UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RODRIGO LUDERMIR DE OLIVEIRA

Detecção de Posicionamento em *Tweets* sobre Covid-19 no Brasil utilizando métodos de Aprendizagem de Máquina

Recife

2022

RODRIGO LUDERMIR DE OLIVEIRA

Detecção de Posicionamento em *Tweets* sobre Covid-19 no Brasil utilizando métodos de Aprendizagem de Máquina

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Cleber Zanchettin

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

O48d Oliveira, Rodrigo Ludermir de
Detecção de posicionamento em *tweets* sobre Covid-19 no Brasil utilizando métodos de aprendizagem de máquina / Rodrigo Ludermir de Oliveira – 2022.
97 f.: il., fig., tab.

Orientador: Cleber Zanchettin.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências.

1. Inteligência computacional. 2. Detecção de posicionamento. 3. *Tweets*. 4. Covid-19. 5. Aprendizado de máquina. I. Zanchettin, Cleber (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2023 – 75

RODRIGO LUDERMIR DE OLIVEIRA

"**Detecção de Posicionamento em *Tweets* sobre Covid-19 no Brasil utilizando métodos de Aprendizagem de Máquina**"

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 10/03/2022.

Orientador: Cleber Zanchettin

BANCA EXAMINADORA

Prof. Dr. Adriano Lorena Inacio de Oliveira

Centro de Informática / UFPE

Prof. Dr. Renato Vimieiro

Departamento de Ciência da Computação / UFMG

Prof. Dr. Cleber Zanchettin

Centro de Informática / UFPE

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que, direta ou indiretamente, contribuiram para esse trabalho seja com seu conhecimento técnico ou com seu apoio afetivo.

Especialemnte ao professor Cleber Zanchettin, por ter acreditado e confiado em mim para o desenvolvimento desta pesquisa.

Aos professores que durante minha vida acadêmica me ensinaram muito.

Ao CNPq pelo apoio financeiro.

RESUMO

A onipresença da pandemia de Covid-19 durante os últimos dois anos acarretou na urgência de ações responsivas contra o avanço da contaminação do novo coronavírus e em estratégias de imunização da população, através de políticas de saúde pública e medidas sanitárias preventivas por parte das autoridades responsáveis e também da sociedade civil. No Brasil, esse processo foi profundamente politizado, suscitando discussões polarizadas que inundaram as redes sociais com opiniões e posicionamentos acerca das medidas adotadas contra a Covid-19 e suas repercussões. Enquanto um paradigma emergente no campo de mineração de opiniões nas redes sociais, sistemas de detecção de posicionamento têm produzido resultados frutíferos, principalmente quando os objetos de classificação estão segmentados por um tópico alvo sobre o qual o posicionamento é realizado. Desse modo, esta dissertação investiga a utilização de métodos de aprendizagem de máquina no desenvolvimento de sistemas de detecção de posicionamento em *tweets* - publicações na rede social *Twitter* - de usuários brasileiros comentando as medidas relacionadas à Covid-19, exercidas por eles próprios e pelo governo brasileiro em seus diferentes órgãos e níveis de atuação. O trabalho envolve três partes principais: (1) Construção da base de dados, na qual houve o levantamento de mais de 6 milhões de *tweets* e *retweets* em português que mencionam palavras relacionadas à Covid-19 entre Janeiro de 2020 e Outubro de 2021, das quais mais de 350 mil *tweets* foram rotulados (*pseudo-labels*), através de métodos de anotação fraca (*weak supervision*), em "favoráveis" ou "contrários" às medidas do governo federal frente à pandemia. (2) Limpeza, análise exploratória e segmentação da base rotulada por tópicos mais relevantes e frequentes. (3) Avaliação de modelos de Aprendizagem de Máquina tradicionais e de aprendizagem profunda - sobretudo *Transformers*, na detecção de posicionamentos. Utilizando o modelo de linguagem de domínio geral em português-brasileiro BERTimbau, que segue a arquitetura base do BERT, foram realizados experimentos com: (1) adaptação de domínio, usando os dados não rotulados; (2) uso de dados relacionais dos usuários (rede de interações - *retweets*, *mentions* e *replies*); (3) Aprendizado via *Multi-tasking*, realizando o ajuste-fino em todos os tópicos ao mesmo tempo. Os experimentos realizados demonstraram que os modelos inicializados usando BERTimbau e treinados combinando as três abordagens citadas acima se sobressaem sobre os demais em seu desempenho diante da variedade de tópicos relacionados à Covid-19 no contexto brasileiro.

Palavras-chaves: detecção de posicionamento; *tweets*; covid-19; aprendizado de máquina.

ABSTRACT

The ubiquity of the Covid-19 pandemic has resulted in the urgency of responsive actions against the advance of the contamination of the new coronavirus and in strategies to immunize the population, through public health policies and preventive health measures by the authorities in charge and also civil society. In Brazil, this process was deeply politicized, giving rise to polarized discussions that overflowed social media with opinions, views, and stances regarding these measures taken against Covid-19 and its repercussions. As an emerging paradigm of opinion mining in social media, stance detection has yielded accurate and fruitful results, especially when classification objects are segmented by a target topic on which the stance positioning is performed. Thus, this dissertation investigates the use of machine learning methods in the development of stance detection systems in Tweets - publications on the social network Twitter - of Brazilian users commenting on measures related to Covid-19, taken by themselves and their government in its different bodies and levels of action. The work involves three main parts: (1) Construction of the database, in which there was a survey of more than 6 millions Tweets and Retweets in Portuguese that mention words related to Covid-19 between January 2020 and October 2021, of which more than 350,000 Tweets were labeled (pseudo-labels), through weak annotation methods (weak supervision), as 'favorable' or 'contrary' to the federal government's measures against the pandemic. (2) Cleaning, exploratory analysis and segmentation of the base labeled by the most relevant and frequent topics. (3) Evaluation of traditional and deep learning Machine Learning models - especially Transformers, in stance detection. Using the Brazilian-Portuguese domain-general language model BERTimbau experiments were carried out with: (1) domain adaptation, using unlabeled data; (2) use of users' relational data (interaction network - retweets, mentions and replies); (3) Learning via Multi-tasking, fine-tuning all topics at the same time.The experiments carried out showed that the models initialized using BERTimbau and trained by combining the three approaches mentioned above stand out from the others in their performance in the face of the variety of topics related to Covid-19 in the Brazilian context.

Keywords: stance detection; tweets; covid-19; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Na virada do século XIX para XX, o sociólogo Gabriel Tarde, teórico da opinião pública e da imitação, atencioso às transformações dos meios de comunicação em voga - mídia impressa e telégrafo - e à rápida propagação de ideias e opiniões que essas mudanças tinham como consequência, atestou ao sufrágio universal um valor latente e inexplorado: A decisão da população na escolha dos seus representantes era sobretudo um *"trabalho intermitente de estatística política pelo qual uma nação é chamada a tomar consciência das mudanças que se operam em seus desejos e suas opiniões sobre questões vitais [...], uma tarefa de ordem social: tomar consciência de si enquanto comunidade"* (TARDE, 1992).

A teoria tardeana da opinião, portanto, aponta que os indivíduos vão às urnas não apenas para decidir seus representantes, mas para se informar sobre as mudanças na sociedade, apontando o sufrágio enquanto um instrumento estatístico. A estatística desse período, para o lamento de Tarde, não dispunha de instrumentos técnicos capazes de fornecer muitas informações e a sociedade de sua época carecia de elementos que gozavam do "privilégio da mensurabilidade", tornando-os incapazes de perceber certas mudanças sociais e alterações cruciais em seu interior.

Para Tarde, é importante aperfeiçoar os instrumentos estatísticos porque eles são muito mais que um meio de observação: *"ela [a estatística] é capaz de demonstrar os efeitos favoráveis ou prejudiciais produzidos pela imitação de certas tendências e assim, influir sobre a propensão que os indivíduos teriam a seguir ou não estes ou aqueles exemplos. [...] A estatística tem a propensão de ser tornada pública, publicada a fim que a sociedade conheça a si mesma, saiba o que ela é e aquilo em que se transforma"*(TARDE, 1992).

Passados anos desde tais afirmações, os instrumentos estatísticos foram aperfeiçoados, assim como a ciência da opinião pública, a ciência da computação e os meios de comunicação. Com o processo de digitalização, as opiniões e as conversações obtêm novos contornos, registros, formas e substâncias, adquirindo também o "privilégio da mensurabilidade", impensável para os tempos de Tarde. No entanto, é impossível ignorar as consequências desastrosas do processo de digitalização e plataformização dos meios de comunicação, como desinformação, polarização e o surgimento de novos tipos de exploração, em parte gerada precisamente pela extensão da capacidade de mensurabilidade das nossas ações (será mesmo um privilégio?).

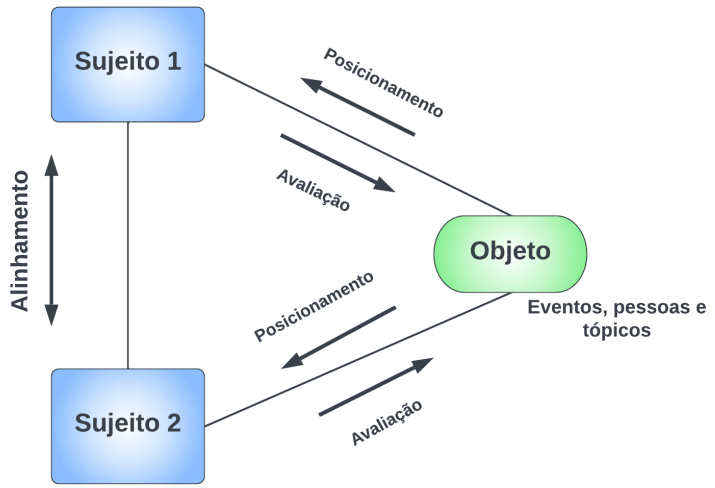
Desse modo, a proposta deste estudo, a de aperfeiçoar e implementar sistemas de aprendizado estatístico para detectar posicionamentos e opiniões durante a pandemia do coronavírus, além de buscar contribuições para a área de mineração de opiniões e aprendizado de máquina, tem como referência as declarações de Tarde: A nossa pesquisa também almeja colaborar, ao longo prazo, com seus experimentos e resultados, e mesmo que em doses mínimas e incipientes, numa tarefa de ordem social, para que "a sociedade [brasileira] conheça a si mesma, saiba o que ela é e aquilo em que se transforma".

1.2 DETECÇÃO DE POSICIONAMENTO

Atualmente as redes sociais constituem um espaço importante de interações sociais e são consideradas ferramentas que expandem a capacidade e velocidade de disseminação de informações, além de serem ambientes propícios para expressar, compartilhar e consumir opiniões e pontos de vista sobre qualquer assunto de interesse público ou pessoal. Desse modo, muitos indivíduos utilizam essas ferramentas como a principal fonte de notícias, já que é possível obter atualizações instantâneas sobre os últimos acontecimentos, e um meio crucial para se conectar com o mundo. Nesses espaços virtuais, os usuários podem dialogar, ao expressar seus próprios pontos de vista, sobre os vários tópicos que emergem interna ou externamente a esse ambiente e observar a percepção do público acerca desses mesmos tópicos, ao receber *feedbacks* instantâneos dos outros usuários. A dependência e ampla utilização dessas plataformas como principal fonte de comunicação na sociedade digital, permite aos pesquisadores estudar diferentes aspectos do comportamento humano online, entre eles, o posicionamento público em relação a várias questões sociais e políticas.

O posicionamento ou posição pode ser definido como a expressão do ponto de vista ou julgamento do interlocutor em relação a uma determinada proposição (BIBER; CONRAD; REPPEN, 1998). Este processo de formulação do posicionamento é um fenômeno subjetivo e intersubjetivo, afetado por fatores pessoais e não-pessoais, como normas culturais e aspectos sociais (DU-BOIS, 2007).

Figura 1 – Reprodução do Triângulo do Posicionamento de Du Bois



Fonte: Autor

A Figura 1 apresenta o triângulo de posicionamento formulado por Du-Bois (2007), e demonstra como o processo de formulação do posicionamento é baseado em três fatores principais: (1) avaliação de objetos (eventos, pessoas e tópicos no geral); (2) posicionamento do sujeito principal; (3) alinhamento do posicionamento com outros sujeitos. Um usuário nas redes sociais, portanto, pode se posicionar expressando sua opinião diretamente em relação a um alvo/objeto, de modo explícito ou implícito, em uma publicação. Mas também é necessário considerar que, conforme explica Du-Bois, o posicionamento também é formulado em alinhamento com outros sujeitos e assim, também pode ser inferido indiretamente através das interações de um usuário com os outros usuários da rede.

Além disso, de uma perspectiva sociolinguística (JAFFE et al., 2009), tem-se argumentado que não existe um posicionamento completamente neutro, pois as pessoas tendem a se posicionar por meio de seus textos a favor ou contra o objeto de avaliação. Isso cria mais uma complexidade na identificação do posicionamento dos indivíduos, já que estes nem sempre constam de modo transparente no texto, mas às vezes deve ser inferida implicitamente a partir de uma combinação de interações sociais e contexto histórico.

O processo de detecção de posicionamento também pode ser conhecido como detecção de perspectiva (BEIGMAN-KLEBANOV; BEIGMAN; DIERMEIER, 2010) e detecção de ponto de vista (TRABELSI; ZAIANE, 2018); (ZHU; HE; ZHOU, 2019), nos quais opiniões são identificadas ao expressar posições em relação a um tema controverso (MOHAMMAD; SOBHANI; KIRITCHENKO, 2017). Dentro do campo de mineração de opiniões é importante diferenciar essa tarefa da

análise de sentimentos, já que a última pode ser aplicada a uma sentença sem nenhum alvo específico, enquanto a detecção de posicionamento geralmente exige um determinado alvo para que seja possível medir o ponto de vista do usuário em relação a ele (ALDAYEL; MAGDY, 2021).

Os primeiros estudos de detecção de posicionamento tinham como fonte de opiniões os debates estabelecidos em fóruns online, que eram predominantes antes da plataformização da internet, ambiente onde os usuários debatiam em volta de tópicos específicos, com um fluxo de informações delimitado por esses temas (BELKAROUI; FAIZ; ELKHLIFI, 2014). Os trabalhos mais recentes, em contraste, concentram-se nas perspectivas que circulam nas redes sociais, especialmente o *Twitter*, onde as discussões sobre um determinado tópico são mais dispersas, mas que podem ser vinculadas e concentradas através de *hashtags* (MOHAMMAD et al., 2016). Os dados coletados em redes sociais como *Twitter*, no entanto, possuem atributos adicionais que os tornam muito mais ricos, como rede de interações entre usuários, utilizadas de modo recorrente no desenvolvimento de técnicas para detecção de posicionamento, que não dependem apenas dos atributos textuais das opiniões. Desse modo, considerando que o *Twitter* tornou-se um lugar comum para os usuários compartilharem suas opiniões acerca de vários assuntos, as pesquisas na automação de classificação de posicionamento foram direcionadas a essa plataforma. Algumas das aplicações que podem fazer uso da classificação de posicionamento são: estudo da evolução das opiniões, monitoramento do debate político, detecção de polarização e de boatos (ALDAYEL; MAGDY, 2019).

A detecção de posicionamento desempenha um papel importante em estudos que objetivam mensurar e monitorar a opinião pública nas mídias sociais, particularmente quando tratam-se de assuntos políticos e sociais, levando em consideração que é característico dessas questões a controvérsia e o conflito de ideias, nas quais as pessoas expressam opiniões opostas. Questões socio-políticas como aborto, mudanças climáticas e feminismo já foram utilizadas como tópicos-alvo para detecção de posicionamento nas redes sociais (MOHAMMAD et al., 2016), assim como tópicos nomeadamente políticos, como referendos e eleições (FRAISIER et al., 2018). Pesquisas na área de detecção de posicionamento têm rendido resultados precisos e calibrados, especialmente em tópicos políticos e sociais, mas em sua maioria no contexto americano e na língua inglesa. Além disso, ainda são poucas as pesquisas para tópicos relativos às medidas sanitárias contra a Covid-19, tema devido ao caráter recente da pandemia e por esse motivos faltam base de dados benchmarks nesse tema (GLANDT et al., 2021).

Recentemente, alguns estudos realizaram uma análise geral das pesquisas no campo de

detecção de posicionamento em redes sociais, indicando lacunas a serem investigadas por trabalhos futuros ((ALDAYEL; MAGDY, 2021); (GHOSH et al., 2019); (KÜÇÜK; CAN, 2020); (WANG et al., 2019)).

O estudo de (WANG et al., 2019) destacou os métodos existentes para mineração de opinião de clientes acerca de produtos. Apesar de discutir os métodos de detecção de posicionamento, o principal interesse do trabalho concentrou-se em métodos que utilizam recursos textuais das postagens em redes sociais e na aplicação de sistemas de detecção de posicionamento no contexto de produtos. O segundo estudo, de (GHOSH et al., 2019), fornece uma análise comparativa de modelos de detecção de posicionamento, investigando a reprodutibilidade desses modelos usando dois conjuntos de dados: SemEval 2016 stance dataset (MOHAMMAD et al., 2016) e conjunto de artigos de notícias online (SEN et al., 2018). Este trabalho indica que o uso de Modelos de Linguagens com aprendizagem por transferência como BERT fornecem as melhores pontuações nos dois conjuntos de dados utilizados em comparação com outros modelos de classificação. Embora esse estudo comparativo seja de grande utilidade na comparação de várias abordagens de detecção de posicionamento, eles se concentraram apenas em técnicas de Processamento de Linguagem Natural (PLN), isto é, usando apenas os recursos textuais desses conjuntos de dados. Da mesma forma, o terceiro estudo, de (KÜÇÜK; CAN, 2020), investigou sistemas de detecção de posicionamento focando principalmente nas técnicas de PLN e nos diferentes aspectos da mineração de opiniões e da detecção de posicionamentos, como sarcasmo, controvérsia e emoções.

O estudo realizado por (ALDAYEL; MAGDY, 2021), principal referência desta dissertação no que concerne às pesquisas sobre detecção de posicionamentos, investiga os resultados do estado da arte nas bases de dados benchmarks e discute as abordagens mais efetivas e eficazes no campo de detecção de posicionamento em redes sociais. Também exploraram as tendências emergentes, as diferentes aplicações existentes e concluem o estudo elencando as principais lacunas nas pesquisas realizadas, indicando possíveis direções para pesquisas futuras em detecção de posicionamento. Algumas dessas lacunas serão tratadas na próxima seção, como motivações que podem guiar o estudo realizado nesta dissertação.

1.3 MOTIVAÇÕES

Neste trabalho estudaremos um paradigma emergente no campo da mineração de opiniões, nominalmente, sistemas de detecção de posicionamento, que usaremos aqui para analisar e,

em suma, classificar automaticamente as posições de usuários em discussões nas redes sociais sobre as medidas políticas e sanitárias contra a Covid-19 no Brasil.

No curto prazo, dadas as circunstâncias atuais, estudar e desenvolver abordagens automatizadas para detecção de posicionamentos da população nas redes sociais em relação a determinados tópicos, como as medidas políticas e sanitárias em relação à Covid-19 - diga-se, vacinação, uso de máscaras, isolamento social, entre outras - podem auxiliar governos, autoridades sanitárias e a própria sociedade civil e suas organizações a tomarem medidas apropriadas para mitigar os efeitos da pandemia (GLANDT et al., 2021).

Em relação ao campo de estudos de detecção de posicionamentos, algumas das lacunas apontadas por ALDayel e Magdy (ALDAYEL; MAGDY, 2021) servem como diretrizes para o desenvolvimento deste trabalho. Os autores apontam a necessidade de novos conjuntos de dados para a tarefa de detecção de posicionamento, capazes de providenciar uma quantidade suficiente de dados para treinar e testar os modelos existentes. Desse modo, esta tarefa de classificação poderá ser explorada em novos domínios e línguas. O nosso estudo, nessa direção, propõe um novo conjunto de dados de publicações (*tweets*) em português de usuários brasileiros que se posicionam acerca das várias medidas de enfrentamento à pandemia no país. Essa base servirá para explorar a tarefa de detecção de posicionamentos em um novo domínio e língua, haja vista a escassez de conjunto de dados rotulados para essa tarefa de classificação em português que têm como tópico principal as medidas relacionadas à pandemia.

Além disso, ao declarar que as *features* relacionais (atributos sociais) são mais precisas no desempenho geral dos modelos em comparação com as *features* textuais, ALDayel e Magdy indicam, através de um survey do estado da arte de detecção de posicionamentos, que o primeiro tipo de *features* são mais custosas e pouco práticas, principalmente na coleta dessas informações. Há, portanto, uma necessidade de investigação em como alcançar modelos mais precisos (eficazes), com menos recursos (eficientes), utilizando atributos sociais e relacionais. Em vista disso, este trabalho também pretende integrar os atributos relacionais dos *tweets*, diga-se usuários mencionados (*mentions*), usuários retuitados (*retweets*) e os usuários os quais a publicação responde (*replies*) e extrair destas *features* um grafo de relacionamentos que auxilie o sistema de detecção a enriquecer a representação dos atributos textuais das publicações, tornando-o mais eficaz de modo eficiente.

Nas pesquisas publicadas que utilizam sistemas de aprendizagem de máquina para a detecção de posicionamentos em redes sociais (ALDAYEL; MAGDY, 2021), uma miríade de modelos supervisionados são experimentados, tais quais *Support Vector Machines* (SVM) (VAPNIK,

1999), Árvores de Decisão (FACELI et al., 2011) e algumas abordagens de Aprendizado Profundo (GOODFELLOW; BENGIO; COURVILLE, 2016), como *Recurrent Neural Network* (RNN) e *Long short-term memory* (LSTM).

Apenas mais recentemente arquiteturas baseadas em *Transformers* e *Transfer Learning* foram exploradas para detecção de posicionamento, como em Ghosh et al. (2019), Giorgioni et al. (2020) e Kawintiranon e Singh (2021). Essas pesquisas demonstram como o uso do BERT Devlin et al. (2018) e suas variações linguísticas como UmBERTo e de domínio como BERTweet (NGUYEN; VU; NGUYEN, 2020), além da possibilidade de adaptação de domínio e outras metodologias de *transfer learning*, oferecem desempenhos mais eficazes em bases como o benchmark SemEval 2016 *stance dataset* (MOHAMMAD et al., 2016), em inglês, e no EVALITA 2020 (CIGNARELLA et al., 2020), em italiano.

Além das experimentações com as metodologias do estado da arte de PLN, diante do contexto global pandemia de Covid-19, o campo de detecção de posicionamento também recebeu contribuições no que diz respeito a este tópico de discussão. Isso ocorreu através da construção de novas bases de dados com posicionamentos em redes sociais acerca de algumas das medidas adotadas para o enfrentamento da Covid-19, que serão analisadas de modo mais significativo na próxima seção, e por meio de estudos com análise exploratórias e realização de experimentos para detecção de posicionamento nestas bases ((HOSSAIN et al., 2020a); (MUTLU et al., 2020); (MIAO; LAST; LITVAK, 2020); (GLANDT et al., 2021)). No entanto, todas essas pesquisas foram realizadas para detectar posicionamentos em textos em inglês e os tópicos sobre os quais os posicionamentos são realizados limitam-se ao contexto americano. Por esse motivo, convém transpor essa análise para o enfrentamento da pandemia no contexto brasileiro e com posicionamentos em textos em português.

Neste trabalho procuramos investigar técnicas supervisionadas de classificação do estado da arte que possam contribuir para o aprimoramento de sistemas de detecção de posicionamento em redes sociais, a serem utilizadas em publicações em português brasileiro no *Twitter*. São abordadas técnicas já consagradas na literatura em PLN e na tarefa de detecção de posicionamentos em redes sociais, mas que ainda não foram experimentadas para o contexto brasileiro quando se refere a essa tarefa de detecção de posicionamento. Modelos baseados em *Transformers*, como BERTweet, pré-treinado em milhões de *tweets* em inglês e RoBERTa (LIU et al., 2019b), versão otimizada do BERT estão presentes no topo do *leaderboards* do TweetEval (BARBIERI et al., 2020) - Unified Benchmark and Comparative Evaluation for Tweet Classification, além de abordagens como o Retweet-BERT (JIANG; REN; FERRARA, 2021) que

incorpora a rede de *retweets* no treinamento do modelo, utilizando-se *features* textuais e dados relacionais (*retweets*), sendo um modelo de alto desempenho para estimar a polaridade política do usuário.

1.4 OBJETIVOS

Esse trabalho estuda os sistemas de detecção de posicionamento em redes sociais, as diversas abordagens propostas na literatura para esse problema, analisando suas vantagens e desvantagens, e as aplica ao contexto da pandemia no Brasil, realizando experimentos diante de um domínio e idioma diferentes, algo possível através da construção de nova uma base de dados em português e do treinamento de novos modelos de aprendizagem de máquina.

Portanto, dentre os objetivos específicos da pesquisa está a construção de uma nova base de dados de *tweets* em português com posicionamentos acerca das medidas políticas e sanitárias contra a Covid-19 no Brasil. Através de métodos baseados em regras e conhecimento de especialistas foi automaticamente realizada a rotulação dos *tweets* da base, além da segmentação desta em tópicos mais específicos relacionados à Covid-19, como vacinação, uso de máscaras, tratamento precoce, *lockdown*, governadores e prefeitos, e CPI da covid.

Além disso, objetiva-se neste estudo a validação dos resultados obtidos em trabalhos anteriores, que utilizam modelos baseados em *Transformers* na tarefa de detecção de posicionamento no *Twitter* (GHOSH et al., 2019), (GIORGIONI et al., 2020) e (KAWINTIRANON; SINGH, 2021), diante da teoria sobre o enfrentamento da pandemia. Também objetiva-se a experimentação de novas abordagens de modelagem, que incorporam representações de relações sociais nas redes, como o Retweet-BERT, proposta por Jiang, Ren e Ferrara (2021).

O principal objetivo da pesquisa é experimentar abordagens do estado da arte em PLN e detecção de posicionamento diante das opiniões sobre as medidas políticas e sanitárias contra a Covid-19 nas redes sociais, que englobam tanto a etapa de modelagem - seleção dos atributos, configuração e experimentação de modelos eficientes e robustos - quanto a etapa construção e preparação dos dados.

1.5 ESTRUTURA DA DISSERTAÇÃO

O restante deste documento está organizado da seguinte forma: o Capítulo 2 apresenta detalhadamente a elaboração da Base de Dados construída nesta dissertação, descrevendo

trabalhos semelhantes de construção de bases de dados de *tweets*, o processo de aquisição dos dados para construção da base, o método escolhido para rotulação dos *tweets* e segmentação por tópicos e os procedimentos utilizados para limpeza e pré-processamento dos dados. No Capítulo 3 são apresentados os modelos utilizados nesta dissertação, iniciando com a descrição de alguns métodos de representação numérica de textos, seguido de uma seção que descreve os modelos tracionais de aprendizagem de máquina que foram utilizados como baseline e finalizando com a descrição dos modelos baseados em *Transformers*. No capítulo 4 serão apresentados os detalhes dos métodos implementados neste trabalho, incluindo os métodos e métricas usados para avaliação dos modelos. O objetivo deste capítulo é uma delinear como as abordagens de aprendizagem profundam foram aplicadas e avaliadas diante do problema de detecção de posicionamento em *tweets*. O Capítulo 5 contém os resultados experimentais para as arquiteturas propostas e para outras abordagens testadas ao longo do processo de desenvolvimento desta pesquisa. Finalmente, o Capítulo 6, descreve as considerações finais e direcionamentos futuros.

2 ELABORAÇÃO DA BASE DE DADOS

Diante de uma quantidade incessante e volumosa de opiniões e posicionamentos, reproduzidos nas redes sociais em tempo real, discutindo cada estágio da pandemia e seus desdobramentos, elaborar uma base de dados representativa, robusta e bem distribuída, em termos de atributos disponíveis e variedade de tópicos discutidos, faz-se uma tarefa desafiadora. Desse modo, tornou-se pertinente um breve estudo de trabalhos relacionados à construção de uma base de dados, considerando principalmente os trabalhos nos quais foram desenvolvidos conjuntos de dados alinhados aos objetivos e ao escopo desta dissertação. Em seguida a esse estudo inicial, neste capítulo será relatado todo o processo de construção da base, desde a aquisição dos dados originais - sem nenhum tratamento - até o seu produto final, passando pelo processo de rotulação, segmentação por tópicos e limpeza. Com a base pronta para modelagem, mas também durante o seu processo de montagem, foram realizadas análises exploratórias dos dados, que serviram tanto como guia do processo de construção da base como referência para as estratégias de modelagem do problema de classificação adotadas em seguida.

2.1 TRABALHOS RELACIONADOS

Tendo em vista os objetivos desta dissertação, investiga-se nesta seção a disponibilidade e variedade de bases de dados relacionadas ao problema em questão, além de explorar um pouco do processo e das metodologias aplicadas no seu desenvolvimento, que de modo semelhante a base que pretende-se desenvolver: (1) utilizam o mesmo tipo de dados (*tweets*); (2) abordam o mesmo domínio (Covid-19); (3) possuem o mesmo idioma (Português); (4) são desenvolvidos para mesma tarefa (Detecção de Posicionamento).

Coletar dados do *Twitter* é, com certa regularidade, a principal escolha dos pesquisadores da área de mineração de opiniões, devido à acessibilidade e facilidade do processo, que garante aos usuários de sua API um grande volume de dados - apesar de certas restrições -, rastreando e configurando as requisições através de palavras-chaves ou IDs dos *tweets*/usuários de seu interesse. Não é à toa que muitas das bases de dados utilizadas na área de mineração de opiniões - considerando aqui também sistemas de análise de sentimento - são de dados extraídos do *Twitter*.

Na área de Análise de Sentimentos, entre as bases mais utilizadas para esse problema de

classificação, existem conjuntos de dados extraídos de postagens no *Twitter*. A maior e mais popular delas é a Sentiment140 (GO; BHAYANI; HUANG, 2009), usada para classificar o sentimento, em "negativo" ou "positivo", relacionado a uma marca, produto ou tópico. Com 1.6 milhão de *tweets*, os autores da base usaram algoritmos de aprendizado de máquina para anotar o sentimento dos *tweets* usando *distant supervision* ao invés de trabalhar com abordagens mais comuns durante processo de rotulação automática em texto, como as abordagens baseada em dicionário (léxicos) ou palavras-chave. Assim, a base Sentiment140 usa os resultados da classificação como rótulos e é amplamente usada pela comunidade de PLN, principalmente para treinar modelos usados em plataformas de gerenciamento de marca.

Do mesmo modo, na área de detecção de posicionamento, podemos citar como principal exemplo o SemEval *dataset* (MOHAMMAD et al., 2016), que possui conjuntos de *tweets* para duas tarefas de classificação, uma supervisionada, com 4870 *tweets* rotulados manualmente sobre cinco tópicos alvo distintos e outra de supervisão fraca, com 707 *tweets* rotulados e 78 mil não rotulados sobre o ex-presidente americano Donald Trump. Essa base proporcionou o primeiro conjunto de dados de benchmark para detecção de posicionamento nas redes sociais, e é referência para vários artigos que abordam o problema (ALDAYEL; MAGDY, 2019) e (MOHAMMAD et al., 2016).

No que se refere às bases de *tweets* relativos à pandemia de Covid-19, recentemente foram publicados muitos trabalhos que se propuseram a desenvolver conjuntos de dados sobre o tema, alguns destes na área de detecção de posicionamento, convergindo com as metas desta dissertação em relação ao domínio de aplicação, ao tipo de dados e aos sistemas de classificação desenvolvidos, oferecendo, desse modo, importantes contribuições para este estudo.

Glandt et al. (2021) fizeram a coleta e anotação de posicionamentos expressos pelos usuários do *Twitter* com respeito aos temas que giram em torno da pandemia, criando um novo conjunto de dados de detecção de posicionamento, chamado COVID-19-Stance. A base COVID-19-Stance consiste em 6.133 *tweets* em inglês de posicionamentos em relação a quatro tópicos alvo que abordam as medidas sanitárias da Covid-19 adotadas no contexto americano. Os tópicos alvo da base são: "*Anthony S. Fauci, M.D.*", "*Keeping Schools Closed*", "*Stay at Home Orders*", e "*Wearing a Face Mask*". Os *tweets* foram manualmente anotados de acordo com três categorias de posicionamento: "a favor", "contra" e "nenhum".

Outro estudo que se propôs a montar uma base para detecção de posicionamentos relacionados à Covid-19 no contexto americano foi Miao, Last e Litvak (2020). Os autores desenvolveram um conjunto de dados focado no posicionamento do usuário em relação ao

lockdown na cidade de Nova York. Usando palavras-chave relacionadas a "*lockdown*" e "*New York City*", extraíram 38.169 *tweets* relevantes de uma grande base de *tweets* relacionados à Covid-19, publicado por Chen et al. (2020). Os autores investiram menos no processo de anotação manual dos rótulos, anotando apenas uma pequena base de posicionamento de 1629 *tweets* e inspirados na ideia de destilação de dados (LIU et al., 2019a; XIE et al., 2020) da área de Visão Computacional, aplicaram uma estratégia de data augmentation usando, além do pequeno conjunto anotado pelos autores, dois conjuntos de dados rotulados e públicos de tarefas relacionadas - as bases já citadas acima, SemEval e Sentiment140 - e o modelo de linguagem BERT (DEVLIN et al., 2018).

Além desses estudos, é possível citar mais dois trabalhos que se dedicaram no desenvolvimento de bases de *tweets* com posicionamentos no contexto da pandemia, mas com alvos e propósitos mais específicos.

Mutlu et al. (2020) publicaram um conjunto de dados de 14.374 *tweets*, chamado COVID-CQ, que foram anotados manualmente em relação ao posicionamento do usuário em relação ao uso da hidroxicloroquina no tratamento de pacientes com Covid-19. O COVID-CQ contém posicionamentos no *Twitter* em relação a um tópico bem específico da pandemia, a alegação de que "cloroquina e hidroxicloroquina são a cura para o novo coronavírus". Apesar da especificidade do tópico, é o maior conjunto de dados de detecção de posicionamentos anotados por humanos para redes sociais até o momento da publicação desta dissertação. Além disso, ao concentrar-se em apenas uma única reivindicação, esse conjunto de dados pode ser utilizado para investigar a dinâmica das opiniões ao longo do tempo em relação ao uso desses medicamentos.

Embora com um propósito levemente diferente, por se tratar de um estudo de detecção de desinformação em relação à Covid-19 nas redes sociais, Hossain et al. (2020b) dividem essa tarefa de classificação em duas subtarefas: (1) A coleta de *misconceptions* relevantes em postagens que são verificadas quanto à sua veracidade e (2) A detecção de posicionamento para identificar se as postagens concordam, discordam ou expressam nenhuma posição em relação às *misconceptions* coletadas. Deste modo, desenvolvem um conjunto de dados, CovidLies, de 6.761 *tweets* anotados por especialistas para avaliar o desempenho de sistemas de detecção de desinformação em 86 tipos diferentes de desinformação relacionada à Covid-19.

Em relação à tarefa de detecção de posicionamento em *tweets* em português, Christhie et al. (2018) desenvolveram um estudo que pode ser considerado o primeiro a abordar esta tarefa aplicada no contexto brasileiro. Em circunstância das eleições presidenciais de 2018, os

autores fizeram uma coleta de *tweets* entre Outubro de 2017 a Janeiro de 2018, através de *hashtags* selecionadas manualmente e com forte viés ideológico que indicavam ser favoráveis ou contrárias aos candidatos que estavam sendo cogitados para concorrer ao cargo de presidente da República: Alckmin, Dória, Bolsonaro e Lula. Como resultado, obtiveram uma base de dados com *tweets* favoráveis ou contrários para cada alvo, neste caso, os possíveis presidenciáveis. É importante evidenciar que os autores removeram as *hashtags* que serviram como *pseudo-labels* dos *tweets* para não enviesar o processo de modelagem e também removeram *tweets* duplicados. No geral, o conjunto de dados final possui aproximadamente 60 mil *tweets*, embora as classes e os alvos estejam muito desbalanceados, sendo o maior alvo Lula, seguido de Bolsonaro, este último possuindo muitos *tweets* favoráveis e poucos contrários. Por tal motivo, na composição dos dados de treinamento, os autores optaram por estratégias como considerar os *tweets* favoráveis a Lula como contrários a Bolsonaro, o que é coerente tendo em vista que são candidatos antagônicos no cenário político brasileiro.

Ainda no contexto brasileiro, mas agora tratando-se de *tweets* semelhantes ao domínio desta pesquisa, o estudo de Brum et al. (2020) realizou um trabalho pertinente e volumoso na caracterização de *tweets* em português relacionados à Covid-19, ao analisar 56 milhões de publicações feitas entre 23 de Abril e 02 de Julho de 2020. O estudo concentrou-se no volume das publicações, seu conteúdo textual, localização e os principais elementos textuais dos *tweets* como URLs e *hashtags*, além de caracterizar os perfis dos usuários.

Uma contribuição crucial do trabalho de Brum et al., que afeta diretamente a configuração do problema de pesquisa aqui apurado, é na identificação do aspecto predominantemente político das discussões, detectados sobretudo através das *hashtags* mais comuns que se concentravam no apoio ou crítica ao presidente Jair Bolsonaro (#fechadocombolsonaro e #forabolsonaro, respectivamente).

Como atestado pelos autores, o consumo de informações sobre o coronavírus nas redes sociais são acompanhados de opiniões em tempo real em relação ao tema, com os usuários repercutindo a peculiaridade da situação, a necessidade e as dificuldades do distanciamento social e sobretudo as medidas tomadas pelos seus governantes para combater o avanço da pandemia. O artigo conclui que o debate no *Twitter* sobre Covid-19 no Brasil ficou mais centrado nas políticas de saúde pública adotadas pelos governantes, indicando uma pretensão futura de caracterizar os posicionamentos e argumentos políticos dos usuários em relação às medidas tomadas pelo Governo Federal e por seu Ministério da Saúde, algo que está diretamente relacionado aos objetivos desta dissertação.

Outro estudo que demonstrou como o debate em torno da Covid-19 nas redes sociais foi altamente politizado, com preferências políticas influenciando nas crenças e descrenças sobre o vírus, mas com foco no contexto americano, foi publicado no artigo de Jiang, Ren e Ferrara (2021). Neste trabalho, os autores exploram o processo de politização e polarização das opiniões relacionadas à Covid-19 nos EUA através da ideia de câmeras de eco, estruturas formadas por bolhas de filtro das redes sociais, nas quais só circulam informações com as quais os usuários já estão de acordo, reforçando o viés de confirmação de sua percepção acerca de determinado tema.

O estudo de Jiang, Ren e Ferrara, apesar de se propor a classificar os usuários e não as postagens, além de investigar sistemas de classificação de polaridade/ideologia política e não de detecção de posicionamentos, teve uma influência considerável no desenvolvimento desta dissertação, tanto na etapa de construção da base como na de modelagem. Como neste capítulo o objetivo é relatar o processo de construção da base, as influências no processo de modelagem serão descritas com mais detalhes no próximo capítulo e aqui será relatado brevemente o processo de elaboração dessa base, que também será retomado nas próximas seções do capítulo.

Jiang, Ren e Ferrara reassumem o debate em torno das estratégias de representação dos dados do *Twitter*, muito frequente nos trabalhos de Aldayel e Magdy (2019, 2021) sobre detecção de posicionamento e já discutidas no capítulo anterior. Tendo em vista as duas abordagens possíveis, baseada no conteúdo textual - *hashtags*, texto da publicação, descrição do perfil, etc. - ou baseada nos atributos relacionais - seguidores, menções, curtidas e retuítes - ainda é possível levar em consideração o tipo de entidade a ser analisada: as postagens (*tweets*) ou seus usuários (perfis). A construção da base de Jiang, Ren e Ferrara levou em consideração os dois tipos de atributos e concentrou-se na análise dos usuários. Deste modo, os atributos textuais foram as descrições do perfil (bios) e os atributos relacionais foram os metadados acerca de contas retuitadas e mencionadas pelos usuários, sobre os quais tornou-se possível gerar uma grafo de interações entre os usuários da base.

Jiang, Ren e Ferrara, assim como o estudo já citado de Miao, Last e Litvak, optaram por usar um grande conjunto de dados do *Twitter* sobre Covid-19 coletado por Chen et al., contendo, até o momento da publicação de seu estudo, dados de 21 de janeiro a 31 de julho de 2020.

Na base de dados de Chen et al., todos os *tweets* foram coletados através de palavras-chave relevantes para Covid-19 como: “coronavírus”, “corona”, “COVID”, “covid19”, “COVID-19”,

"Wuhan", "China", entre outros. Os *tweets* podem ser um *tweet* original, *retweets*, *tweets* citados (*quoted tweets* - *retweets* com comentários) ou *tweets* de respostas (*replies*). Cada *tweet* também contém a descrição do perfil do usuário, o número de seguidores, a localização fornecida pelo usuário e muitas outras informações contidas nos metadados de cada publicação.

Como é possível deduzir a partir das palavras-chave escolhidas para coleta, em sua maioria utilizadas por usuários de vários idiomas para descrever e comentar os eventos relacionados a pandemia de Covid-19, além de ser uma característica que é ressaltada no próprio artigo de divulgação da base (CHEN et al., 2020), as buscas realizadas por este estudo resultam num conjunto de dados multilíngue, isto é, composta por *tweets* de vários idiomas, sendo os mais frequentes inglês, espanhol, português e francês.

Além das já citadas características que tornam essa base significativa e passível de consideração para uso, como a multimodalidade de atributos e o multilinguismo de suas postagens, dois outros fatores tornam-se relevantes para a escolha de sua utilização como ponto de partida para construção da base deste trabalho:

1. A base é continuamente atualizada, possuindo publicações desde Janeiro de 2020 até o momento da publicação deste trabalho, Fevereiro de 2022. Isso é crucial para coleta de *tweets* antigos, tendo em vista que a API do *Twitter* restringe a busca com palavras-chaves até 7 dias precedentes à requisição. Além disso, pela amplitude do período de busca e por seu caráter contínuo, ela possui atualmente um volume absurdo de 2.2 bilhões de *tweets*, dentre estes 93 milhões em português (4% da base).

2. O Outro fator que a distingue de outras bases dessa magnitude é o fato dela ser pública e os IDs dos *tweets* coletados serem divulgados quinzenalmente em um repositório no Github[1]. Através dos IDs é possível coletar qualquer *tweet* da base, de qualquer período da pandemia[2].

Apesar de suas qualidades, há algumas ressalvas para com o uso dessa base. As restrições do número de requisições da API do *Twitter* tornam o processo de coleta demorado diante do grande volume de dados, que pode demorar meses para ser finalizado. Além disso, a impossibilidade de discriminar o idioma através dos IDs dificulta o processo de aquisição dos dados de interesse - neste caso, os *tweets* que estão em português - com maior eficiência. Outro fator

[1] https://github.com/echen102/COVID-19-TweetIDs

[2] De acordo com os Termos de Serviço do *Twitter*, fornecer conteúdo a terceiros, incluindo conjuntos de dados para download de conteúdo ou uma API que retorne conteúdo, só é permitido através do download de IDs de *tweets* e/ou IDs de usuários.

limitante é relativo aos *tweets* que foram excluídos pelos usuários ou pela própria plataforma durante os dois anos de coleta, acumulando na base IDs de muitos *tweets* que não existem mais. O processo de aquisição dos dados, no entanto, será mais detalhado na próxima seção do capítulo.

2.2 AQUISIÇÃO DA BASE

Como já mencionado, devido às restrições de uso da API, a coleta de bilhões de *tweets* através de seus IDs[3] pode levar muito tempo. Por este motivo, antes de escolher a base de Chen et al. como ponto de partida, um teste foi efetuado em Setembro de 2021 com uma amostra aleatória de apenas 1% dos dados, investigando o tempo total para coleta e explorando a qualidade e variedade dos dados disponíveis, depois de filtrá-los por idioma. Atualmente o *Twitter* limita a quantidade de requisições que um usuário de sua API pode realizar, tendo recentemente atualizado sua cota para 900 requisições em 15 minutos ou 360.000 *tweets* por hora[4]. Na época, a base completa possuía na ordem de 1.8 bilhão de *tweets* e a coleta de apenas 1% dos *tweets* durou 2 dias. Destes, apenas 609.578 eram em português[5].

A partir dessa amostra de dados, foi realizada uma análise exploratória das características principais dos *tweets* que pudessem indicar um posicionamento em relação às medidas do governo para com a pandemia, como *hashtags* assinaladas na publicação, na descrição do perfil ou no nome do usuário, tendo em vista que virou prática comum dos usuários brasileiros no *Twitter* se posicionar, através do uso *hashtags* na descrição e no nome, como, por exemplo, "Nome_do_usuário#VacinaJá" ou "Nome_do_usuário#UsePFF". Além disso, também foram consideradas as contas mais retuitadas pelos usuários presentes na base. De antemão, nesta etapa de exploração não foi realizado nenhum pré-processamento, apenas a remoção de duplicatas de usuários quando analisadas as *hashtags* contidas na sua descrição de perfil e nome.

Comparando as *hashtags* mais frequentes nas postagens, descrição e nome, nas Tabelas 1,2,3, respectivamente, percebe-se que as duas últimas possuem um maior número de referências políticas, quase uma unanimidade, e também algumas campanhas relacionadas

[3] A coleta gerada a partir de IDs de *tweets* também é referido ao termo re-hidratação de *tweets*. Tendo em vista que o compartilhamento de *tweets* não é permitido pelo Termos de Serviço do *Twitter*, os repositórios com o de Chen et al. são compostos apenas de identificadores de *tweets* que são possíveis de re-hidratar.

[4] Cada requisição coleta em torno de 100 *tweets*.

[5] Lembrando que alguns *tweets* que possuem seus IDs na base foram deletados na plataforma e o número estimado da coleta sempre é maior do que a quantidade que é de fato recuperada.

Tabela 1 – As 10 *hashtags* mais frequentes nas publicações coletadas

Hashtags	**Frequência**
#COVID19	5.204
#Covid19	3.209
#coronavirus	3.069
#covid19	2.321
#G1	2.144
#ForaBolsonaro	775
#Coronavirus	726
#CPIdaCovid	719
#pandemia	668
#covid	607

Fonte: Autor

Tabela 2 – As 10 *hashtags* mais frequentes nas descrições dos perfis dos usuários

Hashtags	**Frequência**
#ForaBolsonaro	714
#forabolsonaro	310
#Bolsonaro2022	287
#FechadoComBolsonaro	227
#EuAutorizoPresidente	148
#EsquerdistasSeguemEsquerdistas	147
#BTS	146
#방탄소년단	139
#Lula2022	131
#FORABOLSONARO	129

Fonte: Autor

a pandemia, como é o caso da *hashtag* "#VacinaJá", que, como já citado, virou tendência nos nomes dos perfis. Também aparecem com frequência *hashtags* críticas ou de apoio ao presidente como "#ForaBolsonaro" ou "#FechadoComBolsonaro", algo que já havia sido assinalado no estudo de Brum et al..

Em relação aos usuários mais retuitados descritos na Tabela 4, percebe-se a predominância de políticos e comentaristas políticos ligados ao presidente como Osmar Terra, Guilherme Fiuza, Carla Zambelli e Rodrigo Constantino, além do próprio do presidente, Jair Bolsonaro. Além destes perfis, observa-se a presença de plataformas de notícias como G1 e Revista Oeste e, como o mais retuitado - e nada surpreendente - o perfil de Átila Iamarino, divulgador científico

Tabela 3 – As 10 *hashtags* mais frequentes nos nomes dos usuários

Hashtags	**Frequência**
#ForaBolsonaro	841
#FechadoComBolsonaro	293
#MarcoTemporalNão	281
#Lula2022	271
#forabolsonaro	253
#VacinaJá	237
#VotoAuditavelJa	186
#FORABOLSONARO	146
#Bolsonaro2022	138
#LulaLivre	136

Fonte: Autor

Tabela 4 – Os 10 perfis mais retuitados da amostra de 650 mil *tweets*

Screen Name (@)	**Frequência**
@oatila	6.222
@jairbolsonaro	4.654
@revistaoeste	4.057
@OsmarTerra	3.599
@GFiuza_Oficial	2.892
@g1	2.884
@SigaGazetaBR	2.553
@CarlaZambelli38	2.405
@AiltonBenedito	2.384
@Rconstantino	2.303

Fonte: Autor

que durante a pandemia ficou conhecido no Brasil por fazer campanhas de conscientização, análises e previsões relacionadas a pandemia de Covid-19.

Explorando mais a fundo e manualmente as *hashtags* contidas nas postagens e utilizando um método similar ao adotado por Christhie et al., foram selecionadas as *hashtags* que, dentre as 100 mais frequentes, implícita ou explicitamente, indicavam um posicionamento para com as medidas implementadas pelo Governo Federal no enfrentamento da pandemia e assinaladas, provisoriamente, em "favoráveis" e "contrárias" a essas medidas.

Como é possível observar na Tabela 5, existe um grande volume de *hashtags* em publicações nas quais o posicionamento político para com as medidas tomadas pelo governo em relação

Tabela 5 – Hashtags com posicionamento em *tweets* mais frequentes

# Contrárias	**Frequência**	**# Favoráveis**	**Frequência**
#ForaBolsonaro	775	#BolsonaroTemRazao	235
#BolsonaroGenocida	357	#GloboLixo	186
#ForaBolsonaroGenocida	297	#OperaçãoCOVID19	141
#FiqueEmCasa	281	#CPIdoCirco	89
#VacinaParaTodos	159	#BolsonaroAte2026	88
#ImpeachmentBolsonaroUrgente	142	#STFVergonhaNacional	88
#fiqueemcasa	121	#RenanSuspeito	85
#FicaEmCasa	120	#ForaDoria	82
#WearAMask	109	#BolsonaroPresidenteAte2026	79
#VacinaSim	107	#TratamentoPRECOCESalvaVida	79

Fonte: Autor

à Covid-19 estão manifestas, como evidenciaram Brum et al., em sua base de dados e Jiang, Ren e Ferrara na mesma base, de Chen et al., mas em outro idioma.

Deste modo, visando ampliar o conjunto de dados foi implementado o mesmo procedimento de coleta anterior, selecionando uma amostra aleatória de IDs de *tweets*, mas dessa vez de 10% do total da base, durante o período de Janeiro de 2020 a Outubro de 2021, o que resultou em aproximadamente 200 milhões de *tweets* e *retweets* [6]. No entanto, após a filtragem dos *tweets* em português este número foi reduzido para cerca de 6 milhões, uma quantidade considerada suficiente para dar prosseguimento a construção da base de dados deste trabalho.

2.3 EXPLORAÇÃO DOS DADOS E ROTULAÇÃO DA BASE

A base de dados coletada contém, em termos mais exatos, $6.074.289$ de *tweets*, sendo $4.001.936$ (65,88%) destes, *retweets* (re-postagens de *tweets* de outros usuários). A proporção de *retweets* é semelhante à base de Brum et al. - 64,30% - mesmo que a quantidade total de sua base de *tweets* seja 10 vezes maior e que o intervalo seja bem menor, de 70 dias. É possível, assim, supor que os *tweets* em português relacionados à Covid-19 seguem um mesmo padrão de proporcionalidade entre *tweets* e *retweets*. Sendo assim, de modo semelhante ao artigo de Brum et al., aqui será brevemente analisado o volume de *tweets* e *retweets* ao longo do tempo.

[6] Destacando, novamente, que devido às restrições da API e a impossibilidade de filtragem do idioma antes da coleta, 10% foi a porcentagem limite, levando em consideração que o processo de re-hidratação dos *tweets* demorou mais de 1 mês para ser finalizado.

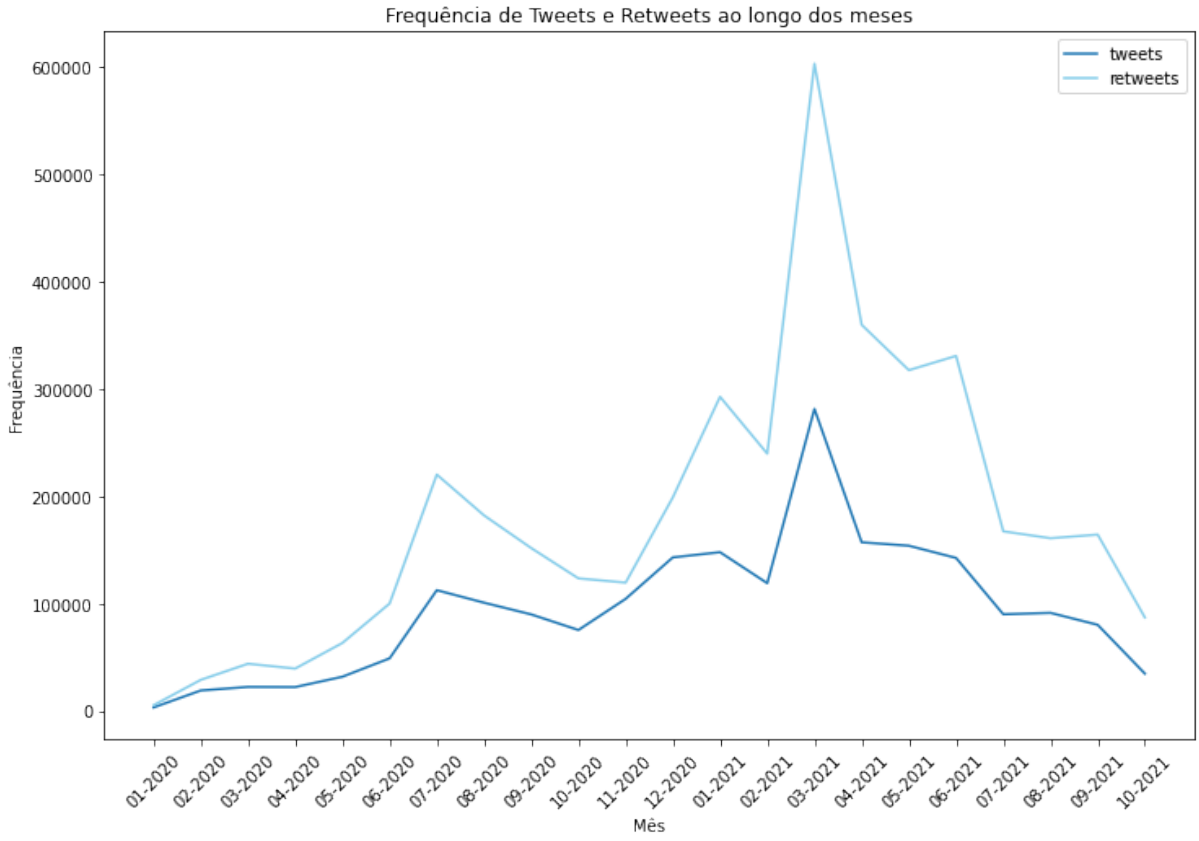
Tabela 6 – Os 10 meses com maior quantidade de *tweets* e *retweets*

Mês	**Total**	***Tweets***	***Retweets***
Março-2021	884.891	281.483	603.408
Abril-2021	517.126	157.173	359.953
Junho-2021	473.535	142.589	330.946
Maio-2021	471.622	153.975	317.647
Janeiro-2021	440.800	147.945	292.855
Fevereiro-2021	358.852	1190.26	239.826
Dezembro-2020	342.170	143.102	199.068
Julho-2020	332.861	112.536	220.325
Agosto-2020	283.150	100.872	182.278
Julho-2021	257450	90053	167397

Fonte: Autor

Levando em consideração que os 6 milhões de *tweets* publicados durante o período de Janeiro de 2020 a Outubro de 2021, os dados foram divididos em 22 partes, relativas a cada mês.

Figura 2 – Quantidade de *tweets* e *retweets* por mês



Fonte: Autor

Tabela 7 – Bi-grams mais frequentes da base

Bi-grams	**Frequência**
cpi covid	54.056
corona virus	52.645
covid nao	49.203
contra covid	49.131
vacina contra	40.202
mortes covid	26.837
vacinacao contra	23.834
tratamento precoce	23.445
vacina covid	22.651

Fonte: Autor

Através da Figura 2 e da Tabela 6, observa-se um aumento considerável na quantidade de *tweets* e *retweets* entre Janeiro e Junho de 2021, sendo o maior pico no mês de Março do mesmo ano. Durante esse período houveram vários eventos relacionados à Covid-19 de grande repercussão nas redes sociais: o aumento de casos e números recordes de mortes por covid, o início do processo de vacinação da população, novos mandatos de *lockdown* e fechamento parcial do comércio diante dos novos casos e o início das investigações da CPI da Covid. Alguns destes eventos, inclusive, serão considerados como tópicos específicos na próxima seção do capítulo, devido ao grande volume de *tweets* que geraram e também por dividir opiniões a seu respeito de acordo com o alinhamento em relação às medidas tomadas pelo governo, passíveis de extração de posicionamentos políticos favoráveis e contrários a estes tópicos alvo.

Após a remoção de *stopwords*, *hashtags*, URLs e alguns outros métodos de pré-processamento, foram calculados os uni-gramas e bi-gramas (dois termos em sequência) mais frequentes das publicações. Pode-se observar alguns dos temas de grande repercussão citados na Tabela 7, como a CPI da Covid, a vacinação da população, o número de mortos e o uso do tratamento precoce, este último aparentemente mais recorrente durante todo o período analisado.

A modelagem e segmentação por tópicos, no entanto, será explorada mais profundamente na próxima seção do capítulo.

Em relação as *hashtags* em *tweets*, nomes e descrições de perfil, além dos usuários mais retuitados observados na primeira amostra (Tabelas 8, 9, e 10), o mesmo padrão se repete aqui, variando apenas na frequência em que foram citadas. Na seção anterior, foram rotuladas as *hashtags* das publicações, também reproduzido na base aumentada, porém mais extensiva-

Tabela 8 – Hashtags Contrárias e Favoráveis ao Governo mais frequentes nas descrições

Contrárias ao Governo	**Frequência**	**Favoráveis ao Governo**	**Frequência**
#ForaBolsonaro	8.724	#Bolsonaro2022	2.979
#forabolsonaro	3.485	#FechadoComBolsonaro	2.213
#Lula2022	1.739	#BolsonaroAte2026	1.320
#FORABOLSONARO	1.648	#EuAutorizoPresidente	1.233
#EsquerdistasSeguemEsquerdistas	1.293	#VotoAuditavelJa	1.104
#LulaLivre	1.244	#QueroBolsonaroAte2026	978
#LulaPresidente2022	1.161	#DireitaSegueDireita	754
#BlackLivesMatter	1.042	#DireitaUnida	751
#EleNão	1.014	#BolsonaroPresidenteAte2026	542
#ForaBolsonaroGenocida	691	#Dia07VaiSerGIGANTE	517

Fonte: Autor

mente. 98 *hashtags* alinhadas à oposição e 73 alinhadas ao governo e suas medidas relativas à Covid-19 foram selecionadas entre as mais frequentes da base. As dez mais frequentes são semelhantes às da Tabela 5, mas com maior volume.

Este método para geração de *pseudo-labels* foi baseado no artigo de Christhie et al., como explicado na seção anterior. No entanto, diante de uma base mais diversa, não só de dados textuais da publicação, mas com metadados relativos aos usuários - textuais e relacionais - e suas interações no *Twitter* e baseando-se no métodos de supervisão fraca adotados por Jiang, Ren e Ferrara, também foram anotadas como "favoráveis" ou "contrárias" às medidas adotadas pelo governo e ao governo em si, as *hashtags* presentes nos nomes e nas descrições de perfil dos usuários da base.

Foram anotadas 123 *hashtags* contrárias e 101 favoráveis nas descrições de perfil dos usuários. Apesar da maior parte das mais frequentes fazerem apenas referência a política no âmbito institucional, muitas das *hashtags* anotadas, principalmente as contrárias ao governo, são de posicionamentos explícitos contra as medidas ou ausência de medidas do Governo Federal e seu Ministério da Saúde, como por exemplo, "#BolsonaroGenocida", "#VacinaParaTodos", "#FiqueEmCasa", "#VacinaJá", "#VacinaSim", "#ForaGenocida" e "#UseMáscara".

Das *hashtags* retiradas dos nomes dos usuários da base, foram anotadas 180 contrárias ao governo e 97 favoráveis. Apesar das *hashtags* estritamente políticas serem predominantes, assim como no caso das descrições do perfil, geralmente relacionadas aos candidatos Lula ou Bolsonaro, aqui nota-se uma presença ainda maior de campanhas de combate ao coronavírus já nas *hashtags* mais frequentes, como "#VacinaJá" e "#UsePFF2" nos usuários

Tabela 9 – Hashtags Contrárias e Favoráveis ao Governo mais frequentes no nomes

Contrárias ao Governo	**Frequência**	**Favoráveis ao Governo**	**Frequência**
#ForaBolsonaro	2.952	#FechadoComBolsonaro	635
#forabolsonaro	782	#Bolsonaro2022	521
#Lula2022	761	#EuAutorizoPresidente	126
#MarcoTemporalNão	649	#QueroBolsonaroAte2026	109
#FORABOLSONARO	538	#EuConfioNoPresidente	63
#VacinaJá	465	#FechadocomBolsonaro	59
#MarcoTemporalNao	253	#BolsonaroReeleito	55
#ForaBozo	230	#ForaDoria	54
#UsePFF2	228	#BolsonaroAté2026	52
#ForaBolsonaroGenocida	226	#BolsonaroTemRazao	49

Fonte: Autor

contrários e "#ForaDoria" nos favoráveis[7]. Além disso, muitas das *hashtags* anotadas como contrárias às medidas do governo demonstram um posicionamento explícito: "#VacinaSim", "#PL490NÃO", "#ForaGenocida", "#VacinaJá", "#UsePFF2", "#FiqueEmCasa", "#VacinaJá", "#BolsonaroGenocida", "#TodosPelasVacinas" e "#VacinaParaTodos".

Um método alternativo às *hashtags* é fazer uso de perfis influentes mencionados nos *tweets* dos usuários por meio de menções ou retuítes. Os estudos de Ferrara et al. (2020) e Jiang, Ren e Ferrara, por exemplo, fazem uso dos meios de comunicação proeminentes no *Twitter* e associam um retuíte explícito de uma conta oficial do *Twitter* a um endosso da publicação em questão e às ideias associadas ao usuário/perfil retuitado. Nessa direção, foram rotulados 150 perfis influentes no *Twitter* brasileiro de acordo com seu posicionamento para com o governo federal e as medidas que este adotou durante a pandemia. Neste caso, consideramos não só meios de comunicação e plataformas de notícias, mas também perfis influentes - geralmente verificados[8] - de políticos, ministros, jornalistas, entre outros.

É importante ressaltar que, com exceção de Atila Iamarino, e na contramão do que é possível observar nas *hashtags* - onde o volume de posicionamentos contrários é maior - os perfis favoráveis ao governo e suas medidas são mais frequentes nos retuítes e menções do que os perfis contrários.

Além de explorar os dados e metadados como nas Tabelas 8, 9, e 10, outras duas informa-

[7] João Dória, governador de São Paulo, liderou a campanha e aplicação da vacinação em seu estado, o que repercutiu em todo país e o posicionou em antagonismo com o Presidente Jair Bolsonaro, que inicialmente subestimou essa medida no enfrentamento da pandemia, levando inclusive, a contínuas substituições no Ministério da Saúde de seu governo.

[8] Explicar o que é ser verificado no *Twitter*

Tabela 10 – Dez perfis contrários e favoráveis mais retuitados da base

Perfis Contrárias	Frequência	Perfis Favoráveis	Frequência
@oatila	63.535	@jairbolsonaro	46.081
@brasil247	16.175	@revistaoeste	41.250
@GuilhermeBoulos	15.691	@GFiuza_Oficial	38.339
@DCM_online	15.587	@OsmarTerra	34.766
@randolfeap	13.191	@SigaGazetaBR	27.830
@felipeneto	13.125	@BrazilFight	24.001
@MarceloFreixo	13.018	@CarlaZambelli38	23.398
@revistaforum	11.500	@Rconstantino	22.798
@jandira_feghali	8.861	@AiltonBenedito	21.675
@slpng_giants_pt	8.699	@BolsonaroSP	21.647

Fonte: Autor

Tabela 11 – n-gramas Contrários e Favoráveis presentes na descrição do perfil

N-grams contrários	Frequência	N-grams favoráveis	Frequência
forabolsonaro	4.350	conservador	11.212
feminista	3.515	direita	10.933
esquerdista	3.424	cristao	8.748
direitos	1.998	patriota	8.555
lula	1.972	crista	4.516
antifascista	1.878	conservadora	4.418
petista	1.764	patria	3.055
cientista	1.644	bolsonarista	2.913
science	1.483	cristao conservador	1.739

Fonte: Autor

ções foram coletadas acerca dos usuários da base: os n-gramas das descrições do perfil e emojis contidos nos nomes dos usuários que foram manualmente selecionadas como "favoráveis" ao governo ou "contrários" - geralmente favoráveis a candidatos de oposição ou politicamente alinhados à esquerda do governo, que também podem servir de indicadores sobre o possível posicionamento político do usuário e, consequentemente, de suas publicações (Tabelas 11 e 12.

Na revisão da literatura sobre construção de bases de dados com *tweets*, realizada na primeira seção do capítulo, foram descritos alguns métodos de supervisão fraca para anotação automática de postagens. Christhie et al.. selecionaram *hashtags* presentes em publicações com forte viés ideológico para rotular publicações com posicionamentos políticos. Enquanto

Tabela 12 – Emojis Contrários e Favoráveis presentes no nome do perfil

Emojis Contrários	Frequência	Emojis Favoráveis	Frequência
	11.277		1.545
	4.253		1.097
	3.304		356
	1.937		

Fonte: Autor

isso, Jiang, Ren e Ferrara usaram duas estratégias fracamente supervisionadas para encontrar os *pseudo-labels* de tendências políticas para um subconjunto de usuários. No primeiro método, os autores reuniram as *hashtags* mais usadas nos perfis de usuário e anotaram como inclinados para a esquerda ou para a direita dependendo de qual partido político ou candidato eles apoiam ou se opõem. E no segundo método, já descrito anteriormente, os autores fazem uso dos portais de notícias proeminentes no *Twitter* e associam um *retweet* a essas contas como um endosso ao seu alinhamento político.

Além disso, Jiang, Ren e Ferrara, baseando-se no estudo de Conover et al. (2021) sobre polarização política no *Twitter*, não consideraram as *hashtags* usadas em postagens, indicando que estas podem ser usadas para injetar conteúdo oposto no feed de outros usuários. Em vez disso, apoiaram-se nos estudos de Badawy, Ferrara e Lerman (2018) sobre manipulação de política digital e Addawood et al. (2019), que tratam da identificação de *trolls*, para atestar que as *hashtags* que aparecem nos dados textuais e relacionais dos perfis capturaram com mais precisão a verdadeira filiação política do usuário.

Levando em considerações estes diferentes e, em certos pontos, contraditórios, métodos de supervisão fraca para o processo de rotulação automática da base, em "contrários" e "favoráveis" às medidas do governo federal para com a pandemia do coronavírus, foram contemplados seis atributos, atribuindo diferentes pesos a cada um, descritos na Tabela 13.

Visando diminuir a proporção de falsos positivos/negativos e considerando que Jiang, Ren e Ferrara atestam que as *hashtags* que aparecem nos dados textuais e relacionais dos perfis capturaram com mais precisão a verdadeira filiação política do usuário, as *hashtags* em descrições de perfil e no nome tem um peso maior do que *hashtags* em publicações, já que estas últimas nem sempre são usadas por pessoas que estão de acordo com o posicionamento da hashtag. Também possui um peso maior as postagens que retuitaram ou mencionaram contas rotuladas com algum posicionamento, já que *retweets* - principalmente quando não tem comentários adicionais (*quoted*) - geralmente implicam em apoio do *tweet* original. Os

Tabela 13 – Pesos para os atributos anotados

Atributos Rotulados	**Peso**
Hashtags em publicações	1
Hashtags na descrição do perfil do usuário	2
Hashtags no nome do usuário	2
Contas retuitadas ou mencionadas	2
n-gramas da descrição do perfil do usuário	1
Emojis no nome do usuário	1

Fonte: Autor

outros dois atributos, n-gramas rotulados da descrição e emojis no nome possuem um peso menor porque são baseados apenas em observações do comportamento dos usuários na rede, sem nenhuma validação conhecida por algum estudo, mas que se apoia nas afirmações de Badawy, Ferrara e Lerman (2018) e Addawood et al. (2019) de que os atributos do usuário capturam com maior precisão o alinhamento a algum tópico ou ideologia.

Os rótulos de "favorável" ou "contrário" às medidas do governo federal para o enfrentamento da pandemia - leia-se as classes "a favor" e "contra", respectivamente - foram geradas de acordo com estes atributos e pesos, conforme o Algoritmo 1.

É importante ressaltar que o objetivo de assinalar os pesos para os atributos considerados no processo de rotulação automática de *tweets* foi única e exclusivamente de evitar uma quantidade significativa de falsos positivos e falsos negativos, como descrito acima, de acordo com os estudos já realizados sobre essas características. De acordo com o Algoritmo 1, se, por exemplo, um *tweet* contém em sua publicação apenas uma *hashtag* favorável ao governo, isso não será suficiente para considerá-lo como um *tweet* favorável ao governo e o mesmo serve para um t*tweet* que possui uma *hashtag* contrária às medidas do governo. Já a apresentação de uma *hashtag* que indique o posicionamento no nome ou descrição do perfil do usuário responsável pela publicação de um *tweet* é suficiente para considerá-lo detentor desse posicionamento. Essa atribuição, no entanto, carece de uma análise exploratória mais contundente para assinalar o grau exato da importância dessas características para rotulação, o que é uma limitação da metodologia aqui adotada e que poderá ser explorada em trabalhos futuros.

Através dessas regras foi possível rotular 352.958 postagens em "a favor" ou "contra" as medidas do governo contra a Covid-19. Dentre estes *tweets*, 210.803 são contra e 142.136 são a favor.

Algoritmo 1: Geração da Base

para *Todas as amostras* **faça**

1 Extrair as carateristicas:

(a) #tweet

(b) #desc

(c) #nome

(d) @user retuitado

(e) emoji nome

(f) n-gramas da descrição do perfil

2 **para** *Todas as característica extraidas no Passo 1 com os rótulos da respectiva caracateristica* **faça**

- **se** *característica* $\in$ *AFAVOR* **então**
  - AFAVOR:=AFAVOR + w
- **senão**
  - **se** *característica* $\in$ *CONTRA* **então**
    - CONTRA:=CONTRA + w
- TOTAL:=AFAVOR+CONTRA

3 **se** *TOTAL* $>$ *1* **então**

- AFAVOR:=AFAVOR/TOTAL
- CONTRA:=CONTRA/TOTAL

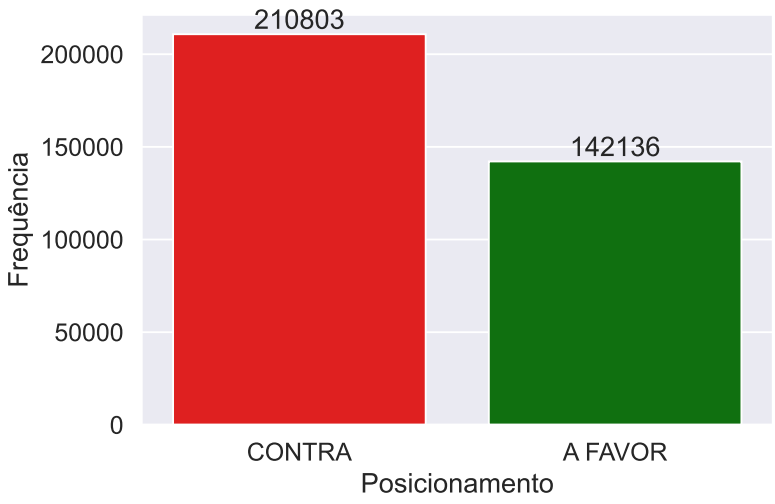
4 **se** *AFAVOR*$>=$ *0.7* **então**

- AMOSTRA É DA CLASSE AFAVOR

senão

- **se** *CONTRA* $>=$ *0.7* **então**
  - AMOSTRA É DA CLASSE CONTRA

Figura 3 – Gráfico com distribuição das classes



Fonte: Autor

Os 352.958 *tweets* rotulados oferecem uma quantidade de amostras considerável para realizar experimentos com sistemas de detecção de posicionamento. São mais amostras do que qualquer outro trabalho que foi descrito na primeira seção, em parte por terem sido anotadas automaticamente através de um sistema de regras e de supervisão fraca, em contraste com a metodologia de rotulação manual, adotadas para construir a base COVID-CQ (MUTLU et al., 2020), o conjunto de dados sobre *lockdown* em Nova York de Miao, Last e Litvak (2020), o conjunto de dados COVID-19-*Stance* de Glandt et al. (2021) e do SemEval *dataset* (MOHAMMAD et al., 2016).

No entanto, é importante ressaltar que o intervalo de tempo analisado é muito grande, de Janeiro de 2020 a Outubro de 2021, no qual uma série de eventos relacionados à pandemia decorreram. A título de exemplo e comparação, no conjunto analisado por Brum et al. os tópicos mais relevantes foram "quarentena", "covid", "aglomeração", "distanciamento" e "hidroxicloroquina". Estes tópicos refletem devidamente as discussões do início da pandemia, mas não contemplam, por exemplo, o processo de vacinação da população, o aumento no número de casos no início de 2021 e a CPI da Covid-19, que aparecem em grande volume no conjunto de dados gerados nesta dissertação, como foi possível observar através das *hashtags* e n-gramas usados para rotulação.

Diante do grande intervalo temporal e a partir do conjunto já rotulado de amostras, foram verificados os tokens, n-gramas (uni-gramas e bi-gramas) e *hashtags* nas publicações, em busca dos tópicos principais relacionados à Covid-19 através dos quais fosse possível fazer uma segmentação por tópicos, algo que será explorado e analisado na próxima seção do capítulo. Para alguns temas, inclusive, a segmentação do conjunto de dados rotulado em tópicos é também uma divisão temporal, tendo em vista que eventos como "vacinação", "cpi" e "*lockdown*" surgem como tópico de conversas e debates nas redes sociais apenas em momentos específicos da pandemia.

2.4 SEGMENTAÇÃO POR TÓPICOS

Após realizar um pré-processamento mais abrangente nos textos das postagens - removendo *stop-words*, *hashtags*, URLs, pontuação, acentos e arrobas - foram calculadas os n-gramas mais frequentes. Também a partir das postagens rotuladas, foram calculadas as *hashtags* mais frequentes, após extraí-las dos textos.

Através dos tokens e *hashtags* presentes nas postagens pode-se notar a presença de tópicos

Tabela 14 – Os vinte n-gramas e *hashtags* mais frequentes dos *tweets* rotulados

n-grama	**Frequência**	**Hashtags**	**Frequência**
covid	95.536	#COVID19	9.408
contra	36.251	#ForaBolsonaro	5.608
bolsonaro	28.869	#coronavirus	5.515
vacina	28.610	#Covid19	5.461
lockdown	25.111	#covid19	4.042
mortes	24.413	#BolsonaroGenocida	2.684
coronavirus	21.827	#ForaBolsonaroGenocida	2.293
cpi	20.078	#CPIdaCovid	2.224
pandemia	19.902	#FiqueEmCasa	2.122
governo	18.486	#BolsonaroTemRazao	1.396
pessoas	17.928	#VacinaParaTodos	1.396
casos	15.902	#FicaEmCasa	1.303
presidente	14.976	#ImpeachmentBolsonaroUrgente	1.196
pais	13.764	#covid19brasil	1.090
dia	12.936	#coronavírus	1.084
vacinas	12.599	#Covid_19	896
virus	12.175	#VacinaSim	840
tratamento	12.050	#COVID-19	783
cpi covid	11.909	#VacinaJa	777
milhoes	11.823	#ImpeachmentDeBolsonaroUrgente	734

Fonte: Autor

que fizeram parte do debate e dividiram opiniões entre os usuários das redes sociais como:

1. Vacinação: “vacina”, “vacinas”, “#VacinaParaTodos”, “#VacinaSim”, “#VacinaJa”
2. *Lockdown* e Isolamento Social: “*lockdown*”, “FiqueEmCasa”
3. CPI da covid: “cpi”, “cpi covid”, “cpidaCovid”
4. Tratamento precoce: “tratamento”

Explorando as 500 *hashtags* mais frequentes e os n-gramas (uni-gramas e bi-gramas) com frequência mínima de mil, foram rotuladas aquelas que se enquadraram nestes 4 tópicos e em mais 2, selecionados de acordo com a relevância no debate sobre a pandemia nas redes sociais no Brasil: a atuação de prefeitos e governadores, que muitas vezes fizeram um contraponto às medidas do Governo Federal e Ministério da Saúde, e o uso de máscara e sua obrigatoriedade em locais públicos.

Tabela 15 – Os 5 n-gramas mais frequentes em *tweets* relacionadas a cada tópico

vacinação	**lockdown**	**CPI da covid**	**tratamento**	**governadores**	**máscaras**
vacina	*lockdown*	cpi	tratamento	governadores	mascara
vacinas	isolamento	cpi covid	cloroquina	prefeitos	mascaras
vacinacao	fiquemcasa	cpidacovid	tratamento precoce	governadores prefeitos	use mascara
doses	quarentena	presidente cpi	ivermectina	doria	uso mascaras
vacina contra	*lockdown* nao	aziz	kit	estadual	usar mascara

Fonte: Autor

Tabela 16 – As dez *hashtags* mais frequentes em *tweets* relacionadas a cada tópico

Tópico	**Hashtags mais frequentes que definem o tópico**
Vacinação	"#VacinaParaTodos", "#VacinaSim", "#VacinaJa", "#vacina", "#VacinaParaTodosJa", "#VacinaJá", "#VacinasSalvamVidas", "#Vacina", "#vacinasim", #PátriaVacinada
***Lockdown* e Isolamento social**	"#FiqueEmCasa", "#FicaEmCasa", "#fiqueemcasa", "#*lockdown*", "#ficaemcasa", "#*Lockdown*Nao", "#stayhome", "#STAYHOME", "#StayHome", "#quarentena"
CPI da Covid	"#CPIdaCovid", "#CPIdoCirco", "#CPIdaPandemia", "#CPIdoGenocidio", "#RenanVagabundo", "#CPIdaPandemiaJa", "#RenanSuspeito", "#CPIdaVergonha", "#CPIDaCovid", "#RenanSabiaDeTudo"
Tratamento Precoce	"#TratamentoPRECOCESalvaVidas", "#TratamentoPrecoceSalvaVidas", "#cloroquina", "#Cloroquina", "#hidroxicloroquina", "#Ivermectina", "#TratamentoImediatoSalvaVidas", "#TratamentoPrecoce", "#ivermectina", "#CloroquinaSalvaVidas"
Governadores e Prefeitos	"#ForaDoria", "#DoriaGenocida", "#GovernadoresGenocidas", "#DoriaTemQueCair", "#DoriaMentiroso", "#ForaDoriaDitador", "#AgripinoDay", "#ReageSP", "#DoriaOMaiorVirusDoBrasil", "#CPIComGovernadoresEPrefeitos"
Uso de Máscaras	"#useMascara", "#UseMascara", "#usemascara", "#WEARAMASK", "#UseMáscara", "#WearAMask", "#mascara", "#mascarasalva", "#PFF2paraTodos" "#semáscarasempre"

Fonte: Autor

Tabela 17 – Frequência total de amostras e frequência dos rótulos para cada tópico

Tópico	Total	A favor	Contra
Vacinação	68.489	25.211	43.276
***Lockdown* e Isolamento Social**	40.854	19.464	21.390
CPI da Covid	28.723	12.392	16.331
Tratamento Precoce	30.490	14.724	15.765
Prefeitos e Governadores	13.969	10.794	3.175
Uso de Máscaras	11.643	3.093	8.550

Fonte: Autor

Baseando-se nas *hashtags* e n-gramas anotados manualmente que definem os seis tópicos determinados, as amostras da base de *tweets* rotulados foram enquadradas em cada tópico de acordo com uma regra simples: quando uma *hashtag* ou n-grama que define o tópico está contida na amostra, o *tweet* é considerado como pertencente àquele tópico. Deste modo 194.168 mil *tweets* foram enquadrados dentre os seis tópicos mais relevantes.

É importante recapitular que a base inicial foi rotulada de acordo com o posicionamento favorável (a favor) ou contrário (contra) as medidas tomadas pelo governo federal diante da pandemia. Portanto, na Tabela 17, as classes "a favor" ou "contra" não se referem diretamente ao tópico em si, mas às medidas do governo e seu alinhamento para com este tópico. Isto é, no tópico "Vacinação", as amostras contrárias não são contrárias à vacinação per se, mas contrárias às medidas (ou ausência de medidas) que o governo federal adotou em relação à vacinação.

Nota-se que, em consequência da classe majoritária do conjunto de dados rotulado ser "contra" as medidas do governo, este padrão se repete nos tópicos, com exceção do tópico "(atuação de) Prefeitos e Governadores", tendo em vista que foi um tema posto em pauta pelos apoiadores e usuários favoráveis ao governo federal. Já o desbalanceamento das classes, é mais significativo em três tópicos, onde a classe mais frequente é quase o dobro ou mais que o dobro da classe menos frequente: "vacinação", "Prefeitos e Governadores" e "Uso de Máscaras".

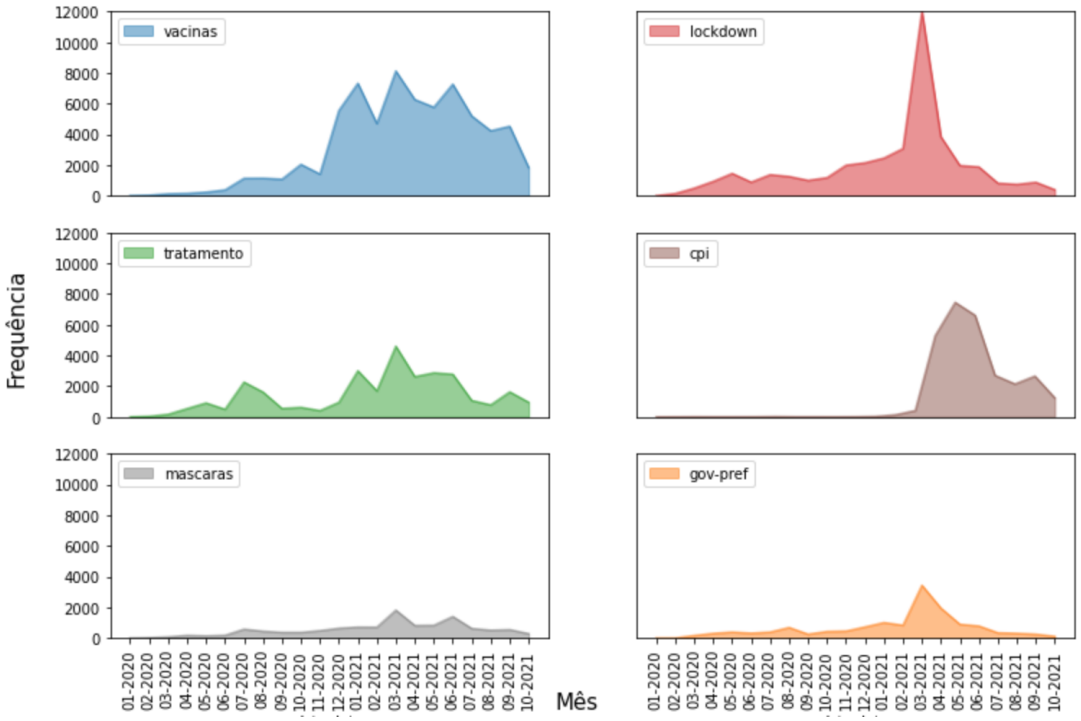
Numa etapa anterior à segmentação por tópicos através de n-gramas e *hashtags*, foi experimentada uma metodologia do estado da arte de segmentação automática de tópicos, o BERTopic[9]. BERTopic consiste em alguns componentes: os documentos são passados como entrada para um modelo *sentence transformer* que dará como saída um único vetor denso.

[9] https://maartengr.github.io/BERTopic/index.html

Esse vetor multidimensional (384 ou 768 dimensões) é uma representação numérica do significado por trás de cada documento. Antes de agrupar os documentos vetorizados, primeiro eles são comprimidos para um menor espaço dimensional (duas ou três dimensões) e para isso utiliza-se UMAP. Enfim, para clusterização utiliza-se o HDBSCAN e para identificar as palavras centrais, que são relevantes para cada tópico, e representar dos *clusters* utiliza-se c-TF-IDF.

Essa abordagem foi aplicada para detectar os tópicos da base de dados rotulada desta dissertação, mas não obteve êxito na representação de tópicos relevantes no contexto da pandemia de Covid-19 no Brasil, suscetíveis de um posicionamento por parte dos usuários do *Twitter*. Um dos problemas foi a quantidade elevada de *outliers*, variando entre 100 a 150 mil, dependendo dos parâmetros escolhidos (quantidade mínima de documentos por tópicos, número de tópicos, entre outros). Além disso, os tópicos gerados eram representados de maneira genérica como: "coronavirus-corona-virus-novo-coronavirus" ou "brasileiros-brasileiro-brasileira-povo-brasileiro". Diante de tais limitações usando *topic modeling*, optou-se pela abordagem descrita acima. Na abordagem adotada neste trabalho, os tópicos escolhidos foram baseados na frequência que esses temas aparecem na base de dados, além da regularidade e importância que eles tiveram no debate público durante a pandemia no Brasil. Em trabalhos futuros, no entanto, pretende-se aplicar outras metodologias de *topic modeling* como AOBTM e o próprio BERTopic, com outras configurações.

Figura 4 – Frequência de cada tópico ao longo da pandemia



Fonte: Autor

A partir da Figura 4, que apresenta a distribuição da frequência de cada um dos tópicos ao longo dos meses, é possível confirmar a suposição levantada na seção anterior, de que a segmentação do conjunto de dados por tópicos reflete, em alguns casos, uma divisão temporal das postagens. O tópico "vacinação", por exemplo, o mais frequente da base de dados, tem um considerável aumento a partir do fim de 2020, quando as vacinas contra a Covid-19 começam a surgir e, a partir deste momento, um alto volume de publicações acerca do processo de vacinação torna-se constante. Por outro lado, os tópicos "uso de máscara" e "tratamento precoce" se mantiveram relativamente estáveis já que são discussões que atravessam todo período analisado. O segundo, no entanto, obteve momentos de maior relevância e volume de acordo com alguns eventos, como a recomendação do uso do remédio pelo presidente da república e pelo Ministério da Saúde, além da prescrição por parte dos profissionais da saúde, gerando controvérsias[10]. O debate em torno da "Cpi da Covid-19", que investigou uma série de irregularidades na gestão e combate a pandemia, está bastante concentrado entre Abril e Junho de 2021, quando se tornou um destaque nos debates das redes sociais e teve seu ponto alto no depoimento do deputado Luís Miranda, quando indicou um suposto esquema de corrupção

[10] https://oglobo.globo.com/politica/bolsonaro-defendeu-uso-de-cloroquina-em-23-discursos-oficiais-leia-as-frases-25025384

na compra de vacinas por integrantes do governo Bolsonaro[11]. O tópico "*lockdown*" obteve o maior pico no volume de dados da base, no mês de Março de 2021, quando quase todos os Estados e Cidades adotaram medidas de isolamento total diante do aumento do número de casos e mortes decorrentes do coronavírus. Considerando que essas medidas foram implantadas por alguns governadores e prefeitos seguindo as orientações da Organização Mundial de Saúde (OMS) e contrariando às políticas do Governo Federal, o tópico "Governadores e Prefeitos" também obteve um maior volume de publicações neste mês.

2.5 LIMPEZA E PRÉ-PROCESSAMENTO DOS DADOS

Objetivando a remoção de vieses do conjunto de dados rotulados, optou-se por algumas estratégias de limpeza e pré-processamento indicados na literatura, embora este último varie bastante de acordo com os métodos de representação dos *tweets* utilizados para cada modelo proposto.

2.5.1 Remoção de Possíveis *Bots*

Como apontado por Ferrara (2020), é de suma importância determinar se as conversações em redes sociais refletem de fato as conversas de pessoas genuínas ou se são distorcidas pela atuação de contas automatizadas, muitas vezes chamadas de *bots* ou *bots* sociais. No artigo de Christhie et al. também está presente essa preocupação, como pode-se observar no trecho a seguir manifesto na conclusão do estudo:

"Alguns pontos precisam ser observados quanto à validade dos resultados encontrados para os classificadores. O primeiro diz respeito às suposições de existência de bots nas postagens da rede social. Encontramos diversas evidências da existência de postagens automatizadas, embora este não seja um dos objetivos diretos deste trabalho. Porém, uma investigação mais detalhada sobre sua existência e extensão é necessária, uma vez que a frequente repetição de postagens pode influenciar no classificador."

Apesar do desenvolvimento de sistemas detecção de *bots* também não englobar os objetivos principais desta dissertação, para evitar um enviesamento dos sistemas de detecção de posicionamento aqui desenvolvidos por contas inautênticas, planejou-se remover os vieses de *bots* em potencial que se infiltram no conjunto de dados rotulados. Como proposto por Fer-

[11] https://www.poder360.com.br/congresso/em-6-meses-de-cpi-da-covid-titulares-cresceram-59-nas-redes/

rara, a técnica calcula a probabilidade dos usuários serem bots usando a metodologia de Davis et al. (2016), atualmente implementada e aprimorada na aplicação Botometer[12], que estima, baseando-se em características textuais e relacionais, uma pontuação de 0 (provavelmente humano) a 1 (provavelmente *bots*) para os usuários do *Twitter*.

De acordo com o estudo de Ferrara, que se propõe a examinar *bots* em *tweets* sobre a Covid-19 também coletados da base de Chen et al., ao invés de realizar uma classificação binária de contas em *bots* e humanos, evidencia-se de que é mais fundamentado estudar as extremidades da distribuição das pontuações, focando em contas que exibem traços explícitos de *bots* ou humanos, evitando os casos limítrofes, nos quais a classificação pode ser equivocada, e visando assim um número mínimo de falsos positivos. Essa estratégia é corroborada pelas orientações de uso dos desenvolvedores do Botometer, que não recomendam estabelecer um limiar arbitrário para cada conta, mas sim através da distribuição dos scores de todas as contas analisadas, para que a taxa de falsos positivos seja a menor possível[13].

Deste modo, o Botometer foi aplicado neste trabalho para detectar *bots* nos usuários do conjunto de dados - apenas aqueles que possuem *tweets* rotulados e enquadrados em algum tópico[14] - seguindo a mesma estratégia Ferrara de considerar apenas os casos das extremidades da distribuição das pontuações, e utilizando a métrica CAP (*Complete Automation Probability*) do Botometer - que é a probabilidade, de acordo com os modelos usados, de que uma conta com tal pontuação ou superior seja um *bots*[15].

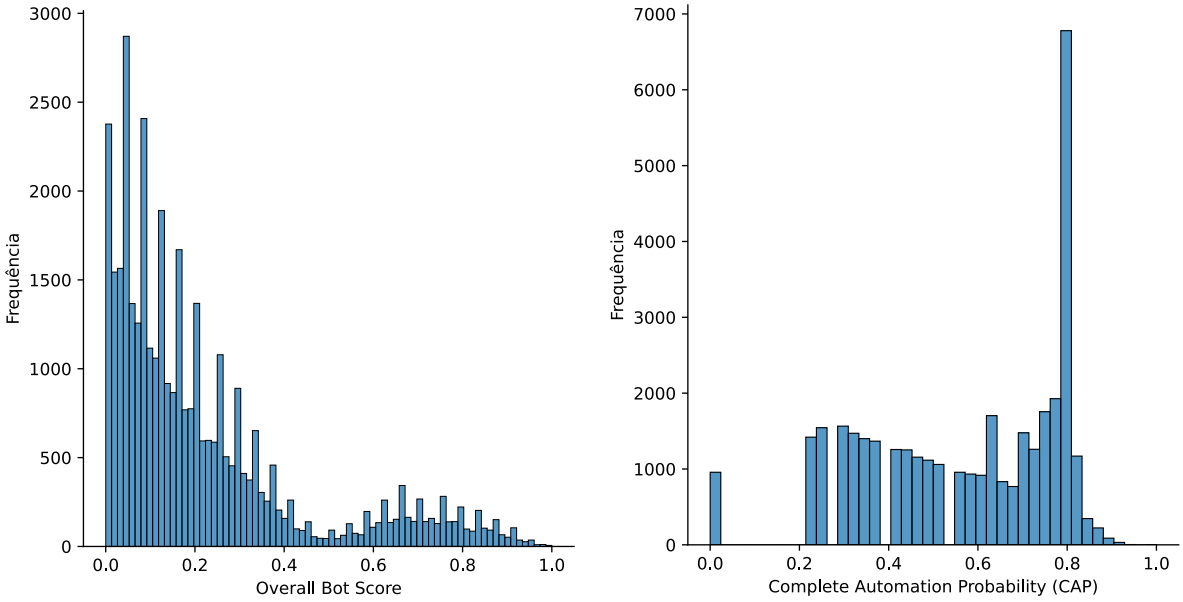
Dos 54 mil usuários presentes nas bases rotuladas e com tópico associado, apenas 63 (0.001%) dos usuários ficaram acima dos 90% de probabilidade de ser *bot* e foram excluídos da base para evitar enviesamento dos experimentos que serão realizados nos capítulos seguintes. Como podemos observar nas Figuras 5a e 5b, há um grande volume de usuários que ficaram acima de 80% de probabilidade de serem *bots*. Mas, como já atestado acima, por outros estudos e pelos próprios autores do detector de *bots*, foram considerados apenas os casos onde a probabilidade é mais extrema.

[12] https://github.com/IUNetSci/botometer-python
[13] https://botometer.osome.iu.edu/faq#bot-threshold
[14] Por conta da limitação de requisições da API do Botometer, a detecção de *bots* só foi realizada nas amostras principais, isto é, aquelas possuíam uma classe e tópico alvo associado.
[15] https://botometer.osome.iu.edu/faq#what-is-cap

Figura 5 – Pontuações dos *bots*

(a) Distribuição das Pontuações

(b) CAP dos Usuários



Fonte: Autor

2.5.2 Pré-processamento dos dados

Levando em consideração que os sistemas de detecção de posicionamento desenvolvidos neste trabalho terão como dados de entrada os textos das publicações e que as *hashtags* contidas nessas postagens foram uma das características utilizadas para a rotulação da base, essas *hashtags* foram removidas dos textos originais para que não enviesassem o treinamento dos classificadores, método de pré-processamento também aplicado por Christhie et al., que também utilizaram as *hashtags* das publicações para rotulação automática.

Outras técnicas de pré-processamento dos textos variaram de acordo com o método de representação das entradas escolhido. Para a adaptação de domínio do BERTimbau, por exemplo, que é realizada com dados não rotulados como continuação da etapa de pré-treinamento do modelo e que será detalhada mais profundamente no próximo capítulo, foram aplicados os mesmos procedimentos de limpeza usados pelos autores do BERTtweet (NGUYEN; VU; NGUYEN, 2020) e BERTweetFR (GUO et al., 2021b). Os textos dos *tweets* foram normalizados, convertendo as menções de usuários e links da web/url em tokens especiais como @USER e HTTPURL, respectivamente.

Outro exemplo de pré-processamento que foi aplicado no conjunto de dados, é referente aos atributos relacionais da base. Os dados relacionais são passíveis de extração através dos

metadados do *tweet*, que informam os usuários mencionados, retuitados e respondidos e serão utilizados para ajustar as representações dos textos das publicações através de um grafo de relações entre os usuários da base. Seguindo os procedimentos aplicados por Jiang, Ren e Ferrara, a rede de interações $G = (V, E)$ foi modelada como um grafo ponderado e não direcionado. Cada usuário $u \in V$ é um nó no grafo e cada aresta $(u, v) \in E$ indica que o usuário u mencionou, retuitou ou respondeu o usuário v ou vice-versa, já que trata-se de um grafo não direcionado. O peso de uma aresta $w(u, v)$ representa a quantidade de vezes que os usuários estabeleceram uma interação através de um retuíte, mention ou reply. Seguindo Garimella et al. (2018), foram mantidas apenas as arestas da rede com pesos de ao menos 2 e também removidos os usuários (nós) com graus inferiores a 5, uma vez que estes são na sua maioria usuários com menos atividade do *Twitter*. O conjunto de dados relacionais final contém 18.637 usuários com 245.760 interações entre eles. O grau médio dos nós da rede de interações foi de $26, 4$.

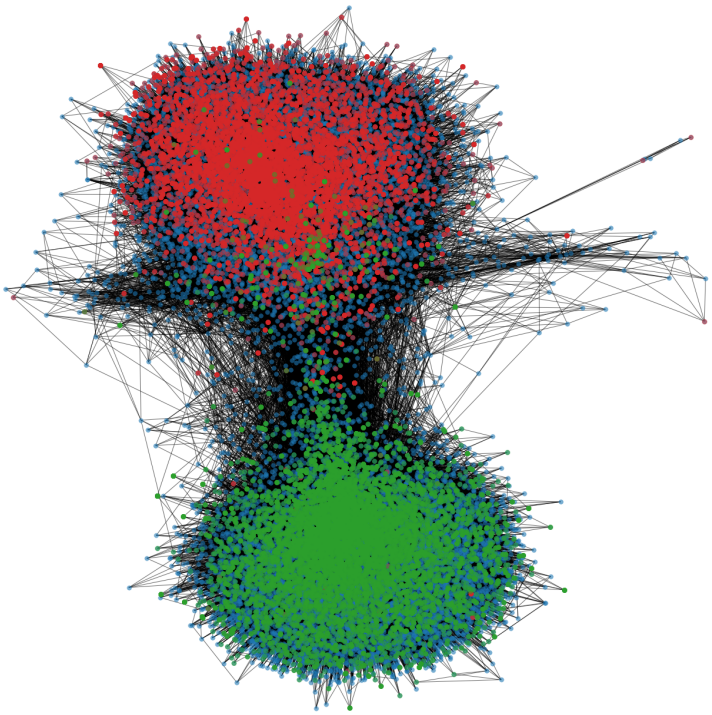
Para visualizar o grafo de interações, foi utilizada a biblioteca NetworkX [16] - uma biblioteca da linguagem de programação Python para estudar grafos e redes - para aplicar algoritmos de posicionamento de nós e desenhar o grafo. Mais especificamente, foi utilizado o algoritmo *force-directed de Fruchterman-Reingold* para posicionamento dos nós, cujo objetivo é posicionar os nós no espaço bidimensional de modo que todas as arestas tenham o mesmo comprimento e o menor número possível de arestas de cruzamento. Isso é possível ao atribuir forças anti gravitacionais entre o conjunto de arestas e nós com base em suas posições relativas e, em seguida, usa isso para simular o movimento das arestas e nós. Uma vez que as posições foram estabelecidas, também foram assinalados com diferentes cores os usuários, a partir de seus IDs, que possuem *tweets* rotulados em "contra", de vermelho, e "a favor", de verde. Os nós em azul são de usuários que não possuem nenhuma publicação rotulada na base de dados.

A representação visual do grafo de interações dos usuários da base de dados apresentada na Figura 6, apesar de não contemplar todos os seus usuários, é bastante elucidativa ao salientar a formação de *clusters* de relações entre usuários com posicionamentos semelhantes. O *cluster* com nós verdes, de usuários favoráveis às medidas do governo, é ainda mais homogêneo do que o *cluster* de nós vermelhos, de usuários contrários às medidas do governo frente à pandemia. Essa estrutura de relações é respaldada pela composição estabelecida entre esses grupos durante a pandemia de Covid-19, quando um bloco politicamente diverso, de pessoas, grupos e lideranças da esquerda à direita se posicionou contrariamente às medidas do governo,

[16] https://networkx.org/

enquanto que um bloco mais restrito e coeso de apoiadores do governo se manteve a favor de suas medidas.

Figura 6 – Visualização bi-dimensional do grafo de interações



Fonte: Autor

Essa ilustração do grafo de interações, com uma separação evidente entre usuários com diferentes posicionamentos e uma confluência de usuários com posicionamentos semelhantes corrobora com as suposições dos autores do Retweet-BERT e justifica sua experimentação nesta dissertação, considerando que o modelo tem como base a hipótese de que os usuários que retuitam uns aos outros são mais propensos a compartilhar ideologias semelhantes, tornando assim mais similares as representações dos textos dos usuários que se relacionam.

3 MODELOS

3.1 *LANGUAGE MODELING AND WORD REPRESENTATION*

Atualmente a área de Processamento de Linguagem Natural (PLN) é fortemente orientada por sistemas de aprendizagem de máquina, tendo em vista que a maioria dos sistemas de PLN seguem um *pipeline* no qual um texto é dado como entrada a um modelo de aprendizagem que produz uma predição qualificando essa entrada ou a descrevendo de alguma forma. Como é possível, portanto, representar e codificar textos numericamente, de modo a serem compatíveis com algoritmos de aprendizagem de máquina, com cálculo do gradiente descendente, entre outras computações?

O método mais simples de representar um texto numericamente é construir um dicionário de tokens com com todas palavras presentes no corpus textual de treinamento, ou ao menos as mais frequentes, e usar *one-hot embeddings*, vetores com a mesma dimensionalidade do vocabulário nas quais apenas um valor é igual a 1, que indica a palavra representada, e todos os outros valores são igual a 0. Embora essa abordagem possa ser útil para tarefas como *topic modeling*, representando os textos através das palavras que eles contêm, ela possui duas grandes desvantagens: (1) Os vetores são esparsos, possuindo uma dimensionalidade muito alta já que é compatível com a quantidade de palavras no vocabulário. (2) As representações são insuficientes para captar qualquer informação acerca da palavra, já que serão vetores de unidade alinhadas aos eixos do espaço vetorial. Além disso, desconsideram a ordem das palavras e a quantidade de vezes que um termo consta em um documento.

3.1.1 TF-IDF

Outro método simples, porém, mais eficaz na representação numérica dos textos é o *Term Frequency–Inverse Document Frequency* (TF-IDF) (RAJARAMAN; ULLMAN, 2011), uma técnica usada em aprendizagem de máquina e mineração de textos que introduz um fator de ponderação para as características textuais de um documento. A ideia por trás dessa medida estatística é que o peso de uma palavra seja calculada a partir de dois componentes: (1) *Term Frequency* - TF: função que aumenta o peso da palavra de modo diretamente proporcional a quantidade de ocorrências da mesma no documento. (2) *Inverse Document Frequency* - IDF: função que faz um contrabalanceamento do peso através do número de vezes que a palavra

aparece em todo corpus textual utilizado. este deslocamento provocado pela função IDF ajuda a remover a importância de palavras muito comuns (*stopwords*), que aparecem de forma frequente em todos os documentos, tendo em vista que estes termos muito comuns em todos os documentos - como “o”, “eu”, “seu”, “como”, “que”, “ele”, “foi” , “para”, “em”, “são”, “com”, etc. - geralmente não fornecem nenhuma informação valiosa sobre o documento que pretende-se classificar. A fórmula matemática do TF-IDF é a seguinte:

$$w_{i,j} = tf_{i,j} \times \log_2 \left(\frac{N}{df_i} \right) \tag{3.1}$$

Nota-se que nessa função o peso de uma palavra em relação a um documento é calculado apenas multiplicando duas outras funções: a função TF - a frequência do termo ($tf_{i,j}$) no documento e a função IDF, que é logaritmo do número total de documentos (N) dividido pela frequência dos documentos que contêm este termo (df_i). Na função IDF o logaritmo é usado para amortecer os resultados da divisão, dando maior peso para as palavras únicas e menos frequentes, que têm maior capacidade de diferenciar um documento de outro.

3.1.2 *Word Embeddings*

Um outro método que enriquece ainda mais a representação das palavras é considerá-las como vetores com valores reais contínuos e distribuídos em um espaço n-dimensional, também conhecido como *word embeddings*. Através deste método os tokens são codificados de modo a representar múltiplos aspectos de uma palavra, como gênero, flexões, parte da palavra, entre outras características e o mecanismo usado para expressar essas relações entre as palavras é a distância entre os vetores gerados. As *embeddings*, de um modo geral, são estruturas de poucas dimensões que formulam uma representação vetorizada de palavras, frases e outras modalidades de dados, como nós em um grafo. As *embeddings* capturam a similaridade semântica das entradas e podem ser pré-treinadas e transferidas entre conjuntos de dados ou tarefas. Para treinar *word embeddings* pode-se utilizar corpus textuais imensos e multilíngues disponíveis e públicos como Wikipédia, e geralmente é necessário formatar estes textos, não rotulados, como uma tarefa supervisionada, na qual o objetivo é prever qual a próxima palavra, dada uma frase incompleta, ou prever qual a palavra de acordo com as palavras em seu entorno. Uma vez treinado, cada palavra pode ser mapeada para uma *embedding* de vetor contínuo, na qual palavras ou frases semanticamente semelhantes compartilham *embeddings* semelhantes

entre si.

Word2Vec (MIKOLOV et al., 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) são considerados os primeiros métodos a introduzir representações distribuídas através de *language modeling*, a tarefa de prever uma parte do texto (palavra ou token) de acordo com o seu contexto. O contexto pode variar de acordo com o método escolhido para geração das *embeddings* e refere-se tanto às palavras precedentes ao token, quanto às palavras ao seu redor. Word2Vec e GloVe são métodos tradicionais de construção de *word embeddings*, que possibilitam o aprendizado de associações entre palavras de um grande corpus de texto sem nenhuma supervisão. Word2Vec considera uma palavra e seu entorno palavras como o contexto em uma frase, enquanto GloVe considera a matriz global de co-ocorrência das palavras. Uma vez treinados, ambos modelos produzem *embeddings* que capturam a semelhança semântica entre palavras.

Apesar destes métodos constituírem um grande avanço e inovação na área de PLN quando trata-se de representação vetorial das palavras, há algumas desvantagens e limitações a este paradigma. A principal limitação é que ao associar um vetor fixo a uma palavra não é possível lidar com casos nos quais essas palavras possuem múltiplos significados. A palavra "banco" nas frases "*O cofre do banco contém apenas dinheiro" e "no banco de trás de minha bicicleta*" são mapeadas ao mesmo vetor de *embedding*, apesar de possuírem significados diferentes. Por este motivo, *word embeddings* como Word2Vec e GloVe são consideradas como representações vetoriais contínuas, porém não contextuais. O ideal, portanto, é que as *embeddings* sejam contextualizadas, refletindo as palavras ao seu redor e que haja assim uma desambiguação em casos de múltiplos significados. A contextualização das *embeddings* tornou-se possível através de redes neurais recorrentes (RNNs) e aprimoradas através de *Transformers* (GOODFELLOW; BENGIO; COURVILLE, 2016), que serão objeto de estudo e investigação nas próximas seções do capítulo.

Nesta dissertação, como *baseline*, serão utilizadas *word embeddings* em português pré-treinadas através de métodos como Word2Vec e GloVe do repositório do NILC (HARTMANN et al., 2017) e também o TF-IDF para representar os conjuntos de *tweets*. Junto a essas representações, modelos supervisionados tradicionais de aprendizagem de máquina, como SVM (VAPNIK, 1999), *Random Forest* (HO, 1995) e MLP (RUMELHART; HINTON; WILLIAMS, 1986) serão experimentados para classificar os posicionamentos. Deste modo, na próxima seção será feita uma breve descrição destes modelos e nas seções subsequentes serão investigados com mais detalhes os modelos baseados em *Transformers*, aplicações de maior foco neste trabalho.

3.2 MODELOS TRADICIONAIS DE APRENDIZAGEM DE MÁQUINA

3.2.1 Classificadores usando MLP (*Multi-layer Perceptron classifier*)

Redes neurais artificiais são sistemas paralelos distribuídos compostos por unidades de processamento simples (neurônios) que computam certas funções matemáticas (funções de ativações), dispostas em uma ou mais camadas, interligadas por um grande número de conexões unidirecionais. Geralmente estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede. O funcionamento destas redes é inspirado numa estrutura física concebida pela natureza: o cérebro humano. Uma das áreas em que redes neurais têm encontrado ampla aplicação prática é em problemas de reconhecimento de padrões, na qual a forma como os problemas são modelados e o paralelismo natural inerente à arquitetura das redes neurais criam a possibilidade de um desempenho superior à dos modelos convencionais.

Em redes neurais, o procedimento usual na solução de problemas passa inicialmente por uma fase de aprendizagem, na qual um conjunto de exemplos representativos das classes de padrões (rotulados ou não) são apresentados para a rede que extrai automaticamente, através de um algoritmo de aprendizagem, as características necessárias para representar implicitamente os padrões. Essas características são utilizadas posteriormente para classificar outros padrões em função do grau de similaridade destes com as representações armazenadas na rede. O algoritmo de aprendizagem mais comum das redes neurais, o *backpropagation* (RUMELHART; HINTON; WILLIAMS, 1986), envolve uma regra de correção de erros baseados em gradientes.

Desde do final dos anos 2000 tem sido possível treinar redes neurais com muitas camadas escondidas, Redes Neurais Profundas (*deep learning neural networks*) (GOODFELLOW; BENGIO; COURVILLE, 2016) e resolver problemas cada vez mais complexos. As redes neurais profundas foram inspiradas pela sensibilidade local e orientação seletiva do cérebro sendo projetadas para que implicitamente extraiam características relevantes das entradas. As redes neurais profundas podem ser treinadas com métodos de aprendizagem como o algoritmo *backpropagation*, entretanto, o uso de certas funções de ativação pode fazer com que o aprendizado que ocorre nas camadas mais baixas da rede não sejam aprendidas pelas camadas mais altas. Algumas das possíveis soluções para resolver este problema são: (1) pré-treinamento camada-a-camada, (2) o desenvolvimento de uma memória longa e de curto prazo e (3) o uso da função de ativação ReLu (Unidade Linear Retificadora) que foi fundamental para o bom desempenho do

treinamento das redes neurais com muitas camadas escondidas em tarefas de reconhecimento de imagem, fala e linguagem natural (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Nas redes neurais profundas, cada neurônio se conecta com um conjunto limitado de neurônios da camada subsequente, restringindo as conexões entre neurônios a janelas limitadas (também conhecidas como filtros ou kernels). Os conjuntos de kernels formam as camadas de convolução nas redes neurais profundas que, por sua vez, são matrizes que definem uma determinada característica que se deseja detectar no padrão da rede. Além das camadas de convolução as redes neurais profundas tem uma camada de *Pooling* que pode ser vista como uma grade de unidades que sumarizam de alguma forma as ativações dos neurônios com que se conectam. Além disso, o emprego de GPUs (unidades de processamento gráfico) para o treinamento das redes neurais também possibilitou o desenvolvimento de redes mais complexas porque aumentou a velocidade de processamento dos computadores permitindo assim o treinamento de redes com um grande número de camadas escondidas.

3.2.2 Floresta aleatória (*Random Forest*)

Floresta aleatória (HO, 1995) é um algoritmo de aprendizado de máquina supervisionado que é construído a partir de algoritmos de árvore de decisão, amplamente utilizada em problemas de regressão e classificação. A ideia por trás do *Random Forest* é combinar - também conhecido como aprendizagem via *ensemble* - muitos classificadores (árvores de decisão) para fornecer soluções para problemas complexos.

As árvores de decisão são fáceis de construir, de usar e de interpretar, mas na prática possuem um aspecto que as impede de ser a ferramenta ideal para a aprendizagem preditiva, a imprecisão (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em outras palavras, árvores de decisão funcionam muito bem com os dados de treinamento usados para criá-las, mas não são flexíveis na hora de classificar novas amostras e tem pouca capacidade de generalização. Para superar essas limitações, as *Random Forest*s combinam a simplicidade das árvores de decisão com flexibilidade e aleatoriedade, resultando em um algoritmo mais robusto que apresenta melhorias consideráveis na precisão, evitando o sobreajuste, que são recorrentes nas árvores de decisão.

O passo inicial para desenvolver essa abordagem é criar um conjunto de dados *bootstrap*, isto é, um conjunto de dados que tenha o mesmo tamanho dos dados originais, mas com amostras selecionadas aleatoriamente. Em seguida, utiliza-se uma árvore de decisão usando o conjunto de dados bootstrap, mas utilizando apenas um subconjunto aleatório das variáveis

em cada etapa.

Percebe-se, portanto, que o funcionamento do algoritmo de *Random Forest* difere-se das Árvores de Decisões, justamente por introduzir aleatoriedade na seleção de atributos ao invés de usar métodos que definem a separabilidade dos atributos como o índice gini ou ganho de informação, usados para avaliar qual atributo deve ser selecionado inicialmente.

Em seguida, esse processo de seleção aleatória se repete, usando uma amostra *bootstrap* e considerando apenas uma subconjunto das variáveis em cada etapa, resultando em uma grande variedade de árvores, já que para cada árvore individual haverá atributos distintos treinados em conjuntos de dados distintos. Além do componente de aleatoriedade, essa variedade de árvores de decisão, com estrutura de características distintas, é o que torna as florestas aleatórias mais eficazes do que as árvores de decisão individuais. Para classificar as amostras com *Random Forest*, são considerados os resultados de todas as árvores individuais de forma agregada, definindo a amostra em questão baseado na classe mais votada ou a média entre as árvores individuais da Floresta Aleatória, processo também chamado de "*bagging*".

Algumas das vantagens deste algoritmo são: (1) Maior robustez; (2) Menos propenso a sobre-ajuste; (3) modelo simples de usar que produz, em geral, bons resultados, mesmo sem a necessidade dos ajustes dos hiperparâmetros; (4) Características importantes nos dados são possíveis de serem constatadas. A principal desvantagem, no entanto, é que demanda um considerável poder de processamento no treinamento e no processo de classificação.

3.2.3 Máquinas de vetores de suporte (*Support Vector Machines* - SVM)

Máquinas de Vetores de Suporte (VAPNIK, 1999) são modelos de aprendizagem embasados pela teoria de aprendizado estatístico com boa capacidade de generalização. Em problemas de classificação com SVM busca-se estimar um objeto geométrico separador das classes através de uma margem suave, a menor distância entre as amostras das diferentes classes e o *threshold*, mas que permite algumas classificações equivocadas, que pode ser uma linha em um espaço bidimensional, um plano em um espaço tridimensional ou um hiperplano, em um espaço n-dimensional de atributos.

A principal ideia por trás das máquinas de vetores de suporte é que com o aumento da dimensionalidade dos dados é possível projetar um classificador de vetores de suporte que possam separar as amostras de cada classe. Para transformar os dados e possibilitar a separação matemática das amostras no espaço dimensional, o SVM utiliza-se de funções do *kernel* (*The*

Kernel Trick) para encontrar sistematicamente classificadores de vetores de suporte numa maior dimensionalidade. Funções de *kernel* como o *kernel* polinomial e o *Radial Kernel* (RBF) são capazes de calcular as relações entre cada par de observação em cada dimensão, e essas relações são usadas para encontrar um classificador de vetor de suporte.

Desse modo, o SVM funciona muito bem com margem de separação clara entre as classes - classes linearmente separáveis - e é eficaz nos casos em que o número de dimensões é maior que o número de amostras. Algumas das vantagens do SVM são: (1) É uma técnica robusta em problemas com muitas dimensões, onde outras técnicas de aprendizado frequentemente apresentam modelos super ou sub ajustados; (2) o modelo encontra uma única solução ótima, diferentemente das redes neurais em que existe a possibilidade do modelo encontrar várias soluções sub-ótimas (mínimos locais); (3) SVM lidam bem com *outliers* e, pelo fato de permitirem erros de classificação, também são capazes de lidar com classificações sobrepostas. Algumas das desvantagens do SVM são: (1) sensibilidade do modelo às escolhas dos parâmetros e (2) dificuldade de interpretação do modelo gerado.

Em um problema de classificação binária, como é o caso desta dissertação, o objetivo da SVM é separar as instâncias das duas classes através de uma função de *kernel* que será obtida a partir dos exemplos conhecidos na fase de treinamento.

3.3 TRANSFORMERS

A arquitetura Transformer, proposta inicialmente no artigo produzido por pesquisadores do Google Research e da Universidade de Toronto, "*Attention Is All You Need*" (VASWANI et al., 2017), está por trás de muitos dos progressos recentes na área de PLN e de Aprendizagem Profunda como um todo, estendendo-se também a área de Visão Computacional. A princípio e de maneira bastante resumida, pode-se afirmar que o *Transformer* é um tipo de modelo de aprendizagem de máquina, uma arquitetura especial de redes neurais profundas. Desde a publicação do artigo original, surgiram uma miríade de modelos tendo esta arquitetura como base, os mais conhecidos são o BERT (DEVLIN et al., 2018) e GPT-3 (BROWN et al., 2020), que vieram a dominar os *leaderboards* da área de PLN em diferentes tarefas, desde a tradução à classificação de textos. Em outros termos, *Transformers* são modelos capazes de traduzir, representar e gerar vários tipos de textos, como poemas, notícias, postagens em redes sociais e até código de programação, além de outras modalidades de dados como imagens, vídeos e som, sendo assim usados para múltiplas aplicações nas muitas áreas do conhecimento, desde

a análise de sequenciamento em Bioinformática até resolução de problemas matemáticos.

No artigo inicial de Vaswani et al., que trata mais especificamente de sistemas de tradução automática de textos, a arquitetura *Transformer* é apresentada através de dois componentes principais, representados na Figura 7: um segmento responsável por codificar a entrada a ser traduzida (*encoder*) e outro que processa a saída do primeiro e decodifica numa saída final (*decoder*), gerando o texto traduzido. Cada um destes componentes da arquitetura original são compostos por seis camadas, que vieram a ser chamadas de *transformers blocks* nos trabalhos subsequentes que tinham o *Transformer* como base. Essa composição *encoder-decoder* é característica dos sistemas *sequence-to-sequence (seq2seq)* ou *text-to-text* - recebe um texto como entrada e produz uma saída também em forma de texto - porque foi desenvolvida objetivando superar os problemas das abordagens anteriores de aprendizado profundo de sequências textuais, como RNN e LSTM.

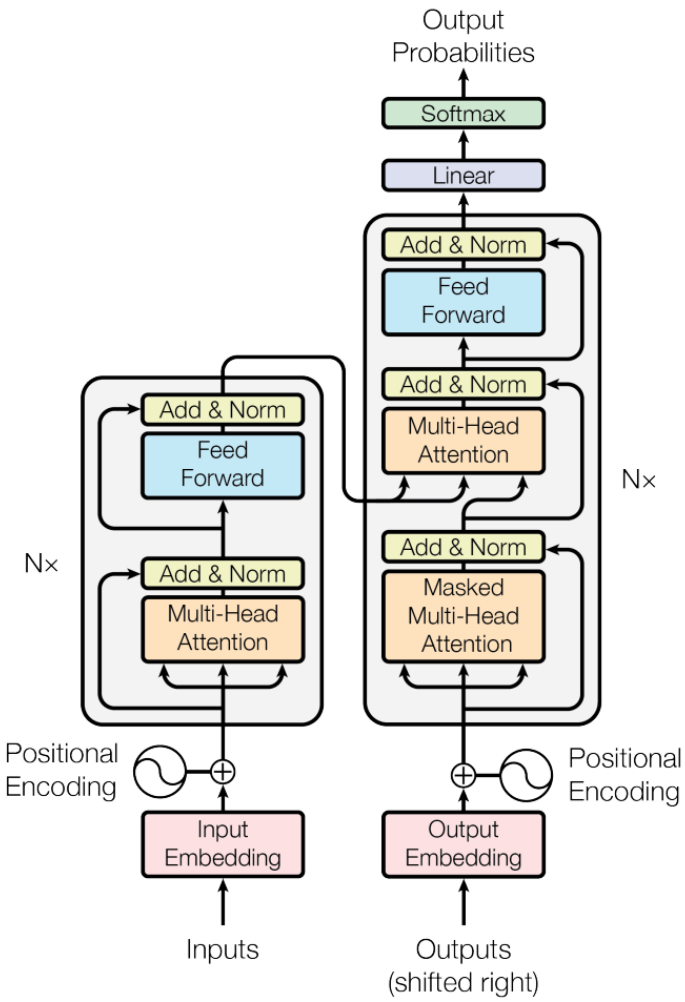
Diferenciando-se dessas abordagens, os *Transformers* conseguem: (1) paralelizar o processamento das sequências, tornando o treinamento mais rápido e (2) representar as sequências de textos com maior qualidade, mesmo que estas sejam longas. Em relação a primeira melhoria, a paralelização do treinamento quando realizada em hardware onde o processamento pode ser dividido e acelerado, como em GPUs, possibilitou o treinamento de modelos cada vez mais largos e escaláveis com quantidade cada vez maiores de dados. O GPT-3, por exemplo, foi treinado em aproximadamente 45 terabytes de dados textuais. A respeito da representação das sequências, o aprimoramento decorre do mecanismo de atenção, que inclusive dá nome ao artigo e também está relacionado a paralelização dos cálculos dos vetores de entrada.

Basicamente, o mecanismo baseado em atenção - ou *Attention* (VASWANI et al., 2017; GUO et al., 2021a)- computa um vetor para cada token da entrada de maneira independente, distinguindo-se do cálculo das redes recorrentes como LSTM, que computam as camadas ocultas da rede baseando-se no token anterior, atualizando o *hidden state* ao longo de toda sequência. Nas camadas de atenção, para cada token da sequência de entrada, na verdade, computa-se dois vetores V e K - value e key - e para cada K calcula-se o produto interno com um vetor de consulta - $query - Q$. O resultado é dividido pela raiz das dimensões de $K - d_k-$ e aplica-se uma função softmax para obter os pesos de cada valor V, como na fórmula a seguir:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.2}$$

Estes pesos, depois de aplicar softmax, serão relativos à uma distribuição de probabilidade que indica um maior ou menor alinhamento do K com o vetor de consulta Q, onde deve-se dar maior "atenção" para extrair os valores V. Todos estes vetores são treinados e ajustados e o mecanismo de atenção é calculado várias vezes ao longo das camadas.

Figura 7 – Arquitetura *Transformer* proposta por Vaswani et al.



Fonte: Vaswani et al. (2017)

O mecanismo de atenção, portanto, possibilitou detectar quais palavras estavam mais alinhadas, sobrepujando o gargalo de informações que era gerado pelas arquiteturas de redes recorrentes, ao indicar para o *decoder* em que palavra ele deveria focar nas camadas do *encoder*. É importante ressaltar, porém, que inicialmente este mecanismo pretendia aprimorar a arquitetura *encoder-decoder* das redes neurais recorrentes, mas sua implementação com algumas modificações tornaram possível uma arquitetura ainda mais potente em qualidade de representação, o que veio a ser o Transformer. Este novo mecanismo de atenção, introduzido

por Vaswani et al. possui três mudanças centrais: (1) *positional encoding*, (2) *self-attention* e (3) *multi-head attention*.

O primeiro foi responsável por substituir uma das vantagens das RNNs em PLN, que era a habilidade de considerar a ordem de uma sequência durante o treinamento, já que a recorrência consistia justamente em considerar um token após o outro e que vai ser desconsiderado pelos *Transformers*. Para mitigar a ausência de senso de ordem, a codificação posicional adiciona a cada *embedding* de entrada adicionando um conjunto de ativações de ondas senoidais variadas, que variam de acordo com a posição do token na sentença e possibilitam que a rede identifique essa ordem antes de processar as entradas.

Self-attention é uma técnica na qual o mecanismo de atenção visto acima é aplicado. No caso das camadas dos *Transformers*, será aplicado entre uma palavra e todas as outras palavras em seu próprio contexto, isto é, comparando as palavras de uma mesma frase ou parágrafo, captando dessa forma não apenas o significado de cada palavra, mas também seu contexto e incorpora essas informações em suas representações, algo que enriqueceu a quantidade de informação que está embutida nas *embeddings*. Já *multi-head attention* é a aplicação a do mecanismo de atenção múltiplas vezes e de modo paralelo, em conjunto. Isso possibilita a representação de vários conjuntos de relações entre as palavras ao invés de apenas um.

No entanto, a composição baseada na arquitetura de camadas em blocos *encoder-decoder*, usadas, por exemplo, em sistemas de tradução automática e de *question-answering* e replicados em arquiteturas como BART (LEWIS et al., 2019) e T-5 (RAFFEL et al., 2019), não é a única possível quando se refere aos *Transformers*. Os modelos de linguagem baseados em *Transformers* usam apenas um dos componentes da arquitetura *encoder-decoder* que são capazes de gerar modelos proficientes per se. Sendo assim, é possível desenvolver arquiteturas baseadas apenas nos blocos de camadas *decoders*, como é o caso do já mencionado GPT-3 e sua versão anterior GPT-2, ou só com blocos de camadas *encoders*, que é o caso do BERT.

Nesta dissertação, visando aplicar a arquitetura *Transformer* em sistemas de detecção de posicionamento em *tweets* em português, optou-se por utilizar o BERTimbau como base, versão em português do modelo de linguagem de domínio geral BERT. Deste modo, a seguir será descrito com maiores detalhes a arquitetura e o funcionamento do BERT.

3.3.1 BERT

Também publicado por pesquisadores do Google, o artigo "*BERT: Pre-training of Deep Bidirectional Transformero artigos for Lan Understanding*" (DEVLIN et al., 2018), apresenta o modelo de linguagem BERT. O BERT toma como entrada sequências de sub-token e gera como saída uma representação dos textos que pode ser utilizada para várias tarefas de PLN com pouco treinamento, já que ele é pré-treinado num corpus textual de tamanho considerável. A novidade do BERT em comparação a outros modelos baseados em *Transformer* e Atenção da época de seu lançamento, como ELMo e o primeiro GPT, era o uso de um *Transformer* bidirecional em seu pré-treinamento, como consta no título do trabalho. Isso significa que as representações textuais geradas pelo BERT são condicionadas conjuntamente pelo contexto à esquerda e à direita da sequência de entrada, em todas as camadas de atenção, incorporando informação de ambos os lados simultaneamente durante o treinamento.

A modelagem ao longo do pré-treinamento, que é realizado de modo não supervisionado, possui uma configuração específica no BERT, que lhe garante essa bidirecionalidade. Tendo em vista que essa arquitetura é composta de camadas *encoders* e que seu uso, diferente dos GPTs, não foi idealizado para geração de textos, o seu treinamento não precisa ser realizado de modo a prever qual é a palavra seguinte a ser gerada. O treinamento, então, é realizado usando duas tarefas: (1) *Masked Language Modeling* (MLM) e (2) *Next Sentence Prediction* (NSP). Na primeira, 15% dos tokens de entrada são mascarados aleatoriamente como no exemplo dado no artigo: "[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]". Neste exemplo, há duas frases subsequentes nas quais um dentre os seis tokens são "escondidos" pelo token "MASK". A tarefa consiste em prever qual é este token e para tal, o modelo terá que usar tanto as informações à esquerda e à direita do token. A segunda tarefa, que é realizada ao mesmo tempo, consiste em prever se a segunda frase está relacionada ou segue à primeira, separadas pelo token [SEP]. No caso do exemplo, é perceptível que são frases relacionadas, então seriam rotuladas no como "IsNext" durante o treinamento e caso não fossem, "NotNext". No treinamento do BERT, 50% das entradas do corpus são de frases relacionadas, enquanto os outros 50% são de frases não relacionadas selecionadas aleatoriamente.

Em relação à representação das entradas do BERT, as *embeddings* de entrada são a soma das *embeddings* dos tokens, as *embeddings* do segmento e as *embeddings* posicionais. As *embeddings* dos tokens são similares ao conceito já conhecido e amplamente aplicado

de *word embeddings* - vetores de palavras - enquanto as *embeddings* de segmento são para identificar de qual frase o token faz parte e as *embeddings* posicionais são para identificar a ordem e posição das palavras já que o *Transformer* não codifica a frase sequencialmente como RNNs e não consegue assim identificar de antemão quão distantes entre si estão os tokens. As entradas, portanto, são construídas através dessas *embeddings* e são passadas para as camadas dos blocos de *transformers* e o modelo é treinado através das duas tarefas que foram descritas acima, MLM e NSP.

As entradas são segmentadas por tokens que não necessariamente representam palavras, mas pedaços de palavras. Isso porque a tokenização das entradas no BERT é realizadas através do método *word piece*, segmentando algumas palavras, como diz o nome, em pedaços, que podem vir a se repetir em outras palavras, como é o caso de alguns sufixos e prefixos comuns quando palavras são flexionadas no gerúndio, infinitivo, aumentativo e outras variações como "#ing" ou "#ndo", gerúndios em inglês e português, respectivamente. Isso é útil porque reduz a quantidade de tokens que terão que ser representados pelo vocabulário do modelo - uma tabela com vetores para cada token - além de ser possível representar palavras que não são comuns ou que sequer foram codificadas no treinamento através dos seus pedaços ao invés de representá-la com um token $[OOV]$ - *out of vocabulary*.

Modelos de linguagem de propósito geral pré-treinados como o BERT são frequentemente utilizados como base para tarefas supervisionadas de aprendizagem de máquina, algo que transformou o panorama das aplicações de PLN nos últimos anos, assim como foi no campo de Visão Computacional com as redes pré-treinadas no *ImageNet*. Este procedimento de *transfer-learning*, que faz uso das representações do BERT, é feito através de um ajuste fino aplicado em dados rotulados específicos da tarefa e o treinamento é realizado adicionando uma camada de classificação às saídas do BERT - basicamente uma regressão logística - e mapeando as sentenças de entrada, que são computadas pelas camadas *transformer* e pela camada de classificação, a uma classe na saída. O ajuste geralmente é realizado em cima do modelo inteiro, mas apenas a nova camada é treinada do zero, o que agiliza o processo de treinamento, que pode ser realizado com menos iterações e com uma menor quantidade de dados rotulados já que aproveita-se das representações adquiridas durante o pré-treinamento.

Nesta dissertação, será aplicada essa abordagem no desenvolvimento de sistemas de detecção de posicionamento. No entanto, o BERT original e suas variações, como RoBERTa (LIU et al., 2019b) e ALBERT (LAN et al., 2020) foram treinados em corpus em inglês e são incompatíveis com o conjunto de dados aqui utilizados, de sentenças de *tweets* em português.

Por este motivo, uma versão do BERT em português brasileiro foi utilizada como base nesta dissertação, descrita na próxima subseção.

3.3.2 BERTimbau

Como explicado acima, modelos de linguagem pré-treinados como BERT apresentam vantagens no desempenho e reduzem a necessidade de grandes quantidades de dados rotulados, o que pode ser bastante valioso especificamente para idiomas onde dados rotulados são limitados mas dados não rotulados são numerosos. Por isso, ao longo dos últimos anos modelos derivados do BERT foram implementados para várias línguas como francês, holandês, espanhol, italiano, além de uma versão multilíngue (mBERT), treinada em mais de 100 línguas. Considerando estes aspectos, (SOUZA; NOGUEIRA; LOTUFO, 2020) treinaram modelos baseados no BERT para o português brasileiro, apelidado de BERTimbau, usando dados do brWaC (FILHO et al., 2018), um grande e diverso corpus de dados extraídos de páginas da web que possui 3.5 milhões de documentos que contêm 2.7 bilhões de tokens, sendo o maior corpus textual disponível em português - após pré-processamento, o corpus de texto possui 17.5 GB.

Em relação aos procedimentos de pré-treinamento e arquitetura, o BERTimbau reproduz de maneira similar o BERT, tendo duas versões: (1) Base - com 12 camadas, 768 dimensões, 12 *attention heads* e 110 milhões de parâmetros (2) Large - com 24 camadas, 1024 dimensões, 16 *attention heads* e 330 milhões de parâmetros. Ambas foram treinadas considerando letras maiúsculas (*cased*) e com tamanho máximo da sequência de 512. O vocabulário foi gerado com o algoritmo BPE a partir de 2 milhões de frases aleatórias extraídas da Wikipédia e possui 30 mil sub-palavras, que são convertidas para o formato *WordPiece*, para serem compatíveis com o BERT. Também replicam as tarefas de treinamento do BERT, MLM e NSP, detalhados na subseção anterior. Uma pequena variação em relação ao BERT original é o uso de "*whole word masking*" na tarefa de MLM, que consiste em "mascarar" todos os tokens relativos à uma palavra caso ela seja selecionada para ser mascarada e seja composta de várias subpalavras. Além disso, a substituição da palavra pelo token [MASK] é realizada em apenas 80% dos casos, tendo 10% de probabilidade de ser selecionada um outro token do vocabulário ou 10% de manter o token original.

Em relação ao pré-treinamento em si, ambos os modelos propostos foram treinados por 1 milhão de iterações, usando taxa de aprendizagem de *1e-4*, *warmup* de 10 mil etapas com decaimento linear da taxa de aprendizagem, seguindo as recomendações do artigo do BERT

original. Para o BERTimbau Base os pesos foram inicializados a partir do mBERT, utilizando as *embeddings* compatíveis com o vocabulário e usando um *batch* de tamanho 128 com sequências de 512 tokens durante o treinamento. Estes detalhes são importantes tendo em vista que neste trabalho foi realizada uma adaptação de domínio em cima do conjunto de *tweets*, utilizando este modelo com continuação do pré-treinamento, que será especificado na próxima seção do capítulo.

Antes de adentrar na próxima seção, é importante ressaltar que o BERTimbau superou o desempenho do BERT multilíngue e outros modelos como LSTM e BiLSTM quando ajustado para tarefas de Reconhecimento de Entidades Nomeadas (NER), Similaridade Textual de Sentenças (STS) e Reconhecimento de Implicação Textual (RTE). Além disso, os autores fazem uma avaliação do impacto da quantidade de iterações em relação ao desempenho em tarefas na qual foram realizadas o *fine-tuning* e demonstram como o desempenho aumenta de forma não-linear em relação a quantidade de iterações, com retornos decrescentes à medida que o pré-treinamento progride.

3.4 *TRANSFER LEARNING*

Há uma ampla variedade de tarefas de PLN, desde classificação de sentimento, reconhecimento de entidade nomeada, *question-answering*, tradução automática e assim por diante. O traço comum em todas essas tarefas é que elas exigem algum conhecimento geral sobre a linguagem como, por exemplo, saber quando uma palavra é um verbo ou um nome próprio, que são úteis independentemente da aplicação específica. O paradigma de *Transfer Learning* (GOODFELLOW; BENGIO; COURVILLE, 2016) é introduzido a partir da constatação de que é muito mais eficaz adquirir este conhecimento uma vez que ele possa ser reutilizado em todas essas aplicações de PLN. Além disso, em aprendizado de máquina também existe uma limitação referente a quantidade de dados rotulados disponíveis considerando que os rótulos são, na maioria das vezes, anotados por humanos, o que pode ser demorado, caro e, em certos casos, precisar do conhecimento de especialistas.

A filosofia do *Transfer Learning* é se aproveitar do amontoado de dados não estruturados e públicos que estão prontamente disponíveis online para pré-treinar um modelo e assim, quando apresentado a um problema no qual há poucos dados rotulados, seja possível transferir o conhecimento adquirido na primeira etapa de treinamento para a segunda. O *pipeline* mais comum em *Transfer Learning* é o pré-treinamento seguido pelo ajuste fino, que consiste em

duas etapas de treinamento aplicadas sequencialmente: (1) treinamento de um modelo de propósito geral usando dados não estruturados e não rotulados, geralmente com uma função de objetivo específica para modelos de linguagem, como é o caso de MLM. (2) Continuação do treinamento com dados rotulados do problema, especializando o modelo para a tarefa de destino específica. Esta técnica é atualmente o estado da arte na maioria das aplicações de PLN. Portanto, o desafio apresentado nesta abordagem é, uma vez que treinamos um modelo de linguagem, como exatamente o aproveitamos para resolver nossa tarefa original? Como fazemos essa transferência de conhecimento? Uma resposta potencial é o paradigma sequencial de pré-treinamento seguido de ajuste fino

A tendência no campo de PLN nos últimos anos tem sido o aumento considerável no uso de recursos para o pré-treinamento de modelos de linguagem de propósito geral, como BERT e GPT. Isso desencadeou no desenvolvimento e ampla utilização de *Transformers* com milhões e até bilhões de parâmetros pré-treinados em corpus textuais extensos e diversos de dados recuperados e extraídos de diferentes domínios da internet, como fóruns e redes sociais, tais quais o *Reddit* e o *Twitter*, de obras literárias e livros digitalizados, de artigos de notícias ou da Wikipédia, entre outros. Deste modo, técnicas de transferência de aprendizagem profundas para PLN surgiram ao ajustar grandes Modelos de Linguagens pré-treinados com arquiteturas de uso geral, substituindo o uso de modelos específicos para cada tarefa. No entanto, o pré-treinamento de modelos de linguagem tem se mostrado semelhante a um objetivo de *multi-tasking*, que permite a aprendizagem *zero-shot* - como muito pouco ou nenhum ajuste - em muitas tarefas.

A pergunta que prevalece, portanto, à volta deste novo paradigma da área de PLN e que será explorada nesta seção é: Os modelos pré-treinados funcionam universalmente, para qualquer tarefa e domínio ou ainda é útil implementar técnicas de *Transfer Learning* como ajuste-fino, adaptação de domínio e aprendizagem via *multi-tasking* para usá-los em tarefas específicas? E se ainda é vantajoso, como fazê-lo?

3.4.1 Adaptação de Domínio

Assim como em *Transfer Learning* , a adaptação de domínio é concernente a uma situação em que o que foi aprendido em uma distribuição X é explorada para melhorar a generalização em uma outra distribuição Y.

Nomeadamente, a adaptação de domínio refere-se, dentro do paradigma de aprendizagem

de representação em redes neurais profundas, a problemas de classificação nos quais a tarefa objetivo - diga-se, o mapeamento ideal $f(x) \rightarrow y$ da entrada para saída - permanece a mesma entre cada configuração, mas a distribuição dos dados em questão é ligeiramente diferente. Considerando, por exemplo, um problema de análise de sentimentos, que consiste em classificar em positivo ou negativo o sentimento declarado em um texto, há uma miríade de tópicos, assuntos e categorias que comentários postados na web podem se enquadrar. Isto é, o vocabulário presente no texto, o estilo - formal ou informal - de escrita, assim como a quantidade de palavras utilizadas em certos espaços podem variar de um domínio para outro, tornando a generalização entre domínios mais difícil (GOODFELLOW; BENGIO; COURVILLE, 2016).

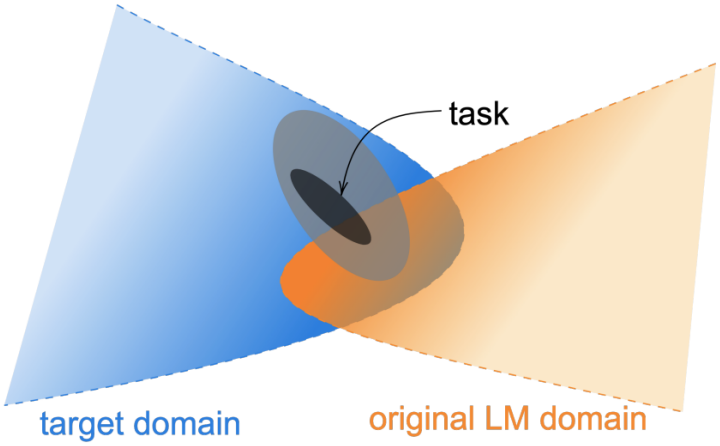
No artigo "*Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*" (GURURANGAN et al., 2020), os autores realizaram um estudo exploratório indagando-se acerca dos benefícios de uma segunda etapa de pré-treinamento, se é vantajoso fazê-la com dados específicos do domínio no qual se pretende-se realizar o ajuste fino ou se o pré-treinamento realizado nos imensos corpus textuais já são o suficiente para adaptação à qualquer tarefa de qualquer domínio. Os autores concluem o estudo afirmando que uma segunda fase de pré-treinamento de domínio, além do ajuste fino para a tarefa em questão, sempre melhora o desempenho do modelo para essas tarefas, tanto em configurações nas quais há muitos recursos quanto onde há poucos. Além disso, adaptação para tarefa também traz ganhos de desempenho, mesmo após ou em combinação com adaptação de domínio. Ambas adaptações são implementadas com a continuação do pré-treinamento - diga-se *Masked Language Modeling* (MLM) - com dados não rotulados de cada conjunto.

Gururangan et al. exploram essa técnica usando o modelo RoBERTa em quatro domínios específicos: notícias, avaliações da Amazon, artigos de Ciência da Computação e artigos de Biomedicina através de dois problemas de classificação para cada um desses domínios, entre eles classificação de sentimento nas avaliações e detecção de tópicos nas notícias, tarefas que se aproximam da que é realizada nesta dissertação.

Através da Figura 8, Gururangan et al. (2020) demonstram como o pré-treinamento contínuo em dados da distribuição de tarefas e da distribuição do domínio podem trazer benefícios para a representação do problema. Na imagem, observa-se como os dados da tarefa são compostos por uma distribuição específica observável, geralmente amostrados de forma não aleatória de uma distribuição mais ampla (elipse em cinza claro) dentro de um domínio de destino ainda maior, que não é necessariamente um dos domínios incluídos no domínio de pré-treinamento do Modelo de Linguagem original - embora uma sobreposição entre eles é

possível.

Figura 8 – Ilustração da justaposição das distribuição dos dados na adaptação de domínio.



Fonte: Gururangan et al. (2020)

Considerando que neste trabalho pretende-se ajustar o BERTimbau para sistemas de detecção de posicionamento em *tweets*, isto é, adaptar um modelo de linguagem de propósito geral para uma tarefa específica (detecção de posicionamento) em um domínio específico (*tweets*), investigar e experimentar as técnicas de adaptação em Modelos de Linguagem, é útil para que essa técnica possa ser implementada no problema tratado neste trabalho.

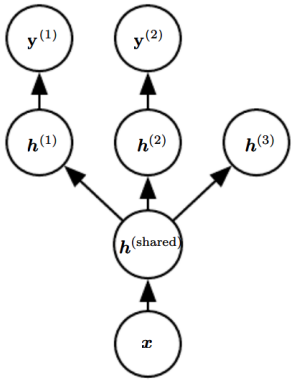
3.4.2 Multi-task Learning

Enquanto em *Transfer Learning* há um processo sequencial no qual a tarefa A é aprendida e depois este conhecimento é transferido e usado para a tarefa B, no aprendizado *multi-task* (CRAWSHAW, 2020) o processo de aprendizagem é realizado simultaneamente, ao configurar uma rede neural para fazer várias tarefas ao mesmo tempo para que, como resultado, cada uma das tarefas auxilie no desempenho de todas as outras tarefas.

Isto é, a aprendizagem *multi-task* é uma forma de melhorar a generalização por meio do agrupamento de amostras decorrentes de várias tarefas, imputando restrições suaves aos parâmetros do modelo. De modo similar aos ajustes dos parâmetros decorrentes de dados de treinamento adicionais que podem gerar uma melhor capacidade de generalização, quando camadas de uma rede são compartilhadas entre tarefas, o resultado pode ser restringir os pesos de modo a produzir uma melhor generalização. Uma forma muito comum de aprendizado

multitarefa é quando diferentes tarefas supervisionadas $(f(x) \rightarrow y)$ compartilham a mesma entrada x e as mesmas representações de nível intermediário h, capturando um conjunto comum de informações (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 9 – Diagrama representando a aprendizagem multi-task.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016)

No diagrama apresentado na Figura 9, é possível observar como um modelo geralmente pode ser dividido em duas partes com tipos específicos de parâmetros associados: (1) Parâmetros específicos da tarefa h^n, que, para obter uma boa generalização, fazem uso dos dados x da tarefa; (2) Parâmetros genéricos $h^{(shared)}$, compartilhados entre todas as tarefas e que podem ser beneficiados através dos dados agrupados de todas as tarefas. Desse modo, normalmente as camadas inferiores de uma rede profunda podem ser compartilhadas entre as tarefas, enquanto os parâmetros específicos de cada tarefa podem ser aprendidos a partir dos parâmetros que produzem uma representação compartilhada.

Do ponto de vista da aprendizagem profunda, a principal hipótese do que torna a aprendizagem multitask benéfica ao desempenho é que dentre os fatores que explicam as variações observadas nos dados específicos a certas tarefas, alguns são partilhados por duas ou mais tarefas (GOODFELLOW; BENGIO; COURVILLE, 2016).

O aprendizado multitarefa faz sentido quando: (1) O treinamento em um conjunto de tarefas pode se beneficiar do compartilhamento de recursos de uma mesma rede, como seus *embeddings* e parâmetros. (2) Em alguns casos: A quantidade de dados que você tem para cada tarefa é bastante semelhante. (3) É possível treinar uma rede neural grande o suficiente para se sair bem em todas as tarefas.

Aprender conceitos para múltiplas tarefas traz dificuldades que não estão presentes no aprendizado de uma única tarefa. Em particular, pode acontecer que diferentes tarefas tenham

necessidades conflitantes. Nesse caso, aumentar o desempenho de um modelo em uma tarefa prejudicará o desempenho em uma tarefa com necessidades diferentes, fenômeno conhecido como transferência negativa ou interferência destrutiva (CRAWSHAW, 2020).

Em relação a aplicação dessa abordagem para Modelos de Linguagem, de acordo com Crawshaw, apesar da popularidade do BERT (DEVLIN et al., 2018), há poucas aplicações do método de codificação de texto para *Multi-task Learning* (MTL). A aplicação de MTL com BERT mais conhecido é a de (LIU et al., 2019a), chamada MT-DNN, que alcançou desempenho do estado da arte em oito das nove tarefas do GLUE (WANG et al., 2019) no momento de sua publicação.

3.5 *SENTENCE TRANSFORMERS*

Além dos métodos de transfer-learning investigados acima, mais uma configuração de treinamento baseada em *Transformers* foi explorada e experimentada nesta dissertação. Baseados no Retweet-BERT (JIANG; REN; FERRARA, 2021), nos experimentos aqui realizados serão incorporadas informações relacionais para ajustar os pesos dos modelos e suas representações textuais de acordo com as interações estabelecidas pelos usuários no *Twitter*. Para tal, como detalhado na seção 2.5, o grafo ponderado e não-direcionado de relações entre os usuários da base foi desenvolvido. O Retweet-BERT, no entanto, é inspirado numa modificação do BERT, o *Sentence Transformer* (S-BERT), proposta por Reimers et al. (2019), que será tratado com mais detalhes, ainda que superficialmente, na próxima subseção para que seja possível compreender o Retweet-BERT e sua implementação nesta dissertação.

3.5.1 S-BERT

Aqui será explorado as diferenças entre um *Transformer* e um *Sentence Transformer* e por quê as representações (*embeddings*) produzidas pelo segundo apresentam uma qualidade diferente que são especialmente úteis para certas tarefas de PLN, principalmente de similaridade ou dissimilaridade entre sentenças.

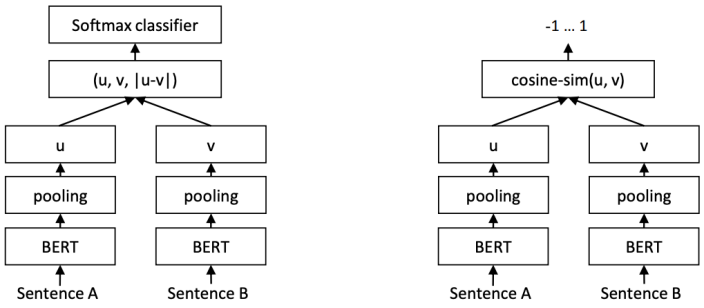
Como relatado nas seções anteriores, os *Transformers* funcionam representam as entradas à nível de subpalavras e tokens, o que pode ser um problema quando pretende-se representar as frases como um todo e fazer comparações entre elas, a nível da sentença. Na tentativa de solucionar este problema, os modelos *cross-encoders* foram apresentados, passando como

entrada para o *Transformer* duas frases simultaneamente e produzindo, através de uma camada de classificação, um valor de saída entre 0 e 1, indicando a similaridade entre o par de frases de entrada, ajustada de acordo com o caso de uso específico. O *cross-encoder*, portanto, assemelha-se ao ajuste fino tradicional realizado com o uso de *Transformers*, adicionando novas camadas *feedforward* nas últimas camadas e realizando o ajuste dos parâmetros. este método funciona em termos de desempenho, mas não é escalável quando busca-se obter as *embeddings* das frases de maneira isolada, já que é necessário computar a inferência da similaridade para cada par de frases, elevando a complexidade temporal da tarefa quando realizado múltiplas vezes. De acordo com o estudo de Reimers et al. (2019), encontrar o par mais semelhante em uma coleção de 10 mil sentenças requer cerca de 50 milhões de cálculos de inferência (65 horas) com BERT. A construção do BERT o torna inadequado para busca de similaridade semântica, bem como para tarefas não supervisionadas como clustering.

Deste modo, tornou-se necessário um modelo que fosse capaz de gerar *embeddings* para sentenças/frases assim como é realizado com tokens e palavras. Com um *Transformer* como BERT, algo que poderia ser feito de maneira a agilizar este processo, era produzir *embeddings* das frases calculando a média entres os valores dos vetores de palavras - também conhecido como *mean pooling*. este método mais simples e mais rápido, no entanto, não gerou resultados bons o suficientes. Conclui-se, assim, que os *Transformers* em si não são adequados para tarefas de grande escala baseadas em sentenças

Para remediar os problemas relativos à tempo e desempenho, Reimers et al. introduziram o primeiro *Sentence Transformers*, S-BERT, que consiste em redes siamesas para produzir *embeddings* de frases semanticamente significativas. S-BERT supera todos os métodos baseados em *transformers* e modelos do estado da arte nas tarefas de similaridade textual, enquanto reduz massivamente a complexidade temporal das operações. A título de comparação, a abordagem reduz o esforço para encontrar o par mais semelhante entre 10 mil sentenças de 65 horas com BERT (*cross-encoder*) para cerca de 5 segundos com SBERT, mantendo a precisão do primeiro. Durante o treinamento, o S-BERT recebe duas sentenças em paralelo através de um *transformer* idêntico (siamês), adiciona um operação de *pooling* para suas saídas, e aprende a prever objetivos de pares de frases pré definidos, assim como medir a semelhança entre as duas frases.

Figura 10 – Arquiteturas do S-BERT propostas por Reimers e Gurevych.



Fonte: (REIMERS et al., 2019)

Na Figura 10 pode-se observar duas arquiteturas de S-BERT, uma com função objetivo de classificação (à esquerda) e outra com função objetivo de regressão, usada para computar a similaridade entre sentenças a partir da similaridade por cosseno aplicada às embeddings da sentença. Ambas arquiteturas reproduzem o grande diferencial dessa abordagem, diga-se a estrutura de rede siamesa, na qual um *Transformer* idêntico - neste caso, o BERT - têm seus pesos compartilhados.

3.5.2 Retweet-BERT

Inspirados no S-BERT, Jiang, Ren e Ferrara propõem o Retweet-BERT, um modelo de sentence *embeddings* que incorpora uma rede de *retweets*. O modelo tem como base a suposição de que os usuários que retuítem uns aos outros são mais propensos a compartilhar ideologias semelhantes. Assim, a intuição do modelo é tornar as *embeddings* das descrições do perfil mais semelhantes para usuários que retuitaram uns aos outros.

Para tal, utilizam o S-BERT, que é capaz de produzir *embeddings* a nível de sentença, e otimizaram a função objetivo, considerando as *embeddings* da descrição do usuário i e cada retuíte positivo do usuário i para j (ou seja, $(i, j) \in E$), de acordo com a fórmula a seguir:

$$\sum_{k \in V, (i,k) \notin E} \max(\|s_i - s_j\| - \|s_i - s_k\| + \varepsilon, 0) \tag{3.3}$$

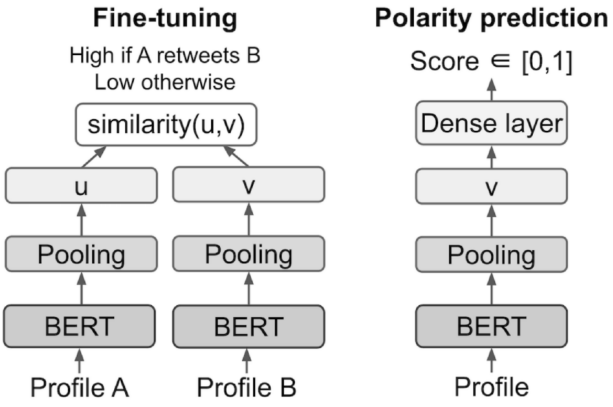
na qual $\|\cdot\|$ é uma métrica de distância e ε é um hiper-parâmetro de margem. Os autores utilizam a configuração padrão do S-BERT, que usa a distância euclidiana e $\varepsilon = 1$. Para otimizar o procedimento de treinamento, eles usaram duas estratégias de amostragem:

1. *Amostragem Negativa* (um-negativo), na qual é selecionado aleatoriamente um outro nó k para cada nó i em cada iteração (MIKOLOV et al., 2013).

2. *Múltipla Amostragem Negativa* (mult-neg), em que os exemplos negativos são todos os outros exemplos no mesmo lote (HENDERSON et al., 2017).
   Por exemplo, se o lote de exemplos positivos for $[(s_{i_1}, s_{j_1}), (s_{i_2}, s_{j_2}), ..., (s_{i_n}, s_{j_n})]$, então os exemplos negativos para par no índice k são (s_{i_k}, s_{j_k}) são todos os $s_{j'_k}$ para $k' \in [1, n]$ e $k' \neq k$.

Vale a pena ressaltar que o Retweet-BERT desconsidera a direcionalidade da rede e considera apenas os vizinhos imediatos de todos os nós. Na prática, porém, os autores demonstram que este modelo é capaz de equilibrar a compensação entre a complexidade do treinamento e o desempenho do teste.

Figura 11 – Diagrama demonstrando funcionamento do Retweet-BERT.



Fonte: (JIANG; REN; FERRARA, 2021)

Na Figura 11, é apresentada uma ilustração do Retweet-BERT, retirada do artigo. Neste diagrama, é indicado à esquerda que primeiro é realizado o ajuste da rede de *retweets* usando uma estrutura de redes siamesas, onde dois BERTs compartilham os pesos. Em seguida, há o treinamento de uma camada densa no topo do BERT para prever a polaridade (direita).

4 METODOLOGIA

Neste capítulo, serão apresentados os detalhes dos métodos implementados neste trabalho, incluindo os métodos e métricas usados para avaliação dos modelos. Na Seção 4.1 serão descritas as métricas utilizadas para avaliar o desempenho dos classificadores. Em seguida, na Seção 4.2 serão descritas as abordagens investigadas utilizando o paradigma de aprendizagem profunda.

4.1 MÉTODOS E MÉTRICAS DE AVALIAÇÃO

4.1.1 Métricas de Avaliação

Neste trabalho usamos diferentes métricas para avaliar o desempenho preditivo das abordagens propostas. Uma vez que o modelo é projetado, para se obter estimativas de desempenho confiáveis, o procedimento usual é treinar o modelo num subconjunto de dados rotulados, sobre os quais o mapeamento $f(x) \rightarrow y$ para indução é obtido e ajustado. Em seguida, utiliza-se um subconjunto de dados diferente para testagem, simulando o algoritmo diante de novas amostras e avaliando seu desempenho, comparando os rótulos reais com os rótulos inferidos para medir a precisão do modelo. Existem várias maneiras de executar o processo de amostragem dos subconjuntos de treinamento e teste, como *holdout*, amostragem aleatória e validação. Um procedimento comum é a divisão do conjunto de amostras de treinamento e teste em uma proporção p para treinamento e $(1 - p)$ para teste, sendo $p = 2/3$ ou $3/4$ (FACELI et al., 2011).

No que se refere às medidas de desempenho, num problema de classificação binário, que é o caso nesta dissertação, é possível qualificar os erros e acertos no conjunto de teste em quatro categorias: Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN). Os exemplos verdadeiros são as predições corretas para as classes positiva ($+$) ou negativa (-), que variam de acordo com a tarefa. No caso deste trabalho, pode-se considerar a classe "a favor" como a classe positiva e "contra" como a classe negativa. Os erros são representados pelos falsos positivos (Tipo I) e negativos (Tipo II). Desse modo, através do conjunto de teste e das predições do classificador, pode-se calcular a acurácia em relação ao resultado esperado, isto é, dentre todas as classificações, quantas o modelo classificou corretamente, através da fórmula a seguir:

$$\text{Acurácia} = \frac{VP + VN(\text{acertos})}{VP + VN + FP + FN(\text{erros e acertos - total})} \tag{4.1}$$

Esta primeira abordagem, embora funcione bem de um modo geral, tem alguns problemas em relação à representatividade das amostras que foram usadas para treinamento e teste e por isso o resultado pode levar a uma interpretação equivocada do desempenho. Se as classes estiverem muito desbalanceadas e uma delas estiver numa proporção de 80% ou 90% em relação ao total dos dados disponíveis, um classificador *dummy* que prever sempre essa classe majoritária, pode obter uma acurácia alta, mesmo quando o classificador está performando mal e não conseguindo discriminar entre as classes. Além disso, mesmo que as classes não estejam desbalanceadas, uma única separação dos dados para treinamento e teste pode ser realizada de modo a desconsiderar a proporção real.

Por esse motivo, existem outras métricas de avaliação e métodos de amostragem para treinamento/teste que pretendem dar maior confiabilidade aos resultados, mesmo quando esses são gerados por conjuntos de dados desbalanceados. Ademais, dependendo do problema de classificação, um tipo de erro pode ser mais relevante do que outro.

Quando um modelo é implementado para prever bons investimentos, conceder crédito, detectar fraudes, *spams* ou desinformação é importante que os exemplos positivos sejam precisamente classificados e detectados, mesmo que gerem alguns erros na classificação dos exemplos negativos. Nesses casos, quando falsos positivos são mais deletérios para o problema do que falsos negativos, o modelo deve desempenhar um bom resultado na classe positiva, então a métrica ideal para mensurar esse desempenho deve considerar dentre todas as classificações da classe positiva reais que o modelo fez, quantas estão corretas. A métrica utilizada para esses casos é a Precisão, que considera apenas os exemplos positivos.

Diante de um problema no qual a situação acima se inverte, isto é, quando os falsos negativos são mais prejudiciais do que os falsos positivos como, por exemplo, na detecção de uma doença viral, onde é necessário detectar com precisão os pacientes com vírus, mas sobretudo ter muita cautela e maior precisão para com os casos negativos. Neste caso, é útil mensurar dentre todas as situações que o classificador indicou como sendo da classe positiva, quantas estão corretas. Essa métrica é conhecida como Revocação ou *Recall*. A fórmula para calcular Precisão e Revocação são descritas a seguir:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{4.2}$$

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{4.3}$$

Para obter um único valor a partir dessas duas medidas de avaliação, utiliza-se o *F1-Score*, que é uma média harmônica da precisão e revocação e é capaz de indicar que o modelo desempenha bem ou não nas duas classes:

$$\textit{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4.4}$$

4.1.2 Validação Cruzada

Para assegurar que todas as amostras sejam consideradas na avaliação do modelo, utiliza-se a validação cruzada, um procedimento de reamostragem usado para avaliar a capacidade de generalização de um modelo diante de um conjunto de dados não visto, ou seja, do conjunto de testes. A técnica de validação cruzada K-*fold* consiste em dividir as amostras de treinamento em K parcelas com quantidade iguais de amostras e depois treinar iterativamente com $K-1$ parcelas e testar em uma das parcelas a cada etapa. Então, uma vez que o processo é repetido K vezes, há K pontuações, das quais pode-se calcular a média. Esta técnica dá resultados mais estáveis porque engloba toda a variedade de amostras disponíveis para avaliar os modelos.

A validação cruzada K-*fold* estratificada é uma extensão da validação cruzada K-*fold* e foi projetada especificamente para conjuntos de dados desbalanceados. O procedimento básico é o mesmo da validação cruzada K-*fold*, uma vez que todo o conjunto de dados é particionado em K partes, com cada partição sendo aproximadamente do mesmo tamanho e, em seguida, uma parte é usada como conjunto de teste e as $K-1$ partes restantes constituem o conjunto de treinamento. O processo é repetido K vezes e utiliza-se como conjunto de teste uma partição diferente a cada vez e depois calcula-se a média das K pontuações. A grande diferença entre a versão estratificada da validação cruzada é que ao invés de particionar os dados aleatoriamente, a divisão dos dados almeja que cada parte mantenha a distribuição original das classes em cada parte.

Levando em consideração que alguns tópicos do conjunto de dados aqui desenvolvido possuem uma distribuição das classes desbalanceada, como apontado no Capítulo 2, será aplicado, para avaliar o desempenho dos modelos, a validação cruzada estratificada, calculando os resultados através acurácia e do *F1-Score* para cada modelo e tópico.

4.2 MODELOS BASEADOS EM *TRANSFORMERS*

É importante ressaltar que embora o BERT e outros modelos baseados em *Transformers* tenham alcançado resultados do estado da arte em uma miríade de domínios e tarefas, como apresentado na Seção 3.3, parte considerável desses resultados são obtidos com Modelos de Linguagem em inglês. Isso decorre do fato que os recursos disponíveis para esse idioma são abundantes, desde corpus textuais acessíveis ao investimento de recursos computacionais para o desenvolvimento de modelos pré-treinados por grandes empresas como *Google*, *Facebook*, *Open AI*, que são tornados públicos nos repositórios do *Hugging Face* [17], por exemplo. Em inglês e francês, por exemplo, há modelos pré-treinados em *tweets* como os já citados BERTweet e BERTweetFR, que são de grande utilidade para tarefas de classificação neste domínio, como é o caso deste trabalho. O BERTweet, inclusive, pré-treinado em milhões de *tweets* em inglês e RoBERTa (LIU et al., 2019b), versão otimizada do BERT estão presentes no topo do *leaderboards* do TweetEval (BARBIERI et al., 2020) - *Unified Benchmark and Comparative Evaluation for Tweet Classification* - que inclui a tarefa de detecção de posicionamento trabalhada nesta dissertação.

No entanto, o mesmo não pode ser dito em relação a outras línguas, como em português, que apesar de possuírem uma grande quantidade de falantes ao redor do mundo e, portanto, dados não rotulados em abundância, não há muitos modelos pré-treinados e poucos recursos são aplicados em seu desenvolvimento, o que leva a modelos pré-treinados disponíves com desempenhos aquém daqueles em idiomas como inglês, francês e italiano. O Modelo de Linguagem mais utilizado e com melhores resultados em português é o BERTimbau, detalhado Seção 3.3.2, e por esse motivo foi o modelo escolhido como base para o desenvolvimento de sistemas de detecção de posicionamento baseado em *Transformers*. Apesar de obter resultados do estado da arte para aplicações em português em tarefas de PLN como NER, RTE e STS, o BERTimbau é um Modelo de Linguagem de propósito geral e não foi concebido especificamente para tarefas de classificação em *tweets*, como o BERTweet ou BERTweetFR, o que torna essa adequação um dos grandes desafios desta dissertação.

Diante deste desafio, foram investigadas as várias abordagens possíveis dentro do paradigma de aprendizagem profunda, através das quais fosse possível aprimorar o desempenho do modelo base, o BERTimbau, o que correspondeu ao escopo das experimentações deste trabalho.

[17] https://huggingface.co/

As abordagens implementadas foram as seguintes:

1. BERTimbau (*portuguese-bert*) *out-of-the-box*, sem nenhuma modificação;
2. BERTimbau com adaptação de domínio;
3. BERTimbau baseado no Retweet-BERT;
4. BERTimbau com adaptação de domínio e baseado Retweet-BERT;
5. BERTimbau com adaptação de domínio e baseado Retweet-BERT, utilizando *multi-tasking* - o ajuste fino é realizado em todos os tópicos ao mesmo tempo.

Nas próximas seções, cada uma destas abordagens será apresentada em detalhes.

4.2.1 Adaptação de Domínio no BERTimbau

Como explicado na Seção 3.4.1, Gururangan et al. apontam em seu estudo sobre adaptação de domínio em *Transformers* pré-treinados que uma segunda fase de pré-treinamento em dados do mesmo domínio, além do ajuste fino para dados específicos da tarefa, tende a melhorar o desempenho do modelo para a tarefa em questão, tanto em configurações onde há muitos recursos quanto onde há poucos. Desse modo, no que se refere à implementação de uma segunda etapa de treinamento, é preciso experimentar algumas configurações dos dados e dos parâmetros, em seguida efetuar um ajuste fino para com a tarefa e depois avaliar uma possível melhora no desempenho do classificador. Isso porque não há solução pré-definida quando trata-se de adaptação de domínio e *transfer learning*, considerando que o êxito da transferência de conhecimento depende do domínio almejado, da quantidade de dados disponíveis e também do modelo pré-treinado usado.

Tendo em vista que o BERTimbau da *Neural Mind* foi treinado com o corpus textual BrWac, atualizado pela última vez em 2017[18], com textos em português oriundos de múltiplos contextos, uma adaptação para o domínio deste trabalho - *tweets* em português comentando as medidas relacionadas à Covid-19 no Brasil - implica em continuar o pré-treinamento com *Masked Language Modeling* (MLM) - a principal função objetivo para o BERT - nos dados não rotulados que sobraram do processo de elaboração da base, um pouco menos de 6 milhões de *tweets*. Para executar a continuação do pré-treinamento e adaptar o domínio com sucesso é

[18] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC#Current_version

importante contemplar certos aspectos, como pre-processamento das entradas e configurações de treinamento, que serão detalhadas na Seção 5.2.

4.2.2 BERTimbau baseado em Retweet-BERT

Para esta dissertação, a rede siamesa do *Sentence Transformer* é implementada através do BERTimbau e tem como entrada as postagens dos usuários, diferenciando-se da aplicação do artigo, usada em descrições de perfis (JIANG; REN; FERRARA, 2021). Outra diferença está na rede de interações utilizadas, que é composta não apenas de *retweets*, mas também de menções e respostas entre os usuários, descrita no capítulo 2. Além disso, o ajuste-fino da segunda etapa é feito para tarefa de detecção de posicionamento que, ao contrário do artigo, ajusta o modelo para classificar a polaridade política dos usuários. Apesar dessas diferenças de implementação, o mesmo *pipeline* é usado e será detalhado a seguir.

Através da readequação do Retweet-BERT para a tarefa de detecção de posicionamento, pretende-se incorporar a rede de interações dos usuários da base de dados no treinamento do modelo, aproveitando-se tanto das *features* textuais quanto dos dados relacionais disponíveis. Para esse fim, um grafo de interações foi gerado - não apenas de *retweets* como é no Retweet-BERT, mas também de replies e mentions - a partir dos mais de 6 milhões de *tweets* da base, como descrito na Seção 2.5.2.

Desse modo, o BERTimbau original e adaptado foram treinados através de uma estrutura siamesa (*Sentence Transformer*), como descrito na Figura 10 da Seção 3.5.2. A pressuposição em torno dessa abordagem é que caso exista uma relação entre um usuário e outro, a função objetivo de similaridade é otimizada para considerar similar as publicações nas quais seus autores se relacionam. Como não há exemplos negativos, de usuários que não se relacionam, e objetivando otimizar o treinamento, utilizou-se *multiple negative sampling*. Assim, o modelo é treinado e as *embeddings* dos textos são ajustadas de modo que fiquem mais similares caso exista um relacionamento entre os usuários, isto é, considerando que devem ser semelhantes os *tweets* dos usuários que possuem um mesmo posicionamento. Em seguida esse modelo é treinado para classificação do posicionamento através de uma camada densa na qual é efetuado um ajuste fino, assim como nos outros modelos.

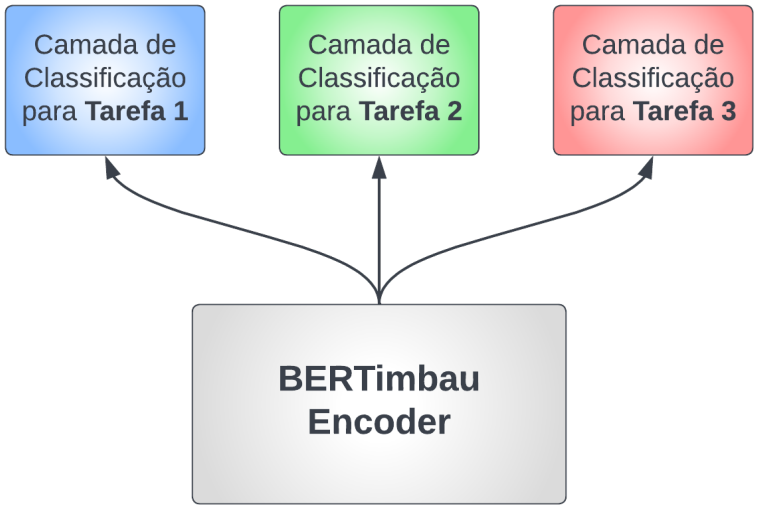
4.2.3 *Multi-tasking* com BERTimbau

Como esclarecido na Seção 3.4.2, o processo de aprendizagem pode ser realizado simultaneamente, ao configurar uma rede neural para fazer várias tarefas ao mesmo tempo para que cada tarefa auxilie no desempenho de todas as outras tarefas. Desse modo, o treinamento em um conjunto de tarefas pode se beneficiar do compartilhamento de recursos de uma mesma rede, como seus *embeddings* e pesos.

Conforme demonstrado no estudo "*Multi-Task Learning with Deep Neural Networks: A Survey*" de Michael Crawshaw (2020), o aprendizado *multi-task* pode oferecer vantagens como maior eficiência de dados, redução de *overfitting* por meio de representações compartilhadas e um aprendizado rápido, aproveitando-se de informações auxiliares.

Existem diferentes abordagens para empregar a aprendizagem multitarefa, que variam de acordo com a arquitetura da rede, a modalidade dos dados e as tarefas envolvidas. A abordagem usada neste trabalho é baseada na adição de uma camada de classificação para cada tarefa por cima do modelo BERTimbau, de modo similar às outras implementações baseadas em *Transformers*, mas dessa vez compartilhando o mesmo *encoder*, como demonstra a Figura 12.

Figura 12 – Representação Visual do Processo de *Multi-tasking* com BERTimbau



Fonte: Autor

Considerando cada tópico do conjunto de dados como uma tarefa específica, experimentou-se aqui realizar um treinamento simultâneo, compartilhando os mesmos blocos *Transformers*, e mantendo apenas a camada de classificação de maneira isolada para cada tarefa. Para

implementação do *Multi-tasking* foi selecionado o modelo de melhor performance entre os *Transformers*, o BERTimbau adaptado baseado na arquitetura do Retweet-BERT.

5 EXPERIMENTOS

Neste capítulo, será apresentada a configuração dos experimentos realizados neste trabalho, além de uma discussão em torno dos resultados obtidos. Na Seção 5.1 serão apresentados os resultados alcançados com os modelos tradicionais. Em seguida, na Seção 5.2 os resultados obtidos com modelos baseados em *Transformers* serão apresentados. Em conclusão, na Seção 5.3 será feita uma discussão em cima dos resultados.

5.1 EXPERIMENTOS COM MODELOS TRADICIONAIS DE APRENDIZAGEM DE MÁQUINA

Neste trabalho, modelos tradicionais de aprendizado de máquina como SVM, *Random Forest* e MLP foram utilizados como *baseline* para avaliar o nível de dificuldade da tarefa de detecção de posicionamento na base de dados construída no capítulo 2 e para comparar sua performance com os modelos do estado da arte que serão detalhados na próxima seção.

Os experimentos com SVM, *Random Forest* e MLP foram executados utilizando o pacote *scikit-learn*, uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python[19]. A *scikit-learn* possui vários algoritmos de classificação, regressão e agrupamento, além de métricas de avaliação dos modelos, técnicas de pré-processamento e métodos de extração de características.

Para representar numericamente os textos dos *tweets* antes de passar para os modelos de aprendizagem, as postagens foram vetorizadas com TF-IDF, transformando os *tweets* em vetores de 30 mil características, cada uma representando um valor associado aos n-gramas (de 1 a 3 palavras) mais frequentes de todo corpus de *tweets*.

O algoritmo de SVM utilizado foi o SVC (*C-Support Vector Classification*) que permite realizar classificação binária e multiclasse em um conjunto de dados. O algoritmo de *Random Forest* utilizado foi o *Random Forest Classifier*, onde é possível controlar o tamanho da subamostra através de um hiperparâmetro do algoritmo. Já o algoritmo MLP utilizado foi o MLP *Classifier*, formado por redes de múltiplas camadas com até 100 camadas escondidas, que possui um treinamento iterativo onde a cada iteração as derivadas parciais da função de perda em relação aos parâmetros do modelo são calculadas para atualização dos parâmetros. Um termo de regularização pode ser adicionado à função de perda para reduzir os valores

[19] https://scikit-learn.org/stable/

Tabela 18 – Resultados (Acurácia e F1-Score) dos Modelos Tradicionais

Tópico	SVM		Random Forest		MLP	
	Acurácia	**F1-Score**	**Acurácia**	**F1-Score**	**Acurácia**	**F1-Score**
Vacinas	0.6318	0.7744	0.7826	**0.8441**	**0.7977**	**0.8410**
Lockdown	0.5235	0.6872	0.7986	0.7992	**0.8320**	**0.8383**
Tratamento Precoce	0.5170	0.6816	0.7882	0.7989	**0.8193**	**0.8269**
CPI da COVID	0.5680	0.7249	0.8181	0.8480	**0.8355**	**0.8576**
Uso de máscaras	0.7343	0.8468	0.8058	0.8817	**0.8507**	**0.9030**
Prefeitos e Governadores	0.7727	0.0000	0.8240	0.4250	**0.8577**	**0.6414**

Fonte: Autor

dos parâmetros do modelo e evitar o *overfitting*. Os valores dos parâmetros utilizados nos experimentos são os valores padrão do *scikit-learn*.

Na Tabela 18 são apresentadas as médias da acurácia e F1-Score obtidas com os modelos SVM, *Random Forest* e MLP através do método de validação cruzada estratificada *5-fold*.

Observa-se na Tabela 18 que os classificadores *Random Forest* e MLP, principalmente este último, obtiveram um melhor desempenho em comparação com o SVM, com uma diferença considerável nos resultados. Tendo em vista que esses modelos foram utilizados como $baselines$ e a variação de seus hiperparâmetros pouco explorada, o SVM, por exemplo, que é sensível à escolha dos seus hiperparâmetros, como, a função de kernel, pode ter tido seu desempenho prejudicado. De modo distinto, o MLP implementado, que possui em sua configuração padrão do scikit-learn 100 camadas escondidas e utiliza ReLu como função de ativação, é um classificador mais robusto, o que pode lhe garantir esse bom desempenho obtido quando comparado aos demais, mesmo sem um ajuste de hiperparâmetros mais robusto.

O desbalanceamento das classes em determinados tópicos, como "Governadores e Prefeitos", aparentam influenciar nos resultados de todos os modelos tradicionais, como pode-se observar na Tabela 18, na qual os F1-Scores deste tópico revela-se muito aquém dos demais tópicos. Essa questão, no entanto, se mostrará transversal a todos os modelos experimentados e por isso, será discutido em detalhes na última seção do capítulo.

Além disso, de antemão, não é possível estabelecer uma correlação entre a quantidade de dados de treinamentos em cada tópico e seu desempenho geral, considerando que o tópico de "vacinas" é o que possui a maior quantidade de amostras e não apresenta bons resultados, en-

quanto "*lockdown*", o segundo maior em termos de amostras, possui um melhor desempenho. O tópico "Uso de Máscaras" possui a menor quantidade de dados de treinamento e apresenta um dos melhores desempenhos. Em seguida, o tópico "Atuação de Governadores e Prefeitos" também possui poucos dados de treinamentos e obteve os piores resultados no F1-Score. Essa análise, porém, será retomada em detalhes no final do capítulo frente aos resultados de todos os experimentos propostos neste trabalho.

5.2 EXPERIMENTOS COM MODELOS BASEADOS EM *TRANSFORMERS*

Dando continuidade aos experimentos desta dissertação, alguns testes foram realizados com modelos de aprendizagem profunda, mais especificamente com modelos baseados em *Transformers* e recorrendo às múltiplas abordagens possíveis dentro desse paradigma, como *transfer learning*, *multi-tasking*, adaptação de domínio e *Sentence Transformers*, descritos nas Seções 3.3, 3.4 e 3.5 e especificados nas Seções 4.2.1, 4.2.2 e 4.2.3 do capítulo anterior.

1. Pré-processamento das entradas: Como o BERTimbau em seu pré-treinamento diferencia letras maiúsculas e minúsculas - *Cased* - o mesmo deve ser feito nessa segunda etapa de pré-treinamento. Também foi levado em consideração no pré-processamento algumas especificidades do pré-treinamento de publicações do *Twitter*, como a remoção e/ou normalização das URLs e "@" em "HTTPURL" e "@USER", seguindo o mesmo procedimento aplicado no BERTweet e no BERTweetFR (NGUYEN; VU; NGUYEN, 2020; GUO et al., 2021b).

2. Configurações do treinamento: Seguindo as configurações do pré-treinamento do BERTimbau original, os dados foram separados em *batches* de 128 amostras. Além disso, objetivando usufruir do conhecimento obtido pelo primeiro treinamento, optou-se por manter o mesmo vocabulário, mesmo que alguns termos cruciais do novo domínio não estivessem contemplados por ele, como "covid", "pandemia", "ivermectina", "*lockdown*", etc. Isso porque: (a) As *embeddings* desses novos termos seriam inicializadas do zero e não iriam aproveitar-se do pré-treinamento já realizado e (b) Como a representação no BERTimbau é a nível de sub-palavras, mesmo que o vocabulário não possua as palavras exatas citadas acima, ele é capaz de representá-las através das sub-palavras como "lock","#down", "co" e "#vid", todas essas presentes no vocabulário do BERTimbau.

Após a adaptação de domínio, um ajuste fino é efetuado através de uma camada densa de classificação no topo do modelo de linguagem gerado, seguindo as recomendações dos autores do BERT Devlin et al. (2019) para *fine-tuning* e o estudo de Sun et al., isto é, com tamanho do *batch* igual a 32 e taxa de aprendizagem igual a $5e-5$, por 3 épocas.

Na Tabela 19 são apresentados as médias da acurácia, desvio padrão e F1-Score obtidas através do método de validação cruzada estratificada 5-fold dos modelos: BERTimbau, BERTimbau adaptado, BERTimbau com Retweet-BERT, BERTimbau adaptado com Retweet-BERT e Multi-task com o melhor modelo.

Tabela 19 – Resultados da Acurácia (AC) e F1-Score (F1) dos 5 modelos com *Transformers*

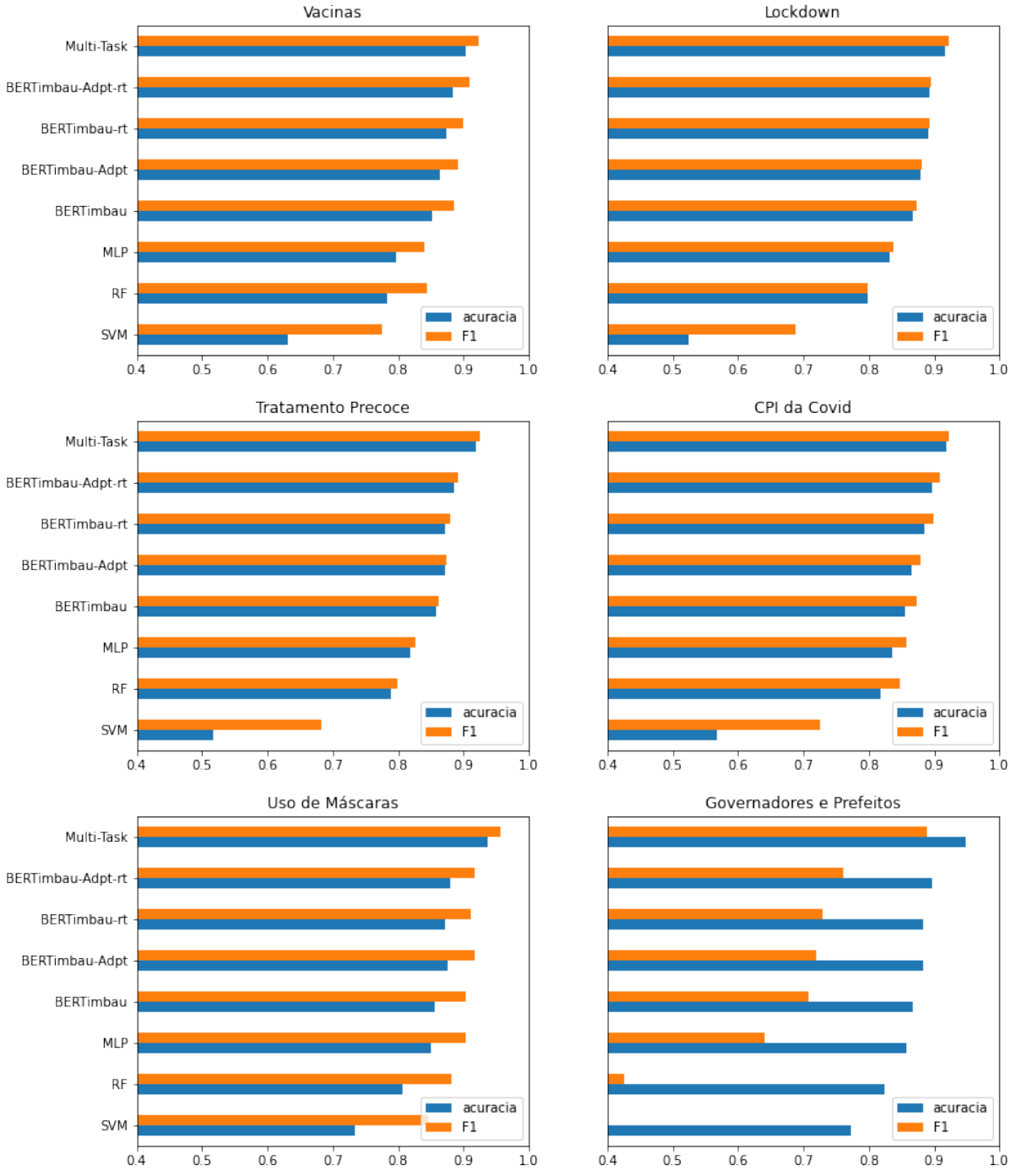
Tópico	BERTimbau		BERTimbau Adaptado		BERTimbau Retweet-BERT		BERTimbau Adaptado Retweet-BERT		Multi-Task	
	AC	F1	AC	F1	AC	F1	AC	F1	AC	F1
Vacinas	0.8524 ± 0.0111	0.8850 ± 0.0095	0.8633 ± 0.0128	0.8917 ± 0.0109	0.8712 ± 0.0056	0.8995 ± 0.0044	0.8770 ± 0.0030	0.9027 ± 0.0029	**0.8842** ± 0.0021	**0.9083** ± 0.0026
Lockdown	0.8671 ± 0.0052	0.8724 ± 0.0062	0.8783 ± 0.0059	0.8817 ± 0.0063	0.8876 ± 0.0071	0.8916 ± 0.0075	0.8936 ± 0.0064	0.8960 ± 0.0063	**0.9046** ± 0.0050	**0.9083**± 0.0039
Tratamento Precoce	0.8581 ± 0.0191	0.8623 ± 0.0188	0.8709 ± 0.0214	0.8741 ± 0.0224	0.8757 ± 0.0093	0.8807 ± 0.0099	0.8848 ± 0.0083	0.8911 ± 0.0072	**0.9000** ± 0.0027	**0.9036**± 0.0029
CPI da COVID	0.8549 ± 0.0135	0.8736 ± 0.0102	0.8655 ± 0.0142	0.8786 ± 0.0137	0.8819 ± 0.0037	0.8972 ± 0.0033	0.8809 ± 0.0093	0.8943 ± 0.0084	**0.8979** ± 0.0025	**0.9095**± 0.0034
Uso de máscaras	0.8558 ± 0.0135	0.9031 ± 0.0105	0.8754 ± 0.0120	0.9164 ± 0.0082	0.8693 ± 0.0135	0.9138 ± 0.0105	0.8797 ± 0.0117	0.9198 ± 0.0090	**0.9203** ± 0.0101	**0.9457**± 0.0074
Prefeitos e Governadores	0.8681 ± 0.0140	0.7070 ± 0.0156	0.8831 ± 0.0082	0.7201 ± 0.0262	0.8829 ± 0.0032	0.7285 ± 0.0197	0.8971 ± 0.0062	0.7613 ± 0.0181	**0.9321** ± 0.0065	**0.8443**± 0.0170

Fonte: Autor

5.3 DISCUSSÃO DOS RESULTADOS

Neste capítulo foram apresentados os experimentos com 8 modelos divididos em dois grupos: 3 modelos tradicionais e 5 modelos baseados em *Transformers*. Na Figura 13 é possível visualizar e comparar o desempenho de cada um dos modelos nos tópicos alvo do problema de detecção de posicionamento desta dissertação.

Figura 13 – Desempenho dos Modelos experimentados em cada tópico do conjunto de dados



Fonte: Autor

Os modelos clássicos selecionados foram SVM, *Random Forest* e MLP. O modelo SVM foi escolhido porque é amplamente utilizado em trabalhos semelhantes na literatura de Detecção de Posicionamento (ALDAYEL; MAGDY, 2021). Os modelos *Random Forest* e MLP foram escolhidos visto que, dentre os algoritmos de aprendizado de máquina tradicionais, apresentam regularmente bons desempenhos e podem ser utilizados como um baseline, estabelecendo um patamar preliminar de desempenho sobre o qual fosse possível traçar comparações com os modelos do estado da arte. Os modelos baseados em *Transformers* foram selecionados considerando que apresentam resultados do estado da arte numa variedade de aplicações na área de processamento de linguagem natural e na tarefa abordada nesta dissertação, como esclarecido na Seção 3.3.

Como observado nas Tabelas 18, 19, e na Figura 13 os resultados experimentais com os modelos baseados em *Transformers* obtiveram melhores resultados e se mostraram mais eficazes para o problema deste trabalho em comparação com os modelos tradicionais. O BERTimbau original, modelo de propósito geral e sem nenhuma modificação, já apresenta uma superioridade de 1 a 6% na acurácia, dependendo do tópico, em relação ao MLP, o melhor dentre os modelos tradicionais e o mesmo se repete para os valores do F1-Score. Os resultados mais equivalentes são os do tópico de "Uso de Máscara", com variações mínimas no desempenho entre o MLP e o BERTimbau.

A adaptação de domínio sobre o BERTimbau, através de uma segunda etapa de pré-treinamento com os *tweets* não rotulados relacionados à Covid-19, se provou vantajosa para a tarefa de detecção do posicionamento nos tópicos avaliados, melhorando o percentual de acerto em ao menos 1% em todos os tópicos.

O mesmo pode ser afirmado em relação a incorporação das características relacionais dos usuários, através do grafo de interações e de uma estrutura em rede siamesa, para ajustar os parâmetros e embeddings do BERTimbau de acordo com uma função objetivo de similaridade e depois aplicar o ajuste fino para detecção dos posicionamentos. Essa abordagem, baseada no Retweet-BERT, aprimorou o classificador em todos os tópicos do conjunto, aumentando o percentual de acurácia em até 2%, tanto na versão base quanto na versão adaptada.

Percebe-se, assim, que as adaptações e adequações sobre o BERTimbau original, através de transferência de conhecimento, advindos tanto de dados textuais não rotulados como de dados relacionais extraídos dos *tweets*, possibilitou um incremento na representação do conhecimento codificados nos pesos do Modelo de Linguagem em relação à tarefa de detecção de posicionamentos em *tweets*, levando a um melhor desempenho neste problema. Esses ajustes e

adaptações experimentados no BERTimbau levaram ao aprimoramento dos modelos em todos os tópicos, sem exceção, o que gerou num desempenho de 5 a 10% mais preciso, dependendo do tópico, em relação aos modelos baselines de melhor desempenho, *Random Forest* e MLP. As melhorias mais significativas na acurácia foram em "vacinas" e "tratamento Precoce", enquanto no F1-Score, apesar de resultados inferiores aos demais tópicos, o desempenho no tópico "Governadores e Prefeitos" teve o maior progresso entre todos os tópicos, de mais de 10 pontos percentuais.

Além dessas adaptações de domínio e ajustes dos parâmetros baseados nas relações entre os usuários na rede de interações da base, ainda foi possível experimentar a abordagem de *multi-tasking* para possíveis aprimoramentos através do compartilhamento de pesos e *embeddings* entre os tópicos, tendo em vista que todos os tópicos consideram mesmo acontecimento, a pandemia de coronavírus no Brasil. Entre todos modelos baseados em *Transformers*, os resultados obtidos com aprendizagem multi-tarefa foram superiores aos demais por alguns motivos: (a) O Modelo de Linguagem utilizado para realizar o treinamento simultâneo entre os tópicos e assim se beneficiar do compartilhamento de *embeddings* e pesos, foi o BERTimbau adaptado e baseado no Retweet-BERT, que apresentava os melhores resultados até o momento; (b) Considerando que sua implementação foi realizada na mesma tarefa (detecção de posicionamento) - mas em diferentes tópicos - e no mesmo domínio (*Tweets* sobre Covid-19 no Brasil), havia uma grande compatibilidade e convergência em termos de contexto, vocabulário e objetivos.

Esta abordagem, diante dessa configuração de consonância entre os problemas, gerou a um aprimoramento do desempenho em 2 a 5% na acurácia em relação aos outros modelos baseados em *Transformers*, dependendo do tópico e levou a significativa melhora de 12 pontos percentuais no F1-Score do tópico "Governadores e Prefeitos". Considerando que o F1-Score deste tópico indicava uma dificuldade de representação da classe minoritária, o compartilhamento dos pesos com outras tarefas e o aproveitamento de informações auxiliares pode ter favorecido uma melhor representação do problema e, consequentemente, a detecção de posicionamento neste tópico.

Dentre os seis tópicos alvo trabalhados, dois se destacam negativamente em relação ao seu desempenho: (a) "vacinas", nos resultados da acurácia, aquém de todos os outros tópicos em todos os modelos, com exceção do SVM; (b) "Governadores e Prefeitos", nos resultados do F1-Score, bem abaixo de todos os outros tópicos, em todos os modelos, sem exceção.

No tópico "vacina", que possui a maior quantidade de amostras do conjunto desta disser-

tação - 68 mil amostras - as classes encontram-se desbalanceadas, mas a classe majoritária não atinge sequer o dobro da minoritária, numa proporção de $63/36$. O desbalanceamento, no entanto, não parece implicar numa deturpação da classe minoritária ou explicar uma menor acurácia frente a outros tópicos, já que os valores do F1-Score são maiores do que o de outros tópicos. Além disso, o tópico "Uso de máscara" possui uma maior desproporcionalidade entre as classes e mesmo assim obteve uma acurácia bem maior. Os resultados, portanto, parecem indicar uma dificuldade de classificação particular ao tópico de vacinação e os *tweets* que o compõem por outros motivos. O tema da vacinação da população foi discutido em excesso ao longo da pandemia desde o final de 2020 até o momento desta dissertação, e os atores políticos envolvidos, principalmente o presidente e seus apoiadores mudaram o discurso em torno do tema várias vezes, refletindo inclusive em 3 mudanças no Ministério da Saúde. Desse modo, os posicionamentos em torno desse tópico provavelmente não possuem tanta coesão como o tópico de "*lockdown*", por exemplo, o qual os apoiadores das medidas do governo federal se opuseram de maneira mais explícita e coesiva durante a pandemia.

Em relação ao tópico que envolve a atuação de "Governadores e Prefeitos", dois motivos aparentam explicar seu baixo desempenho quando avaliado seu F1-Score: (a) Este tópico possui um dos menores conjuntos de treinamento, com apenas 13 mil amostras e é o mais desbalanceado dos tópicos, com a classe majoritária possuindo mais do que o triplo de amostras da classe minoritária, que provavelmente não possui dados suficientes para ser representada adequadamente. (b) Além disso, assim como no caso do tópico "vacinas", há especificidades de como esse tópico foi debatido pelos usuários do *Twitter* - e no debate público em geral - que podem influenciar na complexidade da tarefa de classificação. Isso porque esse tópico de discussão foi introduzido e guiado quase que unilateralmente pelos usuários que eram "a favor" das medidas do governo em relação a pandemia e culpavam a atuação dos governos estaduais e municipais pelos efeitos adversos da situação. Essa predominância de um tipo de posicionamento pode ter afetado a representação da outra classe de posicionamento.

6 CONCLUSÃO

6.1 CONSIDERAÇÕES FINAIS

O desenvolvimento de sistemas de detecção de posicionamento enquanto ferramenta para mineração de opiniões nas redes sociais é uma área de pesquisa emergente e em expansão, diante da ampla utilização dessas plataformas como principal fonte de comunicação na sociedade digital contemporânea.

Desse modo, há um crescente interesse de pesquisa no desenvolvimento de métodos eficazes de detecção de posicionamento em várias áreas de pesquisa como Processamento de Linguagem Natural (PLN), Análise de Redes e Computação Social. A maioria dos esforços direcionados no desenvolvimento dessas aplicações visa produzir tecnologias capazes de gerar conhecimento acerca de uma série de fenômenos da atualidade, como a polarização do debate político, a propagação de desinformação e de boatos, além de possibilitar um abrangente monitoramento da opinião pública.

O presente trabalho está contextualizado na pandemia de coronavírus (Covid-19) na qual foram implementadas uma série de medidas políticas e sanitárias por parte das autoridades responsáveis e também da sociedade civil. No Brasil, de modo semelhante a outros países, esse processo foi profundamente politizado, suscitando discussões polarizadas que inundaram as redes sociais - ocupando agora, mais do que nunca, diante do isolamento social, o centro das discussões sociais e políticas - com opiniões e posicionamentos acerca das medidas adotadas contra a Covid-19 e suas repercussões.

6.2 CONTRIBUIÇÕES DESTE TRABALHO

O objetivo desta dissertação foi desenvolver sistemas de detecção de posicionamento em redes sociais, experimentando as diversas abordagens propostas na literatura para esse problema, analisando suas vantagens e desvantagens, e as aplicando ao contexto da pandemia no Brasil. Além disso, foram realizados experimentos diante de um domínio e idioma diferentes dos existentes na literatura utilizada, algo possível através da construção de uma nova base de dados em português e do treinamento de novos modelos de aprendizagem de máquina.

Através de métodos baseados em regras, uma base de dados de 352.958 *tweets* foi construída e segmentada em tópicos mais específicos relacionados a pandemia no Brasil: "Vacina-

ção", "*Lockdown*", "Tratamento Precoce", "CPI da COVID", "Uso de máscaras" e atuação de "Governadores e prefeitos". A base elaborada será disponibilizada publicamente, conforme os Termos de Serviço do *Twitter*, através do Github para o uso de pesquisadores interessados no problema.

Para o desenvolvimento dos sistemas de detecção de posicionamento, foram experimentadas diversas abordagens, desde modelos tradicionais como SVM, Random Forest e MLP a modelos de aprendizagem profunda, mais especificamente modelos baseados em *Transformers* e recorrendo às múltiplas abordagens possíveis dentro desse paradigma, como *Transfer Learning*, *multi-tasking*, adaptação de domínio e *Sentence Transformers*. Os resultados experimentais obtidos com os modelos baseados em *Transformers* obtiveram melhores resultados em todos os tópicos explorados e se mostraram mais eficazes para o problema abordado neste trabalho em comparação com os modelos tradicionais aplicados.

Através dos experimentos realizados nesta dissertação, é possível afirmar que as adaptações e adequações efetuadas no BERTimbau, através de transferência de conhecimento, advindos tanto de dados textuais não rotulados como de dados relacionais extraídos dos *tweets*, possibilitou um incremento na representação do conhecimento codificado nos pesos do Modelo de Linguagem em relação à tarefa de detecção de posicionamentos em *tweets*, levando a um melhor desempenho neste problema.

6.3 PROPOSTA PARA TRABALHOS FUTUROS

Considerando que a grande maioria das bases de dados de detecção de posicionamento possuem uma classe "None", que representa um texto que não possui um posicionamento em relação ao tópico alvo, pretende-se enquanto trabalho futuro incorporar na base construída neste trabalho *tweets* que representem "nenhum" posicionamento, rotulados como "None". Mesmo que de uma perspectiva sociolinguística (JAFFE et al., 2009), tenha-se argumentado que não existe uma posicionamento completamente neutro, pois as pessoas tendem a se posicionar por meio de seus textos a favor ou contra o objeto de avaliação, a classe "None" pode representar, no contexto de detecção de posicionamento, textos factuais, anedóticos e que, apesar de não serem exatamente neutros, não possuem um posicionamento explícito em relação ao tópico alvo. O desafio dos textos com nenhum posicionamento é que eles são mais difíceis de serem rotulados automaticamente, como foi realizado nesta dissertação, pela ausência de sinais que indiquem a ausência de posicionamento. Esta rotulação "None",

portanto, teria que ser obtida através de uma anotação manual da base construída, um processo bem mais custoso, mas que pode enriquecer a representação das opiniões a serem classificadas.

Algumas outras limitações para com a base de dados construída, que podem ser exploradas em trabalhos futuros são: (1) Análise Exploratória mais conclusiva para determinar a importância dos atributos textuais e relacionais dos *tweets* no processo de geração de *pseudo-labels*; (2) Análise qualitativa da base de dados e dos posicionamentos detectados em cima desta, que investiguem com maior profundidade as opiniões dos usuários acerca das medidas políticas e sanitárias tomadas durante a pandemia no Brasil; (3) Experimentação de metodologias de *topic modeling* para automatizar a identificação de tópicos da base.

Em geral, como a detecção de posicionamento tem sido muito usada como técnica de análise do comportamento dos usuários nas redes sociais para estudar tópicos de discussão que vão desde tópicos políticos, religiosos e sociais, pesquisadores da área (ALDAYEL; MAGDY, 2019); (ALDAYEL; MAGDY, 2021) têm pleiteado a necessidade de incorporar atributos relacionais e "sociais" da rede de interações entre os usuários para uma modelagem mais robusta e precisa dos posicionamentos. Essa concepção, demonstrada nos trabalhos de Aldayel e Magdy (ALDAYEL; MAGDY, 2019), (ALDAYEL; MAGDY, 2021) foi uma referência seminal para este trabalho, principalmente no uso do Retweet-BERT, que faz uso da rede de interações para modelagem, e foi validado ao obter bons resultados na base aqui desenvolvida. No entanto, a incorporação do grafo de relações dos usuários da base no processo de modelagem do Retweet-BERT é muito sútil e apenas considera se houve ou não relação entre os usuários, desconsiderando outros atributos do grafo como direcionalidade e peso das arestas. Em trabalhos futuros, portanto, torna-se necessário uma investigação mais profunda nessa direção, possivelmente com abordagens baseadas em redes (*Network-based*), fazendo uso de *embeddings* de nós, que possam capturar semelhanças na estrutura de rede. Alguns métodos de *network embeddings* que podem ser explorados são o node2vec ((GROVER; LESKOVEC, 2016), GraphSAGE (HAMILTON; YING; LESKOVEC, 2017), *Label propagation* e *Graph Neural Networks* (GNN) (ZHOU et al., 2020) no geral.

Outra possibilidade seria explorar outros aspectos de *Transfer Learning* e Adaptação de Domínio, como a task-adaptation (GURURANGAN et al., 2020), continuando o pré-treinamento não apenas com os dados do domínio, mas também da própria tarefa ou tópico na qual se pretende realizar o ajuste fino. Além disso, considerando que a tarefa de detecção de posicionamento ainda é muito dependente do tópico alvo na etapa de treinamento, trabalhos futuros podem buscar aprimorar o uso das técnicas de *Transfer Learning* para essa tarefa,

almejando uma adaptação de aspectos que não são apenas relativos aos tópicos do posicionamento, e desse modo, melhorar o atual estado dessa metodologia e levar a um classificador de posicionamento geral, como existem em outros problemas de classificação de texto.

REFERÊNCIAS

ADDAWOOD, A.; BADAWY, A.; LERMAN, K.; FERRARA, E. Linguistic cues to deception: Identifying political trolls on social media. In: *Proceedings of the international AAAI conference on web and social media*. [S.l.: s.n.], 2019. v. 13, p. 15–25.

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.

BADAWY, A.; FERRARA, E.; LERMAN, K. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2018. p. 258–265.

BARBIERI, F.; CAMACHO-COLLADOS, J.; NEVES, L.; ESPINOSA-ANKE, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

BEIGMAN-KLEBANOV, B.; BEIGMAN, E.; DIERMEIER, D. Vocabulary choice as an indicator of perspective. In: *Proceedings of the ACL 2010 conference short papers*. [S.l.: s.n.], 2010. p. 253–257.

BELKAROUI, R.; FAIZ, R.; ELKHLIFI, A. Conversation analysis on social networking sites. In: *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*. [S.l.: s.n.], 2014. p. 172–178.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics - investigating language structure and use*. Cambridge Univ. Press, 1998. (Cambridge approaches to linguistics). ISBN 0521499577. Disponível em: <https://www.worldcat.org/oclc/36806763>.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

BRUM, P. V.; TEIXEIRA, M. C.; MIRANDA, R.; VIMIEIRO, R.; JR, W. M.; PAPPA, G. L. A characterization of portuguese tweets regarding the covid-19 pandemic. In: SBC. *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2020. p. 177–184.

CHEN, E.; LERMAN, K.; FERRARA, E. et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, JMIR Publications Inc., Toronto, Canada, v. 6, n. 2, p. e19273, 2020.

CHRISTHIE, W.; REIS, J. C.; MORO, F. B. M. M.; ALMEIDA, V. Detecção de posicionamento em tweets sobre política no contexto brasileiro. In: SBC. *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2018.

CIGNARELLA, A. T.; LAI, M.; BOSCO, C.; PATTI, V.; PAOLO, R. et al. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. In: CEUR. *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. [S.l.], 2020. p. 1–10.

CONOVER, M.; RATKIEWICZ, J.; FRANCISCO, M.; GONCALVES, B.; MENCZER, F.; FLAMMINI, A. Political polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 5, n. 1, p. 89–96, Aug. 2021. Disponível em: <https://ojs.aaai.org/index.php/ICWSM/article/view/14126>.

CRAWSHAW, M. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796, 2020. Disponível em: <https://arxiv.org/abs/2009.09796>.

DAVIS, C. A.; VAROL, O.; FERRARA, E.; FLAMMINI, A.; MENCZER, F. Botornot: A system to evaluate social bots. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016. (WWW '16 Companion), p. 273–274. ISBN 9781450341448. Disponível em: <https://doi.org/10.1145/2872518.2889302>.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

DU-BOIS, J. W. The stance triangle. In: ENGLEBRETSON, R. (Ed.). *Stancetaking in discourse: Subjectivity, evaluation, interaction*. [S.l.]: John Benjamins Publishing Company, 2007, (Pragmatics & Beyond New Series, v. 164). p. 139–182.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: Grupo Gen-LTC, 2011.

FERRARA, E. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, v. 25, n. 6, May 2020. Disponível em: <https://journals.uic.edu/ojs/index.php/fm/article/view/10633>.

FERRARA, E.; CHANG, H.; CHEN, E.; MURIC, G.; PATEL, J. Characterizing social media manipulation in the 2020 u.s. presidential election. *First Monday*, v. 25, n. 11, Oct. 2020. Disponível em: <https://journals.uic.edu/ojs/index.php/fm/article/view/11431>.

FILHO, J. A. W.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The brwac corpus: A new open resource for brazilian portuguese. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. [S.l.: s.n.], 2018.

FRAISIER, O.; CABANAC, G.; PITARCH, Y.; BESANÇON, R.; BOUGHANEM, M. Stance classification through proximity-based community detection. In: *Proceedings of the 29th on Hypertext and Social Media*. New York, NY, USA: Association for Computing Machinery, 2018. (HT '18), p. 220–228. ISBN 9781450354271. Disponível em: <https://doi.org/10.1145/3209542.3209549>.

GARIMELLA, K.; MORALES, G. D. F.; GIONIS, A.; MATHIOUDAKIS, M. Quantifying controversy on social media. *ACM Transactions on Social Computing*, ACM New York, NY, USA, v. 1, n. 1, p. 1–27, 2018.

GHOSH, S.; SINGHANIA, P.; SINGH, S.; RUDRA, K.; GHOSH, S. Stance detection in web and social media: a comparative study. In: SPRINGER. *International Conference of the Cross-Language Evaluation Forum for European Languages*. [S.l.], 2019. p. 75–87.

GIORGIONI, S.; POLITI, M.; SALMAN, S.; BASILI, R.; CROCE, D. Unitor@ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In: *EVALITA*. [S.l.: s.n.], 2020.

GLANDT, K.; KHANAL, S.; LI, Y.; CARAGEA, D.; CARAGEA, C. Stance detection in covid-19 tweets. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. [S.l.: s.n.], 2021. v. 1.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, p. v. 1, n. 12, p. 2009, 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. ISBN 9783540620587.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. In: KRISHNAPURAM, B.; SHAH, M.; SMOLA, A. J.; AGGARWAL, C. C.; SHEN, D.; RASTOGI, R. (Ed.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. [S.l.]: ACM, 2016. p. 855–864.

GUO, M.-H.; XU, T.-X.; LIU, J.-J.; LIU, Z.-N.; JIANG, P.-T.; MU, T.-J.; ZHANG, S.-H.; MARTIN, R. R.; CHENG, M.-M.; HU, S.-M. Attention mechanisms in computer vision: A survey. *arXiv preprint arXiv:2111.07624*, 2021.

GUO, Y.; RENNARD, V.; XYPOLOPOULOS, C.; VAZIRGIANNIS, M. Bertweetfr : Domain adaptation of pre-trained language models for french tweets. *arXiv preprint arXiv:2109.10234*, 2021.

GURURANGAN, S.; MARASOVIĆ, A.; SWAYAMDIPTA, S.; LO, K.; BELTAGY, I.; DOWNEY, D.; SMITH, N. A. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

HAMILTON, W. L.; YING, Z.; LESKOVEC, J. Inductive representation learning on large graphs. In: GUYON, I.; LUXBURG, U. von; BENGIO, S.; WALLACH, H. M.; FERGUS, R.; VISHWANATHAN, S. V. N.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. [S.l.: s.n.], 2017. p. 1024–1034.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. [S.l.]: Springer, 2009. (Springer Series in Statistics).

HENDERSON, M.; AL-RFOU, R.; STROPE, B.; SUNG, Y.-H.; LUKÁCS, L.; GUO, R.; KUMAR, S.; MIKLOS, B.; KURZWEIL, R. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

HO, T. K. Random decision forests. In: IEEE. *Proceedings of 3rd international conference on document analysis and recognition*. [S.l.], 1995. v. 1, p. 278–282.

HOSSAIN, T.; IV, R. L. L.; UGARTE, A.; MATSUBARA, Y.; YOUNG, S.; SINGH, S. COVIDLies: Detecting COVID-19 misinformation on social media. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, 2020. Disponível em: <https://aclanthology.org/2020.nlpcovid19-2.11>.

HOSSAIN, T.; IV, R. L. L.; UGARTE, A.; MATSUBARA, Y.; YOUNG, S.; SINGH, S. Covidlies: Detecting covid-19 misinformation on social media. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. [S.l.: s.n.], 2020.

JAFFE, A. et al. (Ed.). *Stance: sociolinguistic perspectives*. [S.l.]: Oxford University Press USA, Oxford Scholarship Online, 2009.

JIANG, J.; REN, X.; FERRARA, E. Social media polarization and echo chambers: A case study of covid-19. *arXiv e-prints*, p. arXiv–2103, 2021.

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.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012.

KÜÇÜK, D.; CAN, F. Stance detection: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 53, n. 1, feb 2020. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3369026>.

LAN, Z.; CHEN, M.; GOODMAN, S.; GIMPEL, K.; SHARMA, P.; SORICUT, R. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.

LEWIS, M.; LIU, Y.; GOYAL, N.; GHAZVININEJAD, M.; MOHAMED, A.; LEVY, O.; STOYANOV, V.; ZETTLEMOYER, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

LIU, P.; KING, I.; LYU, M. R.; XU, J. Ddflow: Learning optical flow with unlabeled data distillation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 8770–8777.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

MIAO, L.; LAST, M.; LITVAK, M. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. [S.l.: s.n.], 2020.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, v. 26, 2013.

MOHAMMAD, S.; KIRITCHENKO, S.; SOBHANI, P.; ZHU, X.; CHERRY, C. 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.

MUTLU, E. C.; OGHAZ, T.; JASSER, J.; TUTUNCULER, E.; RAJABI, A.; TAYEBI, A.; OZMEN, O.; GARIBAY, I. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, Elsevier, v. 33, p. 106401, 2020.

NGUYEN, D. Q.; VU, T.; NGUYEN, A. T. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

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.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

RAJARAMAN, A.; ULLMAN, J. D. *Mining of massive datasets*. [S.l.]: Cambridge University Press, 2011.

REIMERS, N.; GUREVYCH, I.; REIMERS, N.; GUREVYCH, I.; THAKUR, N.; REIMERS, N.; DAXENBERGER, J.; GUREVYCH, I.; REIMERS, N.; GUREVYCH, I. et al. Sentence-bert: Sentence embeddings using siamese bert-networks. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2019. p. 671–688.

RUMELHART, D.; HINTON, G.; WILLIAMS, R. Learning representations by back-propagating errors. *Nature*, v. 323, p. 533–536, 1986.

SEN, A.; SINHA, M.; MANNARSWAMY, S.; ROY, S. Stance classification of multi-perspective consumer health information. In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. [S.l.: s.n.], 2018. p. 273–281.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 403–417.

SUN, C.; QIU, X.; XU, Y.; HUANG, X. *How to Fine-Tune BERT for Text Classification?* 2020.

TARDE, G. *A opinião e as massas*. 1ª edição. ed. [S.l.]: WMF Martins Fontes, 1992. ISBN 8533600895.

TRABELSI, A.; ZAIANE, O. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2018. v. 12, n. 1.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 1999.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

WANG, R.; ZHOU, D.; JIANG, M.; SI, J.; YANG, Y. A survey on opinion mining: From stance to product aspect. *IEEE Access*, IEEE, v. 7, p. 41101–41124, 2019.

XIE, Q.; LUONG, M.-T.; HOVY, E.; LE, Q. V. Self-training with noisy student improves imagenet classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 10687–10698.

ZHOU, J.; CUI, G.; HU, S.; ZHANG, Z.; YANG, C.; LIU, Z.; WANG, L.; LI, C.; SUN, M. Graph neural networks: A review of methods and applications. *AI Open*, v. 1, p. 57–81, 2020. ISSN 2666-6510. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666651021000012>.

ZHU, L.; HE, Y.; ZHOU, D. Hierarchical viewpoint discovery from tweets using bayesian modelling. *Expert Systems with Applications*, Elsevier, v. 116, p. 430–438, 2019.