; “Não se vacine, enfie a cloroquina no seu @bo!” |
| 1 | 3251 | yuri, é, está chegando, chegando, gomes | “Gandalf Melody Yuri contra todos os Santos sos está chegando Luiz Fernando Diego Souza Atlético Copa Brasil Igor Gomes Rabello Everson Thaciano Yuri”; “Eu não quero, não vou!”; “Ótimo, bolsonaristas não vão tomar: haverá mais vacinas para nós!” |
| 2 | 869 | brasil, brasileiro, pessoas, não, brasileiros | “Parabéns população destemida Búzios ditadores não podem ser obedecidos zombados até que desapareçam espalhem iniciativa Brasil.”; “Prisão de Ditadores Comunistas Brasileiros!”; “STF envia mensagem que queriam dar ao Brasil!” |
| 3 | 654 | doutor, tratamento, tratamento médico, intervenção cirúrgica, submeter envergonhado | “Copie e compartilhe esta arte da constituição - ninguém pode ser obrigado a qualquer tratamento médico ou intervenção cirúrgica.”; “É a lei - ninguém pode ser obrigado a qualquer tratamento médico ou intervenção cirúrgica, meu corpo minhas regras!”; “Aprovação da Anvisa é resultado de corrupção.” |
| 4 | 534 | china, vacina, chinês, vacina chinesa, não | “Confiar na OMS? Confiar na vacina chinesa? De jeito nenhum!”; “Pazuello vai nos comprar vacinas reais.”; “Ditador Dóriá vai nos forçar a nos vacinar, Absurdo. Cidadãos de São Paulo não querem a vacina chinesa!” |
| 5 | 409 | gado, não, vacina de gado, tomar | “Triste em saber que o gado não vai se vacinar.”; “O gado não vai pegar vacina: a fila será menor para nós!”; “Veja o lado bom: não haverá mais manada de gado!” |
| 6 | 384 | senadores, vereadores, deputados, senadores deputados, vereadores senadores | “Governadores prefeitos vereadores senadores deputados. Onde estão os políticos que nos representam? Deitados em um berço esplêndido. Revolta”; “Ministros do STF, não é esse o seu papel!”; “Senadores deputados vereadores, façam alguma coisa!” |
| 7 | 346 | tag, twitter, assistindo, apenas, fazendo login | “Essas pessoas insanas continuam levantando tags.”; “Levantando hashtags. O gado do Bolsonaro está tuitando loucamente.”; “O gado está representado nesta hashtag insana.” |
| 8 | 326 | brasil, vacina, não, brasileiro, brasileiros | “O voluntário que testou a vacina cometeu suicídio. Você acha que os brasileiros são estúpidos?”; “Vamos nos rebelar contra a vacina, o Brasil vai fazer uma decente!”; “Ninguém deve ser obrigado a tomar uma vacina. Os brasileiros não são ignorantes. Evolua, pessoas estúpidas!” |
| 9 | 321 | china, chinês, não, chinês, vachina | “Qualquer nacionalidade, especialmente chinesa!”; “Ele quer vender SP para a China.”; “Capacho da China.” |

A Tabela A.2 resume os tópicos identificados para o movimento “Anti-Vaxxers”. Eles são contra a vacinação obrigatória, questionam a segurança da vacina e focam em

disputas políticas.

As preocupações contra a vacinação obrigatória são principalmente representadas nos Tópicos 2, 6 e 8. A vacinação obrigatória foi defendida por muitos governadores, incluindo o Governador João Dória, como meio de combater o cenário pandêmico. Em geral, há muitas críticas ao Supremo Tribunal Federal por julgar a vacinação obrigatória como constitucional devido a razões sanitárias no contexto da pandemia. As pessoas também expressam forte resistência, cobrando uma posição dos representantes eleitos e convocando os brasileiros a irem às ruas protestar contra políticos e autoridades que apoiam a vacina obrigatória.

O Tópico 3 é dedicado a conscientizar sobre a inconstitucionalidade da vacina obrigatória, onde o artigo 15 do Código Civil[2] ("Ninguém pode ser compelido a submeter-se a tratamento médico ou intervenção cirúrgica com risco de vida") é usado para promover essa ideia.

Os Tópicos 4 e 9 são caracterizados por uma sólida resistência à Coronavac, resultante da parceria entre o laboratório chinês Sinovac e o Instituto Butantan brasileiro. As pessoas expressam preconceito e desconfiança devido à sua suposta "origem chinesa" e críticas a João Dória, que alguns alegam ter explorado a produção desta vacina para fins políticos.

No Tópico 1, existe o uso de engajamento artificial para fortalecer os interesses do movimento "Anti-Vaxxers". Os *tweets* não têm relação com a vacinação e referem-se a atores, jogadores de futebol e equipes, cantores, programas de televisão ou eventos importantes.

Um fenômeno interessante é o confronto evidente dos "Pro-Vaxxers" aos *tweets* postados pelos "Anti-Vaxxers" (falso-negativos). Mesmo seguindo uma inspeção minuciosa das *hashtags*, como detalhado na Seção 6.1.1, essa dinâmica é frequentemente observada devido à presença maciça de respostas nas quais os indivíduos favoráveis à vacinação contestam as publicações ou *hashtags* promovidas pelos grupos anti-vacina. Nos Tópicos 0 e 5, por exemplo, observam-se comentários sarcásticos dos "Pro-Vaxxers", ironizando a maior disponibilidade de vacinas resultante da hesitação dos "Anti-Vaxxers". O Tópico 7, especificamente, revela críticas à tentativa dos "Anti-Vaxxers" de engajar *hashtags* contrárias à vacinação. Em muitos casos, são empregados comentários depreciativos, associando os "Anti-Vaxxers" a apoiadores do atual Presidente. É importante ressaltar que, embora existam *tweets* dentro desses tópicos de indivíduos resistentes à va-

[2]http://www.planalto.gov.br/ccivil_03/leis/l3071impressao.htm

cinação, a narrativa predominante é de confronto. No contexto dos “Pro-Vaxxers”, esse padrão de comportamento é igualmente presente, porém, tende a ser menos perceptível devido ao alto volume de *tweets*, reduzindo assim sua proeminência.

Em conclusão, os “Pro-Vaxxers” estão alinhados com a tradição do Brasil de PNI para manter a saúde da população, com pessoas expressando alegria ou expectativas sobre a vacinação e criticando todas as pessoas que não estão dispostas a se vacinar. Os “Anti-Vaxxers” estão preocupados com a segurança da vacina e a vacinação obrigatória. Ambos os lados expressam suas posturas entrelaçadas com comentários políticos que criticam o governo pela falta de um PNI ou endossam as ações do Presidente e do governo federal. Além disso, ambos os lados tentam envolver a população em suas posturas através de engajaento artificial.

# APÊNDICE B — INSTRUÇÕES PARA ROTULAGEM DE TWEETS

Os *tweets* a serem anotados foram coletadas usando *hashtags* “Pro-Vaxxers” (i.e. #VouTomarVacina) ou “Anti-Vaxxers” (e.g. #NaoVouTomarVacina). Eles estão distribuídos nas abas “antivaxxer” (949) e “provaxxer” (495). Cada coluna corresponde as seguintes informações:

- **texto:** postagem em formato texto bruto.
- **rotulo:** rótulo de posicionamento.

Contudo, nem sempre o texto representa o posicionamento inferido pela *hashtag*, principalmente pelo viés político embutido nas manifestações. No caso dos “Pro-Vaxxers” existem críticas ao governo federal/presidente. No caso dos “Anti-Vaxxers”, há apoio ao governo federal/presidente, e resistência a todas instituições que se manifestam a favor da vacinação obrigatória (governadores, STF, etc).

A anotação tem por objetivo ajudar a identificar os casos com problema. A tarefa é ler cada *tweet* e anotar o posicionamento quando não concordar com o rótulo sugerido pela aba.

- **F (Favor):** O texto do *tweet* permite compreender que a pessoa é a favor da vacinação.
- **C (Contra):** O texto do *tweet* permite compreender que a pessoa é contra a vacinação.
- **N (Nenhum):** O texto do *tweet* não expressa claramente posicionamento contra/a favor, ou não tem relação com o alvo.

Na aba “antivaxxer”, quando não concordar, marque a coluna “rótulo” com N (neutro) ou F (favorável). Se concordar, deixe em branco.Na aba “provaxxer”, quando não concordar, marque a coluna “rótulo” com N (nenhum) ou C (contra). Se concordar, deixe em branco.

**Exemplos de *tweets* com posicionamento favorável (rótulo = F) a vacinação:**

- Gratidão enorme pela Mônica Calazans e por todos que trabalham na Saúde Pública, e lutam por ela Finalmente a vacina vai vir pro Ceará.
- Só sei que hoje minha avó tomou a 2° dose da vacina!
- O plano é deixar bolsominion sem tomar vacina pq aí sobra menos deles!

**Exemplos de *tweets* com posicionamento contra (rótulo = C) a vacinação:**

- Não tomo. E tem mais, negacionista são os que não aceitam tratamento precoce!
- Eu tive a praga chinesa e fui curada com ivermectina. 19 membros de minha família

usaram o mesmo medicamento e estão ótimos, sem sequelas.

- Pois é, não sou COBAIA de vacina experimental emergencial.
- Só digo uma coisa, obrigatória ou não, eu não tomo essa vacina feita a toque de caixa. Vacina obrigatório na @nda do Dória!

**Exemplos de *tweets* com posicionamento neutro (rótulo = N) em relação a vacinação:**

- I will do shopify store redesign, dropshipping shopify store design,website design Kim Namjoon watermelon sugar Meninas Malvadas Adele harrys Luiz Fernando Renat Rodrigo Hilbert James Rodrigues MIRELLA MERECE RESPEITO.
- Eu tive a praga chinesa e fui curada com ivermectina. 19 membros de minha família usaram o mesmo medicamento e estão ótimos, sem sequelas.
- O hómi não tá falando coisa com coisa.
- Tipo “um certo senhor ali” c/ seus seguidores nessa nova falácia.
- Cara@@@ como eu odeio brasileiro! Poderia cair um meteoro agora no Brasil!