



UNIVERSIDADE FEDERAL DE GOIÁS
Escola de Engenharia Elétrica, Mecânica e de Computação

Giulia Borges de Oliveira
Moacir Batista Tavares

**Análise de dados da internet para
acompanhamento das eleições
presidenciais brasileiras em 2022**

Goiânia
2023



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem resarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as):

Giulia Borges de Oliveira

Moacir Batista Tavares

Título do trabalho:

Análise de dados de internet para acompanhamento das eleições presidenciais brasileiras em 2022

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [] SIM [x] NÃO¹

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 24/08/2023, às 16:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Giulia Borges De Oliveira**, **Discente**, em 24/08/2023, às 16:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Moacir Batista Tavares**, **Discente**, em 24/08/2023, às 16:36, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3991487** e o código CRC **C01E070D**.

Referência: Processo nº 23070.024548/2023-11

SEI nº 3991487

**Giulia Borges de Oliveira
Moacir Batista Tavares**

**Análise de dados da internet
para acompanhamento das
eleições presidenciais brasileiras
em 2022**

Tese apresentada ao Programa de Graduação da Escola de Engenharia Elétrica, Mecânica e de Computação da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Engenheiro de Computação.

Área de concentração: Engenharia de Computação
Orientador: Prof. Dra. Deborah Silva Alves Fernandes

Goiânia

2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Oliveira, Giulia Borges de
Análise de dados de internet para acompanhamento das eleições
presidenciais brasileiras em 2022 [manuscrito] / Giulia Borges de
Oliveira, Moacir Batista Tavares. - 2023.
XVI, 16 f.: il.

Orientador: Profa. Dra. Deborah Silva Alves Fernandes.
Trabalho de Conclusão de Curso (Graduação) - Universidade
Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de
Computação (EMC), Engenharia da Computação, Goiânia, 2023.
Bibliografia.
Inclui gráfico, tabelas.

1. Análise de sentimento. 2. Eleições presidenciais brasileiras. 3.
BERT. 4. Machine learning. I. Tavares, Moacir Batista. II. Fernandes,
Deborah Silva Alves, orient. III. Título.



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

ATA DE AVALIAÇÃO DE PROJETO FINAL CURSO

| | | |
|---------------------|---------------------|--|
| () Eng Elétrica | () Eng Mecânica | (X) Eng Computação PFC 1 () PFC 2 (X) |
|---------------------|---------------------|--|

Título do Trabalho

Análise de dados de internet para acompanhamento das eleições presidenciais brasileiras em 2022

Banca Avaliadora

| | |
|----------|-------------------------------|
| Membro 1 | Deborah Silva Alves Fernandes |
| Membro 2 | Sanderley Ramos Pires |
| Membro 3 | Cássio Vinhal |

Data da Defesa

23/08/2023

Discentes

| Matrícula | Nome |
|-----------|---------------------------|
| 201703674 | Giulia Borges de Oliveira |
| 201708988 | Moacir Batista Tavares |

NOTAS

| | Membro 1 | | | Membro 2 | | | Membro 3 | | | | |
|-----------|-----------|-----|-----|----------|-----|-----|----------|-----|-----|-----|----|
| | Matrícula | NPT | NTE | NAA | NPT | NTE | NAA | NPT | NTE | NAA | |
| 201703674 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 201708988 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |

NPT - Nota plano de trabalho;

NTE - Nota do trabalho escrito;

NAA - Nota de apresentação e arguição

Para Eng. Elétrica, Mecânica e PFC2 da Eng. Da Computação: $NF = 0,1 \times NPT + 0,45 \times NTE + 0,45 \times NAA$

Para PFC1 da Eng. Da Computação: $NF = 0,3 \times NPT + 0,7 \times NAA$

* A aprovação do(s) aluno(s) está condicionada à apresentação do trabalho final ao orientador com todas as correções sugeridas pela banca.

Observações:

Preencher com modificações solicitadas, caso existam. Em caso de reaprovação, informar a justificativa.



Documento assinado eletronicamente por **Deborah Silva Alves Fernandes, Professor do Magistério Superior**, em 23/08/2023, às 15:50, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Cassio Dener Noronha Vinhal, Professor do Magistério Superior**, em 23/08/2023, às 16:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Sandrerley Ramos Pires, Professor do Magistério Superior**, em 24/08/2023, às 08:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3985664** e o código CRC **3211985B**.

Análise de dados de internet para acompanhamento das eleições presidenciais brasileiras em 2022

Giulia Borges de Oliveira*, Moacir Batista Tavares†, graduandos em Engenharia de Computação. Deborah Alves Fernandes‡, Professora Associada *EMC/UFG. †EMC/UFG. ‡INF/UFG. E-mails: gi_borges@discente.ufg.br*, moacir.batista@ufg.br†, deborah.fernandes@ufg.br‡

Resumo—Em períodos de eleições a opinião pública e a manifestação dos sentimentos da população com os candidatos é significativa para as previsões dos resultados. Com o desenvolvimento e o crescimento das redes sociais criou-se um espaço para os usuários manifestarem seus pensamentos, julgamentos, elogios e críticas. Dessa forma, as redes sociais conquistaram o interesse para o estudo de análise de sentimento em diversos temas. Este projeto consiste na análise de sentimentos de tweets coletados para acompanhamento e predições das eleições presidenciais brasileiras de 2022. Com a aplicação de técnicas de pré-processamento e o modelo BERT, em uma base de 400 tweets, obteve-se 86% de F1-Score. Portanto, o estudo permitiu observar o cenário político polarizado de 2022 nos dados coletados e classificados. Além disso, o modelo aproximou-se dos resultados reais das eleições no primeiro turno e conseguiu prever o candidato eleito no segundo.

Palavras-chave—Análise de sentimento, eleições presidenciais brasileiras, BERT, machine learning

Abstract—During election periods, public opinion and the expression of the population's sentiments towards the candidates are significant for predicting the results. With the development and growth of social networks, an environment has been created for users to express their thoughts, judgments, praises, and criticisms. Thus, social media platforms have gained interest for sentiment analysis studies on various topics. This project aims to analyze the sentiments of collected tweets to monitor and predict the Brazilian presidential elections of 2022. With the application of pre-processing techniques and the BERT model, on a basis of 400 tweets, an 86% F1-Score was obtained. Therefore, the study allowed observing the polarized political scenario of 2022 in the collected and classified data. In addition, the model approached the real election results in the first round and managed to predict the elected candidate in the second.

Index Terms—Sentiment analysis, brazilian presidential elections, BERT, machine learning

I. INTRODUÇÃO

As informações são cada vez mais disseminadas, de forma acelerada e frenética, em quantidades ilimitadas, e de maneira irrestrita. As notícias do dia a dia têm a necessidade de serem repassadas em instantes, pois os usuários têm pressa, e precisam estar antenados a cada nova notícia que surge. Com essa agilidade como motivador, o *Twitter*, tem revolucionado as possibilidades de irrigar o mundo com conteúdos simultâneos da realidade mundial.

Apesar dos debates na rede social online não serem necessariamente políticos, eles também abordam questões de interesse público, podendo alcançar as pessoas, e, consequentemente, levando a política a elas [1]. Visto que, o acesso à informação

passou a ser personalizado, o que atende aos usuários das plataformas digitais que não desejam ver publicações e anúncios publicitários inadequados às suas preferências, de forma que a maior parte da curadoria é efetivada pelos algoritmos de Inteligência Artificial [2]. Ao ter o primeiro contato com o tópico ele pode tornar-se recorrente, pois os serviços online buscam entender o perfil do consumidor para oferecer conteúdos que facilitam no processo de escolha sobre o que tem maior probabilidade de agradá-lo, caminhando para a hiper personalização [3].

O Serviço Federal de Processamento de Dados (SERPRO) [4] informa que as eleições realizadas em 2014 reforçaram a importância do uso das mídias sociais como um espaço de comunicação e de debate político. De acordo com um levantamento conduzido pelo Instituto Datafolha, 46% dos internautas compartilham conteúdos sobre o tema eleições políticas em seus perfis e 19% assumem que a escolha de candidato foi influenciada por conteúdo das mídias sociais [5]. Além disso, a Pesquisa Nacional por Amostra de Domicílios "Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal", relativa ao ano de 2021, aponta que houve um aumento no número de domicílios com internet [6], mostrando assim a importância e relevância do mundo digital.

Sendo assim, o presente projeto consiste na análise de sentimento dos dados colhidos de mensagens publicadas na rede social *Twitter*. Compondo uma pequena base de dados que será utilizada pelo algoritmo classificador *BERT*, com o intuito de compor predições e análises das eleições presidenciais brasileiras de 2022. Para tanto, utilizou-se também um conjunto de dados de 2018 e avaliou-se o desempenho com os dados de 2022.

Este trabalho está organizado da seguinte forma:

Seção II: abordará trabalhos relacionados;

Seção III: uma visão geral da base teórica de processamento de linguagem natural, com foco em análise de sentimento, e aprendizado de máquina utilizados;

Seção IV: aborda-se o detalhamento da implementação da solução: a coleta dos dados utilizados, o pré- processamento e a rotulação, o processo de classificação e as métricas utilizadas;

Seção V: os resultados e discussões; e na Seção VI a conclusão e as referências ao final.

II. TRABALHOS RELACIONADOS

Nesta seção, serão abordados trabalhos relacionados a utilização de tweets para o estudo de previsões de eleições,

abordando os modelos clássicos de algoritmos *de machine learning* e modelos mais atuais.

Praciano [7] propôs um *framework* para prever os resultados de eleições brasileiras baseado em *tweets*, utilizando como treinamento *tweets* da eleição presidencial de 2014. Os dados foram coletados na semana da eleição utilizando *Web Scraping* feito em *Python*, coletando mais de 100.000 *tweets*. Outros dados também foram salvos como: autor, data, *retweets*, favoritos, menções, *hashtags*, *id* e *link* de publicação. Realizou-se um pré-processamento nos dados para a remoção de informações desnecessárias, normalização e lematização dos textos. Para a classificação utilizou-se *TextBlob* e *OpLexicon* em conjunto com *Sentilex*, no qual o último obteve melhor resultado. Os dados foram classificados em positivo, negativo e neutro. Aplicou-se aprendizado supervisionado onde 20% dos dados dividiu-se para teste e validação e os outros 80% foram utilizados para a fase de treinamento. Como classificadores, selecionaram os algoritmos: *SVM* com *grid search* que apresentou os melhores resultados, *Naive Bayes*, regressão logística e árvore de decisão que obteve os piores resultados e desempenho computacional. As métricas consideradas foram: acurácia, precisão, *recall* e *F1-score*. Escolheram o *SVM* por seus resultados, prevendo corretamente mais de 99% de *tweets* positivos, 86% de *tweets* negativos e 70% de *tweets* neutros, além de obter 98% nas métricas selecionadas. Realizaram também, uma previsão para a eleição presidencial de 2018, onde classificou-se 54.47% de *tweets* positivos para Bolsonaro contra 41.38% para Haddad, e o resultado real das eleições constou 55,13% dos votos para Bolsonaro e 44,87% para Haddad.

Na busca de explorar os algoritmos de *machine learning*, Hasan, Moin, e Karim [8] focaram na adoção de três analisadores de sentimentos para determinar o de melhor acurácia para aprendizado de sentimentos relacionados a eleições. Com o intuito de comparar análise de sentimentos entre *lexicons* (*WWSD*, *SentiWordNet*, *TextBlob*) na busca pelo melhor, a equipe reuniu opiniões através de *hashtags* relacionadas a partidos políticos utilizando a *Tweepy API* a quais representavam a opinião política das pessoas. Os *tweets* coletados eram pertencentes a usuários do Paquistão e, portanto, foi necessária uma conversão de Urdu para Inglês. Após isso, realizou-se um pré-processamento de dados removendo informações desnecessárias, como Urdu romanizado, *urls*, e símbolos de caracteres especiais. Dessa forma, de 100.000 *tweets* coletados permaneceram apenas 6250 *tweets*, que serviram de entrada para os *lexicons* selecionados para obter o cálculo de polarização, onde a pontuação de -1, 0 e 1 representa sentimentos negativos, neutros e positivos, respectivamente. Para validar os resultados obtidos dos *lexicons*, a equipe utilizou a *Waikato Environment for Knowledge Analysis* (Weka) para implementar os algoritmos supervisionados, *Naive Bayes* e *SVM*. Os resultados alcançados para o cálculo de polarização e análise de sentimento foram de 79% de acurácia para o *WSD* utilizando *Naive Bayes* e de 62.67% para o *TextBlob* utilizando *SVM* com 70% dos dados.

Em seu estudo, aplicando técnicas de análise de sentimentos, Cristiani, Lieira e Camargo [9] avaliam se existe uma relação entre a opinião dos usuários do *Twitter* e o

resultado final das eleições presidenciais de 2018 no Brasil. A coleta de dados foi realizada através da *API* pública do *Twitter* e com duas horas de duração durante eventos específicos como debates, entrevistas e os dias de eleição resultando em 903.518 *tweets* coletados. No total, 600 *tweets* foram anotados, manualmente pela equipe levando em conta o conteúdo da mensagem, em três polaridades: positivo (apoio ao candidato), neutro e negativo (rejeição ao candidato). Dessa forma, criou-se um dicionário de palavras após realizarem a divisão e o agrupamento de tweets para seus respectivos candidatos. No pré-processamento, as técnicas aplicadas consistiam na preparação e padronização, tokenização, remoção de *stopwords*, lematização e *TF-IDF* utilizando bibliotecas em *Python*. Ao testar os dois classificadores, *Naive Bayes* e *SVM*, o segundo foi eleito como o melhor para a classificação de todos os *tweets* do conjunto de dados. Dessa forma, os resultados demonstraram que a abordagem é útil para conduzir pesquisas e estudos sobre a opinião dos usuários da rede social referentes às eleições do Brasil.

Brito e Adeodato [10] propuseram um modelo de predição eleitoral utilizando dados de redes sociais em conjunto com pesquisas eleitorais e buscaram prever as eleições presidenciais dos Estados Unidos da América (EUA) de 2016 e do Brasil de 2018. Coletaram dados do *Twitter*, *Facebook* e *Instagram*, considerando-os como: quantidade de curtidas, *retweets* e comentários. Aprimoraram o conjunto de dados de modo que o resultado de uma enquete foi representada por uma função de engajamento. O período de coleta aconteceu durante um longo intervalo, desde 1º de janeiro até o dia da eleição para a previsão do Brasil e por 1 ano para a previsão dos EUA. Utilizaram Regressão Linear, RNA com parâmetros fixos manualmente e RNA com busca de parâmetros em *grid* para a previsão dos votos. Para a avaliação dos resultados utilizaram o erro médio absoluto e o erro percentual médio absoluto. Para a previsão no Brasil, consideraram cinco candidatos, totalizando 18976 *posts* e 21 enquetes, 11 do Datafolha e 10 do Ibope. Aplicaram os dados a uma rede artificial *MLP-PB*, treinada com dados até a última previsão, um dia antes da eleição, e fizeram uma previsão com a parcela de votos final, no qual compararam os resultados com a parcela de votos real. Os resultados mostraram-se melhores que as enquetes sozinhas, tendo a RNA com parâmetros fixos obtido os menores erros. Para a previsão nos EUA, foram 12558 *posts* de dois candidatos e 366 enquetes. Aplicaram os mesmos processos do experimento para a previsão no Brasil. O melhor modelo foi a regressão linear com menores erros, mas tendo resultados piores que a predição das eleições brasileiras.

Ao analisar três modelos de *machine learning*, Capellaro e Caseli [11] possuíam como objetivo responder os seguintes questionamentos: “o desempenho do *BERT* na análise de polaridade supera o de *Naive Bayes* e *SVM* utilizando o mesmo *corpus* e pré-processamento de Cristiani et al. [9]” e “o pré-processamento refinado, com o auxílio da ferramenta *Enelvo*, traz ganho de desempenho de *BERT*, *Naive Bayes* e *SVM* na análise de polaridade?”. Dessa forma, dividiram o pré-processamento em dois tipos: original (onde removeram *hiperlinks*, *retweets* e o texto foi convertido para minúsculo) e *Enelvo Raw* (normalização feita no modo *Enelvo Raw*, mas

removendo *emojis* e *retweets*). Os autores utilizaram o modelo *BERTimbau* realizando o *fine tuning* com o *corpus* de Cristiani et al. contendo dados da eleição presidencial do Brasil de 2018. Como resultado, obtiveram o melhor *F1-score* de 96,6% com o *BERT* e *Enelvo Raw* sem *emoji* e *retweet*, enquanto o *SVM* pontuou 57,8% e o *Naive Bayes* 53,1%. Portanto, o desempenho do *BERT* superou os outros dois algoritmos e, ao aplicar o pré-processamento refinado com o auxílio da ferramenta, houve sim um ganho de desempenho.

III. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, serão apresentadas e descritas as bases teóricas para a execução deste trabalho.

A. Processamento de Linguagem Natural

Segundo cientistas [12], a origem de como a linguagem surgiu é desconhecida e uma incógnita para muitos, ainda que haja diversas hipóteses. Entretanto, sabe-se que a forma de se comunicar modifica-se com o passar dos anos. Do período Pré-histórico a atual Era da Informação [13] a humanidade desenvolveu meios de transmitir seus pensamentos e até mesmo registrá-los de pinturas rupestres à sinais e sons, de sinais e sons à escrita no papel, da escrita à impressão, da impressão ao meio digital que proporciona numerosas possibilidades. É perceptível como a evolução da linguagem é também acompanhada pela tecnologia. Dispositivos, como os *smartphones* e os computadores facilitam, armazenam e intermedeiam a comunicação entre os humanos.

Em suma, a maneira das pessoas de se expressar por escrita e diálogos em diferentes idiomas e culturas pode ser compreendida como linguagem natural que foi desenvolvida naturalmente com o decorrer da sociedade. Essa tornou-se interesse de estudo envolvendo as áreas de ciência da computação, inteligência artificial e linguística, mundialmente reconhecida por Processamento de Linguagem Natural (PLN), do inglês *Natural Language Processing* (NLP). Assim dizendo, PLN é compreendido como uma área que possui interesse em possibilitar que os computadores possuam a habilidade de compreender textos e falas, assim como o ser humano é capaz, combinando a linguagem formal (baseando-se em estruturas gramaticais) com estatística, modelos de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*) [14].

Com a explosão na quantidade de dados gerados e disponíveis na internet, presume-se que cerca de 2.5 quintilhões de dados são gerados diariamente [15]. Além disso, segundo a Faculdade XP Educação [16], estima-se que 80% sejam dados não estruturados, ou seja, postagens em redes sociais, comentários, artigos, imagens, áudio e vídeo, entre outros. Esses dados podem ser mais desafiadores de serem processados e analisados em comparação com os dados estruturados que possuem uma organização conhecida, como tabelas de banco de dados, planilhas, arquivos *CSV/XML/JSON*, entre outros. Portanto, eles permitem realizar operações como buscar informações específicas e realizar análises estatísticas com facilidade. Nesse sentido, com o Processamento Natural de Linguagem, a máquina pode processar a linguagem humana,

inserida por texto ou voz, transformando-a em dados estruturados, e por fim entender o significado, a intenção e o sentimento elaborado pela pessoa.

A Figura 1 ilustra uma linha do tempo contando alguns momentos importantes da história da área de Processamento de Linguagem Natural. No início do campo de estudo de linguagem natural as abordagens tradicionais e os algoritmos utilizados eram baseados em máquinas de estado, sistemas de regras, lógica, modelos probabilísticos, e modelos de espaço vetorial [17]. Dessa forma, figuras de destaque como Alan Turing, Noam Chomsky e Andrei Markov contribuíram para o avanço das pesquisas da época de forma que os primeiros sistemas desenvolvidos obtiveram sucesso, como *SHRDLU* [18].

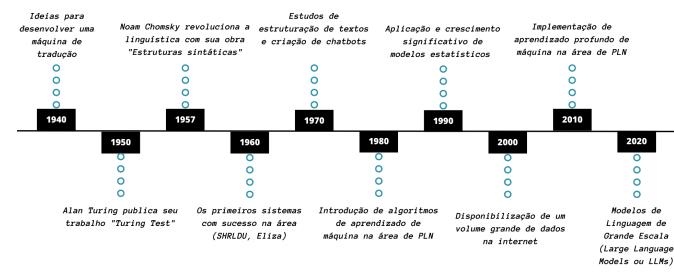


Figura 1: Linha do tempo do NLP.

Adaptado de Kumar [19]

O estudo na área continuou se desenvolvendo e ramificando em campos especializados como Sumarização Automática, Análise de Discurso, Reconhecimento de Entidade Nomeada (NER) e Reconhecimento Óptico de Caracteres (OCR) [20]. Além disso, outro aspecto relevante que contribuiu para o avanço do Processamento de Linguagem Natural foi a popularização da internet e sua transição gradual de um meio restrito ao meio acadêmico e empresas para o público em geral. Essa transição gradual foi acompanhada pela disponibilização de conexões de banda larga, o que resultou na redução de falhas de acesso à rede e quedas de conexão, possibilitando a transmissão de conteúdo pesado, como imagens, vídeos e músicas. Como resultado, programas de navegação automática começaram a surgir, sendo o *Google* um exemplo conhecido, que combinou uma interface simples com um modelo inovador de extração de dados da web (*data mining*) revelando ser mais do que um catálogo da rede: era sua porta de entrada [21].

Análise de Sentimento

O avanço das tecnologias e a disseminação da internet permitiram a expansão do acesso à internet de alta qualidade, impulsionando a popularização dos computadores pessoais e dispositivos móveis. Esse cenário propiciou o crescimento dos comércios eletrônicos (*e-commerce*) e o surgimento de redes sociais, como *Orkut*, *Facebook*, *LinkedIn* e *Twitter*, que viabilizaram o estudo das opiniões expressas no meio digital sobre produtos, temas sociais e comportamento humano. Nesse

contexto, uma nova vertente de pesquisa surgiu na área de Processamento de Linguagem Natural, conhecida como Análise de Sentimentos.

O professor Bing Liu [22], de ciência da computação, define que a análise de sentimento é pesquisada principalmente em três níveis:

- Em nível de documento: o objetivo é classificar se uma opinião expressa em um documento direcionado a uma única entidade possui sentimento positivo ou negativo. Portanto, não utiliza-se em documentos que avaliam ou comparam múltiplas entidades.
- Em nível de sentença: o objetivo é determinar se uma sentença expressa um sentimento positivo, negativo ou neutro. Este nível relaciona-se com o trabalho de classificação de subjetividade, ou seja, distinguir sentenças objetivas que expressam informações factuais de sentenças subjetivas que expressam opiniões e pontos de vista.
- Em nível de entidade e aspecto: diferente dos outros dois níveis, o objetivo é observar a opinião em si e não a estruturação da linguagem (documentos, parágrafos, sentenças, cláusulas ou frases). Baseia-se na ideia de que uma opinião consiste em um sentimento (positivo ou negativo) e um alvo (entidade a qual é direcionada a opinião).

Dessa forma, Liu define opinião como uma quíntupla (e, c, s, d, t) composta por uma entidade (e), (c) como uma característica de (e), (s) o sentimento (positivo, negativo, ou neutro) sobre a característica (c) da entidade (e), (d) trata-se do detentor desse sentimento e (t) o tempo em que a opinião é expressada. Com base nessa definição, o professor conclui que o objetivo da análise de sentimento consiste em: dada uma opinião em um documento D , encontrar os elementos da quíntupla (e, c, s, d, t) em D . Entretanto, reforça-se que essa prática é aplicada apenas em opiniões simples, para opiniões comparativas uma outra definição é aplicada. Pois, as análises tornam-se mais complexas, e.g. “Suco de uva é melhor que suco de laranja”, nessa frase há dois alvos de opinião e o sentimento expresso dependerá do alvo em questão a ser considerado. Portanto, neste trabalho será abordado apenas opiniões simples.

B. Aprendizado de Máquina

Segundo a IBM [23], o aprendizado de máquina, do inglês *machine learning*, pode ser compreendido como uma área da inteligência artificial (IA) e da ciência da computação que se concentra no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão. Com os avanços tecnológicos em armazenamento e capacidade de processamento propiciou-se criações inovadoras baseadas em *machine learning*, como mecanismos de recomendação, tradutores *onlines* e assistentes virtuais.

Outros dois termos bastante envolvidos na área são: redes neurais (*neural networks*) e aprendizado profundo (*deep learning*). O primeiro representa um subcampo do aprendizado de máquina, consistindo em técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de

organismos inteligentes e que adquirem conhecimento através da experiência [24]. O segundo é um subcampo das redes neurais, no qual permite a entrada de dados brutos para ser processado por redes que possuem muitas camadas exigindo muito tempo no treinamento de grandes volumes de dados, e computação de alto desempenho [25].

Classificadores

A área de *machine learning* busca propor soluções também aos problemas de classificação que através das características identificadas e presentes em um conjunto de dados atribuem um valor discreto como saída [26]. Por exemplo, ao analisar uma opinião procura-se determinar se o sentimento expresso é positivo ou negativo não havendo meio termo. Dessa forma, um classificador é um algoritmo treinado para categorizar ou classificar entradas em classes ou categorias predefinidas, aprendendo padrões e relações nos dados de treinamento para atribuir as classes aos futuros novos dados [27]. Esse aprendizado pode ser realizado de duas formas: supervisionado ou não supervisionado. “Os algoritmos de aprendizagem supervisionada relacionam uma saída com uma entrada com base em dados rotulados. Neste caso, alimenta-se o algoritmo com pares de entradas e saídas conhecidos, normalmente na forma de vetores. Para cada saída é atribuído um rótulo, que pode ser um valor numérico ou uma classe. O algoritmo determina uma forma de prever qual o rótulo de saída com base em uma entrada informada. No caso dos algoritmos de aprendizagem não-supervisionada, não é atribuído um rótulo para os dados de saída. Com base em um número grande de dados, o algoritmo busca padrões e similaridades entre os dados, permitindo identificar grupos de itens similares ou similaridade de itens novos com grupos já definidos” [28]. A seguir, apresenta-se mais sobre o classificador utilizado neste trabalho.

Transformer é uma rede neural desenvolvida originalmente por Vaswani et al. e apresentada no artigo ”Attention Is All You Need”[29]. O *transformer* possui uma arquitetura composta por dois elementos principais: codificadores (*encoder*), que processam uma sequência de entrada aprendendo o contexto, e decodificadores (*decoder*), responsáveis por gerar uma saída, utilizado em traduções ou geração de texto por exemplo. O *transformer* utiliza um mecanismo de autoatenção (*self-attention*), permitindo que o modelo dê maior importância a diferentes partes da entrada e consiga relacionar elementos distantes em uma série de dados.

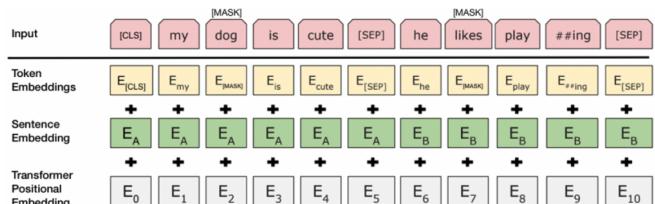


Figura 2: Representação da entrada de uma arquitetura BERT

Fonte: Devlin, 2018 [30]

O BERT (*Bidirectional Encoder Representation from Transformers*) [30] [31] é um avançado modelo de linguagem que

utiliza a camada de encoder do *transformer*, desenvolvido pelo *Google* em 2018. Trata-se de um modelo com pré-treinamento bidirecional, que o permite compreender o contexto de uma palavra considerando as outras palavras tanto à esquerda como à direita em uma frase. Conforme observado na Figura 2, a entrada do *BERT* é composta pela soma de 3 vetores:

- *Token embeddings* - Uma representação numérica de cada *token* contido na frase.
- *Segment embeddings* - Indica a que frase um determinado *token* pertence.
- *Position embeddings* - Denota a posição dos *tokens* no texto.

O treinamento do *BERT* consiste em dois passos, o pré-treino e o ajuste fino (*fine-tuning*), como pode ser observado na Figura 3.

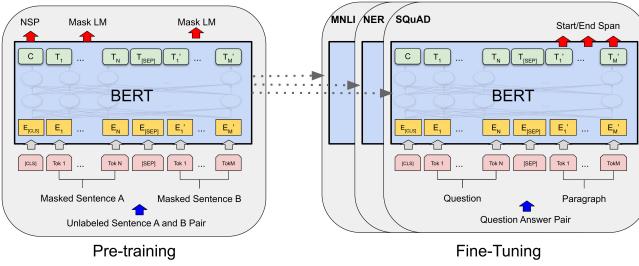


Figura 3: Fluxo de treinamento e *fine-tuning* do BERT

Fonte: Devlin, 2018 [30]

Durante o pré-treino, ele é treinado de maneira não supervisionada com dados não rotulados, com o objetivo de compreender o contexto geral dos dados. Essa etapa exige grandes quantidades de dados, geralmente em um idioma específico, no qual ele é treinado em duas etapas diferentes, sendo elas:

- 1) *Masked Language Modeling(MLM)*, onde alguns tokens são mascarados e ele é treinado para prever a palavra correta.
- 2) *Next Sentence Prediction (NSP)*, onde o *BERT* é treinado para entender se uma sentença é sequência de outra, aprendendo relações entre as duas sentenças.

Após o pré-treino, o modelo pode ser usado em contextos específicos realizando um ajuste fino (*fine-tuning*). No ajuste fino ele é inicializado com os parâmetros do modelo pré-treinado e é treinado com dados específicos e rotulados, onde poderá ser usado para um objetivo definido.

IV. MATERIAIS E MÉTODOS

As atividades realizadas no decorrer do experimento podem ser observadas na Figura 4. As etapas no diagrama da figura serão descritas nas subseções abaixo.

A. Dados

1) Coleta:

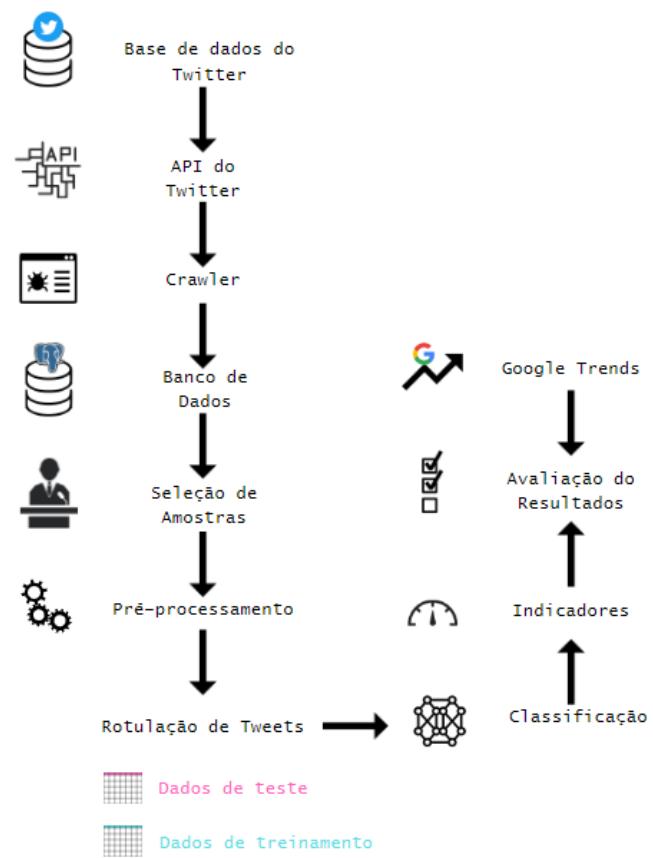


Figura 4: Desenho do Experimento

a) *Tweets*: A predição do resultado das eleições presidenciais brasileiras de 2022 foi executada com dados do *Twitter* e comparadas com os dados do *Google Trends* e de pesquisas tradicionais. Este trabalho foi realizado tendo como base a monografia de [32].

Algumas características da rede social online *Twitter* o torna atrativo para a pesquisa, tais como:

- Os *tweets* são publicações de no máximo 280 caracteres e expressam de forma breve e direta a opinião de seus usuários, simplificando a análise de sentimento;
- Os perfis dos usuários são, em sua maioria, públicos, permitindo que as mensagens publicadas sejam visualizadas pelos demais;
- Os eleitores buscam informar-se e manifestar-se gerando grande atividade na rede social [33];
- A área acadêmica demonstra-se interessada em utilizar o *Twitter* como fonte de dados para predizer eventos, e.g. [34] [9] [35].

O *crawler*, ou *web crawler*, segundo a definição do dicionário de Cambridge é um programa de computador que, de forma automática, realiza buscas de informações na internet, normalmente para indexar (listar) o conteúdo procurado. Dessa forma, com o uso do *crawler* disponibilizado por Fernandes et al. em conjunto com a *API* de pesquisa do próprio *Twitter* [36] a coleta de dados foi realizada. Informa-se ao *crawler* um filtro (conjunto de hashtags, termos e usuários) relacionado a política no Brasil com foco nos candidatos à presidência.

Dessa forma, as informações encontradas são coletadas e armazenadas em um banco de dados *PostgreSQL*.

O período de coleta foi dividido em dois, para contabilizar os *tweets* do primeiro turno de eleição iniciou-se no dia 16 de setembro de 2022 até o dia 02 de outubro de 2022 totalizando 4.268.153 *tweets*. Ao passo que para contabilizar os do segundo turno, iniciou-se no dia 26 de outubro de 2022 até o dia 30 de outubro de 2022 coletando no total 4.281.992 *tweets*.

b) Google Trends: A ferramenta possibilita verificar o volume de pesquisas no serviço de busca do *Google* para palavras-chave específicas ao longo do tempo. Além disso, é possível comparar resultados, verificar as tendências pelos termos, entre outras funcionalidades. Dessa forma, é possível observar o interesse do público através das pesquisas pelos candidatos analisados, e verificar as tendências das buscas para analisar o sentimento para cada um deles.

Para este trabalho, utilizou-se os termos "Lula" e "Jair Bolsonaro" sugeridos pela ferramenta, com o filtro no tópico de "Lei e governo". Selecionou-se o período entre junho e dezembro de 2022 para análise antes, durante e após as eleições. Comparou-se o interesse ao longo do tempo definido para os dois candidatos. Analisou-se também, os temas das pesquisas mais consultadas durante o intervalo selecionado, para cada candidato, com o objetivo de verificar se os termos eram positivos, negativos ou neutros.

2) Pré-processamento: Algumas línguas como o inglês, o português e o francês possuem em sua gramática a presença de artigos, acentuação, conjunções e preposições que facilitam a conexão de sentenças e a fluidez da leitura para os humanos. Além disso, no meio digital encontra-se a modificação dessas línguas para uma comunicação mais veloz e informal muitas vezes, principalmente em redes sociais. Assim, surgem as abreviações, a ressignificação de algumas palavras e o surgimento de novas gírias. Entretanto, para a máquina essas características não são de fácil entendimento e processamento exigindo assim uma etapa que execute a preparação, organização e estruturação dos dados a serem utilizados.

O pré-processamento é uma etapa de extrema relevância durante o processamento de linguagem natural para auxiliar e facilitar a transformação dos dados em uma entrada compreendida pela máquina. Dessa forma, apresenta-se algumas técnicas [37] a serem utilizadas:

- Remoção de ruído: nesta técnica procura-se remover tudo que não acrescenta muito significado à análise conduzida, por exemplo: pontuação, caracteres especiais, *hashtags*, *emojis*, menções (citações), *urls* e tags *HTML*.
- Normalização: nesta técnica procura-se manter um padrão para as palavras convertendo todas em maiúsculas ou minúsculas para que o processo de análise seja mais preciso.
- *Stopwords*: nesta técnica procura-se remover palavras que são comuns na maioria das frases, servindo para conectar ou estruturar o texto, mas não contribuindo para a definição do sentimento. São considerados *stopwords*: conectivos, artigos, preposições, pronomes.
- Lematização: consiste em reduzir as palavras ao seu radical, buscando agrupar palavras diferentes com o mesmo

sentido para reduzir a dimensionalidade dos dados, facilitando o trabalho de classificação do modelo.

Foi definida a linguagem *Python* como ferramenta por ser muito popular para trabalhar com pré-processamento de dados. Com uma ampla gama de bibliotecas e ferramentas disponíveis, *Python* oferece uma variedade de recursos para limpar, transformar e preparar dados antes de serem utilizados em análises.

Com auxílio da biblioteca *SQLAlchemy*, os *tweets* foram carregados do banco de dados e transferidos para um *Dataframe* utilizando a biblioteca *Pandas*, a qual permite realizar análises e manipulações dos dados com maior praticidade. É comum na coleta automática a presença de dados de outros idiomas ou dados que não se tratam do assunto.

Em análise realizada pelo Senado Federal [38], o número de eleitores brasileiros no exterior bateu recorde em 2022 com cerca de 697 mil brasileiros aptos a votar, assim o número é 39,21% maior que o da última eleição. Ao realizar uma busca constatou-se 139 países diferentes mencionados, apesar de uma boa parte dos dados não constarem com informações de localização. Dessa forma, analisou-se através do mapa de calor a origem dos *tweets* coletados.

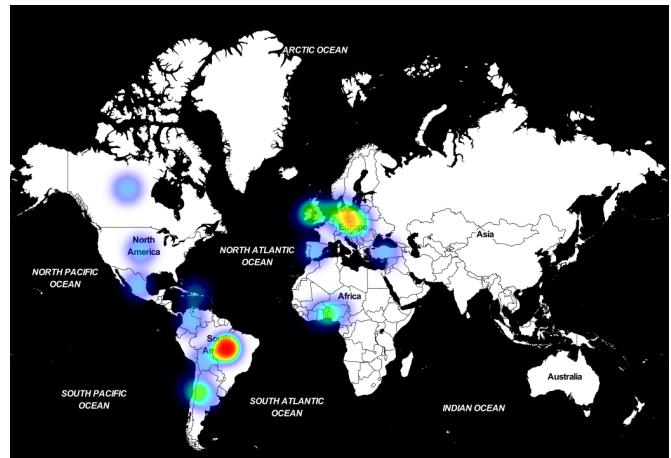


Figura 5: Mapa de Calor dos Dados Coletados

Na Figura 5 é possível constatar que há uma grande concentração de *tweets* com origem no Brasil, Argentina e países da União Europeia. Portanto, com o objetivo de minimizar a quantidade de *tweets* de outras línguas e também remover *tweets* não relacionados ao contexto da análise, utilizou-se o filtro por candidato demonstrado na Tabela I. Assim, textos que não continham nenhuma palavra contida no filtro foram removidos.

| Candidatos | Tokens |
|----------------|---|
| Lula | 'lulaoficial', 'lula', '#13', 'inacio', '@pt' |
| Jair Bolsonaro | 'jairbolsonaro', 'bolsonaro', 'jair', '#22' |

Tabela I: Filtro utilizado para separar os *tweets* a serem utilizados.

O filtro contém o nome dos candidatos, nome de usuário do *Twitter* e termos associados ao contexto das eleições, como

partido e número identificador do partido político dos candidatos. Optou-se por buscar apenas *hashtags* da identificação numérica dos candidatos, pois os números podem ser inseridos em vários contextos. Como também, por adicionar marcações ao partido do candidato Lula, uma vez que o partido e o candidato são comumente associados. Devido à quantidade de dados disponíveis para o treinamento, em uma tentativa de simplificar a análise, os *tweets* que mencionam os dois candidatos foram desconsiderados, devido à maior complexidade de analisar opiniões com múltiplos alvos.

Em seguida, foi feita uma análise nos dados textuais dos *tweets*, onde percebeu-se a necessidade de tratá-los com o intuito de limpar e simplificar esses dados. Uma grande quantidade de *retweets* e ruídos, como links e menções, estão entre os principais problemas. A Figura 6 apresenta uma nuvem de palavras do *corpus* sem a realização de pré-processamento, evidenciando a presença dos problemas mencionados. Em uma análise mais profunda foi possível observar que os textos dos *retweets* estavam incompletos, muitas vezes cortados no meio da frase, portanto, optou-se por removê-los.

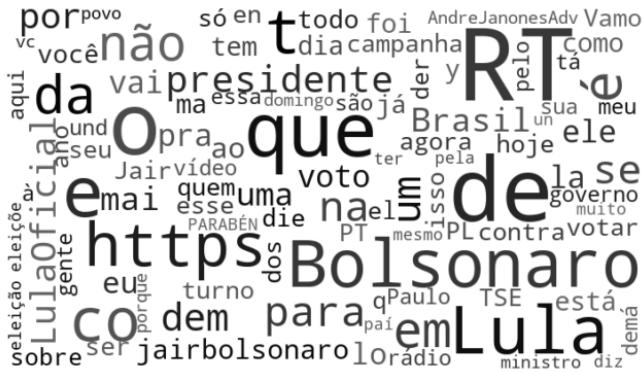


Figura 6: Nuvem de palavras - *Corpus* original

A biblioteca Pandas em conjunto com expressões regulares auxiliou na normalização do *corpus* e na remoção de *hashtags*, *links*, caracteres especiais e menções a outros usuários, com exceção dos perfis oficiais dos candidatos analisados (@lulaoficial e @jairbolsonaro). Dessa forma, com os dados pré-processados realizou-se a separação em *datasets* por candidato, utilizando o filtro da Tabela I. A divisão dos conjuntos dados estão representados na Tabela II.

| Candidatos | 1º Turno | 2º Turno |
|----------------|----------|----------|
| Lula | 287.654 | 361.970 |
| Jair Bolsonaro | 201.315 | 230.444 |

Tabela II: Quantidade de *tweets* dos candidatos separados em turno.

Cada *dataset* foi salvo em um arquivo CSV para classificação posterior. Para observar os termos mais frequentes e relacionados a cada candidato, gerou-se as nuvens de palavras nas Figuras 7 e 8.

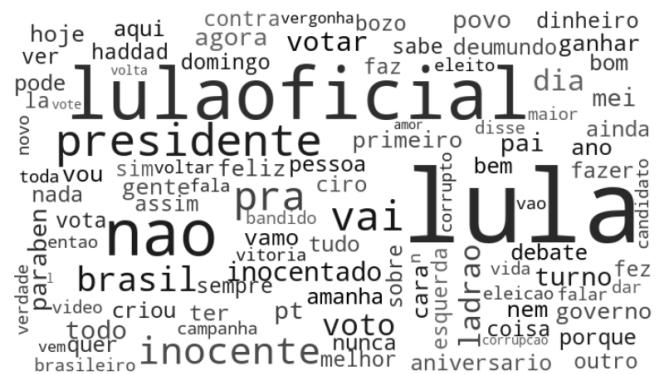


Figura 7: Nuvem de palavras - Lula.



Figura 8: Nuvem de palavras - Bolsonaro.

No geral, algumas etapas de pré-processamento que buscam simplificar o *corpus* e agrupar palavras que são necessárias em classificadores mais simples podem variar de acordo com a tarefa específica a ser realizada no *BERT*, em alguns casos pode até dificultar a classificação. Isso ocorre devido ao modelo ser bidirecional, o que significa que, durante a fase de treinamento, ele aprende informações dos lados esquerdo e direito do contexto de um *token* [39]. Logo, inicialmente aplicou-se o pré-processamento completo, entretanto, percebeu-se dificuldades na classificação do *BERT*, provavelmente, devido às etapas de lematização e remoção de *stopwords* que podem contribuir para a perda de contexto das frases. Optou-se então por remover essas etapas. Dada a quantidade de *tweets* para treino, as etapas de remoção de ruído e normalização foram mantidas com o objetivo de manter o texto mais simples e facilitar a classificação.

Com o intuito de observar os diferentes tipos de pré-processamento citados, separou-se três opções com aplicações de técnicas diferentes:

| Tweet original |
|--|
| <i>Lula é a paz e a dignidade que o povo brasileiro quer e merece!!!</i> |

Tabela III: Exemplo de *tweet* sem pré-processamento.

- Pré-processamento completo: realizou-se todas as técnicas citadas anteriormente (remoção de ruído, normalização,

remoção de *stopwords* e lematização.

| Tweet pré-processado completo |
|--|
| <i>lula paz dignidade povo brasileiro querer merecer</i> |

Tabela IV: Exemplo de *tweet* pré-processado

- Pré-processamento com remoção de caracteres especiais: realizou-se apenas as técnicas de remoção de ruídos e normalização.

| Tweet pré-processado com remoção de caracteres especiais |
|---|
| <i>lula é a paz e a dignidade que o povo brasileiro quer e merece</i> |

Tabela V: Exemplo de *tweet* pré-processado.

- Pré-processamento sem remoção de caracteres especiais: realizou-se apenas as técnicas de remoção de ruídos, com exceção dos caracteres especiais (! / - / . / , / ?) e normalização.

| Tweet pré-processado sem remoção de caracteres especiais |
|--|
| <i>lula é a paz e a dignidade que o povo brasileiro quer e merece!!!</i> |

Tabela VI: Exemplo de *tweet* pré-processado.

3) *Rotulação*: A rotulagem de dados (ou anotação de dados) requer a identificação de dados brutos (ou seja, imagens, arquivos de texto, vídeos) e a inclusão de um ou mais rótulos nesses dados para especificar o contexto deles aos modelos de *machine learning*, a fim de permitir que esses modelos façam previsões [40].

A rotulação de dados desempenha um papel fundamental para o *fine-tuning* do modelo *BERT* para análise de sentimentos. Através da rotulação, é possível fornecer um conjunto de dados de treinamento com informações sobre o sentimento associado, permitindo que o modelo aprenda as relações entre os textos de entrada e as classes de sentimento correspondentes, como positivo e negativo. Além disso, ela é essencial para o reconhecimento de padrões nas palavras e frases que estão relacionadas a diferentes sentimentos.

Outro aspecto importante é a necessidade de utilizar um conjunto de dados de validação ou teste rotulado para avaliar a precisão do modelo. Ao dispor de um conjunto de dados rotulados separados, é possível verificar o desempenho do modelo em relação às classes de sentimento esperadas e determinar a sua precisão na análise de sentimentos.

Para o treinamento, utilizou-se dois conjuntos de dados distintos. O primeiro conjunto consistiu em 600 *tweets* rotulados em positivos e negativos do Cristiani et al. [9], ao realizar uma análise dos dados percebeu-se a presença de *tweets* duplicados removeu-se as suas cópias deixando apenas o *tweet* original no *dataset* totalizando 411 *tweets*. Na imagem 9 é possível observar a distribuição de sentimento dos dados rotulados.

O segundo conjunto consistiu de 400 *tweets* do conjunto de dados coletados pelo *crawler* como explica a Seção 4.1.1,

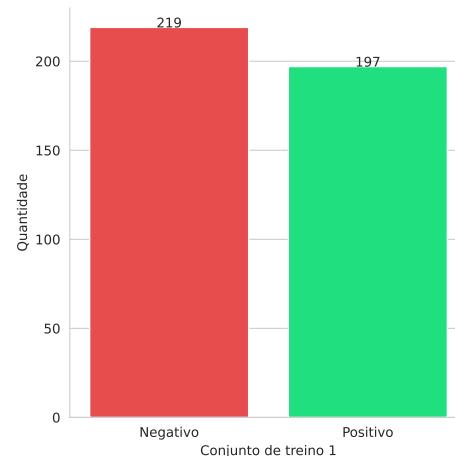


Figura 9: *Dataset* composto pelo Cristiani.

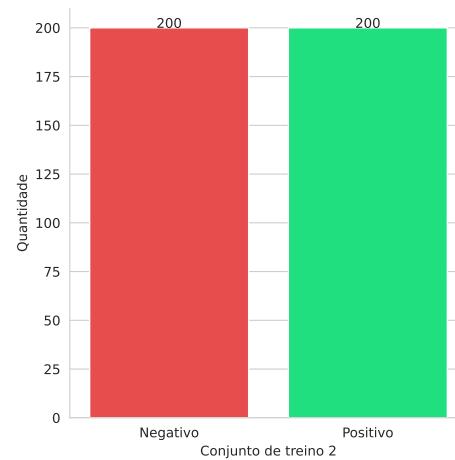


Figura 10: *Dataset* composto para o projeto.

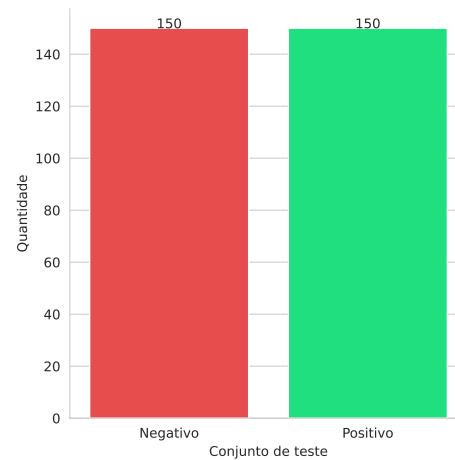


Figura 11: *Dataset* composto para o teste do modelo.

os quais foram classificados manualmente. Classificou-se 200 *tweets* com rótulo positivo e 200 com rótulo negativo. Na imagem 10 é possível observar a distribuição de sentimento dos dados rotulados.

Para o conjunto de teste, foram selecionados e classificados manualmente 300 *tweets* do conjunto de dados coletados pelo *crawler*, contendo 150 *tweets* de cada sentimento. Na imagem 11 é possível observar a distribuição de sentimento dos dados rotulados.

Para este trabalho, definiu-se que para um *tweet* ser considerado como positivo o usuário poderia:

| |
|---|
| Expressar apoio ou encorajamento direto ao candidato |
| “TE AMAMOS BOLSONARO 22” |
| Concordar com suas propostas, ideais e falas |
| “Quero meu país de volta!! Com o fim da FOME, com o brasileiro empregado e vivendo com amor e alegria!! Quero LULA de novo!!” |

Tabela VII: Exemplos de *tweets* positivos presentes no *dataset*.

Um *tweet* negativo por sua vez o usuário poderia expressar:

| |
|--|
| Discordância clara com alguma ideia apresentada |
| “O governo Bolsonaro é um governo de absurdos! E por isso precisamos colocar um fim nele!” |
| Ataques pessoais a uma pessoa ou grupo |
| “@LulaOficial Se o luladrão voltar, já sabemos... muito dinheiro pra turminha” |

Tabela VIII: Exemplos de *tweets* negativos presentes no *dataset*.

A reportagem do Correio Braziliense [41] informava que devido a polarização, as eleições de 2022 caminhavam para um desenho bipolar, com os dois líderes nas pesquisas muito à frente dos candidatos do centro. Dessa forma, descartou-se o sentimento neutro visto que as pesquisas demonstraram o quanto a eleição foi acirrada e que o sentimento era de apoio ou rejeição aos candidatos [42].

B. Classificação

Durante a etapa de classificação, por trabalhar com muitos dados, é necessário uma máquina que tenha capacidade de lidar com processamento intensivo. Neste experimento, fez-se uso da plataforma *Kaggle* quando necessário, e um computador com as seguintes características para os treinamentos:

| Processador | GPU | Memória RAM |
|---------------|-------------------------|-------------|
| Ryzen 7 5700G | NVIDIA GeForce RTX 3050 | 16GB |

Tabela IX: Dados da máquina utilizada para treino.

1) Processo: Como mencionado na Subseção I da Seção B, o desenvolvimento do *BERT* consistiu na criação de um método de representação de linguagem pré-treinado no qual, primeiramente, ele é treinado com uma base contendo grandes fontes de texto. Com os resultados adquiridos do treinamento pode-se aplicá-lo em outras tarefas de processamento de linguagem natural, como a análise de sentimentos, realizando

o *fine-tuning* com a adição de mais uma camada de saída [43]. Relembrando, o ajuste fino consiste em realizar pequenos ajustes no modelo pré-treinado adicionando novas camadas de classificação e de saída a ela, no caso deste experimento a classificação é feita em sentimento negativo e positivo. Dessa forma, com a modificação de parâmetros o modelo é treinado para refinar as camadas antigas e treinar do zero as camadas novas adicionadas [44].

O *BERT* é utilizado em diversas línguas e com o crescimento de pesquisas na área surge-se assim variações e adaptações. Para este experimento utilizou-se o *BERTimbau*, modelo treinado com dados em português. Os dados de entrada para a rede necessitam ser tokenizados, isto é, dividir o texto em um conjunto de unidades menores (*tokens*) [45]. Assim sendo, empregou-se o *BertTokenizer* da biblioteca *Transformers* [46] da empresa *Hugging Face*, para realizar a tokenização dos *tweets*.

Além das bibliotecas citadas anteriormente, utilizou-se a biblioteca *Pytorch* [47] que foi desenvolvida na linguagem *Python* para auxiliar em projetos de aprendizado de máquina. A estrutura do classificador é composta por uma camada do *BERT*, uma camada de *dropout* e uma camada linear de saída. A camada do *BERT* é o modelo *BERT* pré-treinado, que recebe como entrada os *tokens* gerados pelo tokenizador e realiza o trabalho de classificação propriamente dito. A camada de *dropout* busca evitar o *overfitting*, generalizando o aprendizado e evitando que o modelo se ajuste demais aos dados de treinamento. A camada linear é a camada que mapeia a saída gerada pelo *BERT* para as classes do experimento, 0 para sentimento negativo e 1 para sentimento positivo.

Com o conjunto de dados limitado e buscando obter estimativas mais precisas da capacidade preditiva do modelo aplicou-se a técnica de validação cruzada. Para tal, optou-se pelo *K-Fold Cross Validation*, que consiste em dividir o conjunto de dados em k grupos, chamados de *folds*. Dessa forma, o treinamento é realizado com K -1 *folds* e o *fold* que sobrou é utilizado para o teste [48]. Na Figura 12 pode-se visualizar a ilustração do processo.

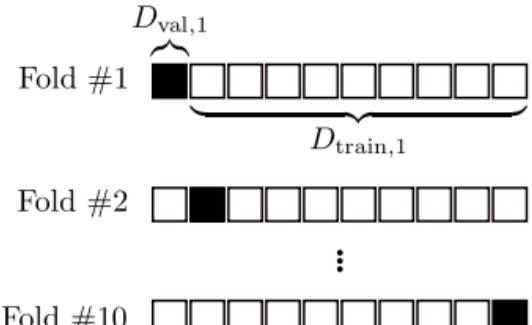


Figura 12: Exemplificação do *Cross Validation*.

Fonte: D. Berrar 2019 [49]

Foram realizados testes para cada conjunto de treino com diferentes tipos de pré-processamento, apresentados nas Seções 4.2.2 e 4.2.3. A princípio, utilizando K = 4, onde 75% dos dados são separados para treino e 25% dos dados são

separados para validação. Como também, $K = 10$, onde 90% dos dados são separados para treino e 10% dos dados são separados para validação. Em cada teste, o modelo com melhor acurácia e menor *loss* para o conjunto de validação foi salvo para análise posterior. Cada modelo salvo foi submetido a uma avaliação utilizando o conjunto de testes, onde foram consideradas as seguintes métricas: Acurácia, *F1-Score*, precisão e revocação. Para todos os casos, os modelos treinados com $K=10$ apresentaram resultados melhores.

A configuração dos parâmetros necessários para a classificação encontra-se na Tabela X. Ao analisar a distribuição de *tokens* no conjunto de dados, o tamanho das palavras para a entrada do *BERTTokenizer* foi definido em 200 ao analisar o tamanho dos textos no *dataset*. Selecionou-se um tamanho de lote (*batch size*) de 8, a taxa de aprendizagem em $2e-5$ [50] [51]. Como é um modelo pré-treinado, o *BERT* compreende o contexto e relaciona o sentimento bem rápido e geralmente obtém bons resultados em poucas épocas. Dessa forma, foi definido o total de 5 épocas para o refinamento após alguns testes.

| Batch size | Tamanho das palavras | Taxa de aprendizagem | Épocas |
|------------|----------------------|----------------------|--------|
| 8 | 200 | $2e-5$ | 5 |

Tabela X: Parâmetros para treinamento

Na Subseção 4.1.3, apresentou-se os *datasets* a serem utilizados neste trabalho. Durante o treinamento averiguou-se qual dos conjuntos obteve melhor desempenho. A Seção 4.2.2, a seguir, apresentará um panorama do cenário.

2) *Avaliação do modelo do experimento*: Para a avaliação dos modelos foram utilizadas as métricas de precisão (Eq. (1)), revocação (Eq. (2)), *F1-score* (Eq. (3)) e acurácia (Eq. (4)) conforme destacado por [52].

Precisão - Define-se como a razão entre a quantidade de positivos verdadeiros (PV) e o número total de previsões positivas realizadas pelo modelo, ou seja, positivos verdadeiros e positivos falsos (PF). Dessa forma, a precisão auxilia na confiança do modelo ao prever, por exemplo, se o *tweet* pertence a classificação de sentimento positivo e com qual porcentagem ele estará certo.

$$\text{Precisão} = \frac{PV}{PV + PF} \quad (1)$$

Revocação - Define-se como a razão entre a quantidade de verdadeiros positivos e o número total de casos positivos, ou seja, positivos verdadeiros e negativos falsos (NF). Dessa forma, a revocação auxilia o modelo a detectar, por exemplo, de todos os tweets que ele poderia classificar como sentimento positivo, quantos ele acertou.

$$\text{Revocação} = \frac{PV}{PV + NF} \quad (2)$$

F1-Score - Define-se como a média harmônica entre as duas métricas citadas acima, precisão e revocação. Quando ambas as métricas possuem relevância para o modelo, o *F1-Score* é utilizado para determinar a qualidade geral do modelo.

$$F1Score = 2 \cdot \frac{\text{precisão} \cdot \text{recall}}{\text{precisão} + \text{recall}} \quad (3)$$

Acurácia - Define-se como a razão entre o número de previsões corretas pelo número total de previsões. Dessa forma, a acurácia pode ser interpretada como a taxa de acerto do modelo.

$$\text{Acurácia} = \frac{PV + NV}{PV + NV + PF + NF} \quad (4)$$

Na Subseção A.3 informou-se dois conjuntos de dados utilizados para o treinamento e na Subseção A.2, informou-se os pré-processamentos realizados. Inicialmente, os treinos foram realizados com o *dataset* do Cristiani et al. [9], porém os resultados não estavam satisfatórios, como demonstram as matrizes de confusão na imagem 13.

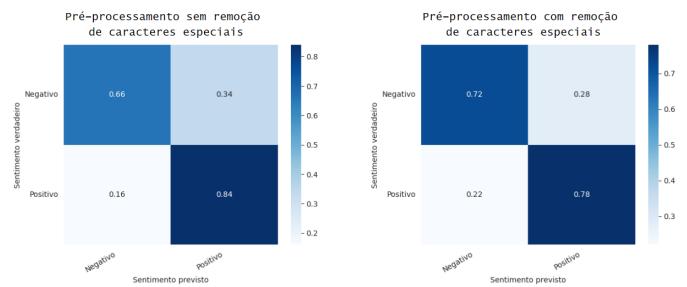


Figura 13: Matrizes de confusão do dataset Cristiani et al.

Dessa forma, iniciou-se um novo treinamento com os dados coletados pelo *crawler*, as Tabelas XI, XII e XIII demonstram as métricas obtidas.

| Sentimento | Precisão | Revocação | F1 score |
|------------|----------|-----------|----------|
| Negativo | 0,77 | 0,75 | 0,76 |
| Positivo | 0,76 | 0,77 | 0,76 |

Tabela XI: Métricas do *dataset* de teste com pré-processamento completo.

| Sentimento | Precisão | Revocação | F1 score |
|------------|----------|-----------|----------|
| Negativo | 0,83 | 0,87 | 0,85 |
| Positivo | 0,86 | 0,82 | 0,84 |

Tabela XII: Métricas do *dataset* de teste sem remoção de caracteres especiais.

| Sentimento | Precisão | Revocação | F1 score |
|------------|----------|-----------|----------|
| Negativo | 0,83 | 0,90 | 0,86 |
| Positivo | 0,89 | 0,81 | 0,85 |

Tabela XIII: Métricas do *dataset* de teste com remoção de caracteres especiais.

Observa-se que o melhor desempenho foi obtido ao utilizar o pré-processamento com remoção de caracteres especiais ao

obter para os rótulos negativo e positivo, precisão de 0,83 e 0,89, além de revocação igual a 0,90 e 0,81, respectivamente. Nas matrizes de confusão das imagens 14, 15, 16 é possível visualizar a percentagem de *tweets* classificados (durante a etapa de teste) corretamente como positivos e negativos, assim como os classificados incorretamente. Portanto, ao treinar o *BERT* com *tweets* com um contexto mais próximo dos dados classificados para o teste percebe-se a melhora do desempenho e os benefícios do modelo bidirecional.

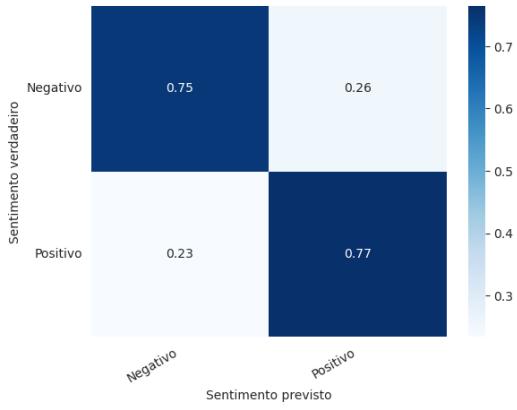


Figura 14: Matriz de confusão do conjunto de dados coletados com pré-processamento completo.

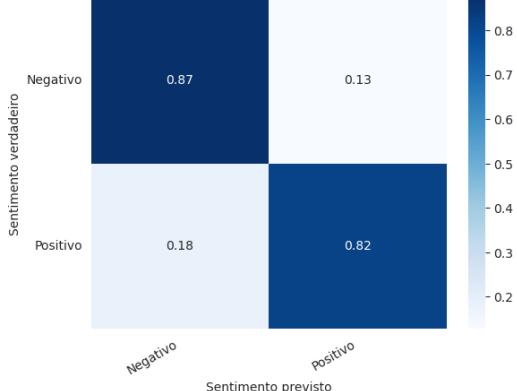


Figura 15: Matriz de confusão do conjunto de dados coletados com pré-processamento sem remoção

V. RESULTADOS

A. Classificação Automática

A Tabela XIV apresenta a classificação dos *tweets* separados por turno para os dois candidatos. Ao analisá-la, pode-se observar que há uma quantidade maior de *tweets* citando o candidato Lula. No entanto, a quantidade total de *tweets* para cada candidato não é um fator decisivo para o resultado final. Observando as Figuras 17 e 18, que apresentam o percentual de *tweets* para cada sentimento em cada turno, é possível notar que os sentimentos não estavam discrepantes para ambos os candidatos.

Apesar disso, o percentual de *tweets* negativos é maior do que o de positivos nos dois turnos para o candidato Bolsonaro,

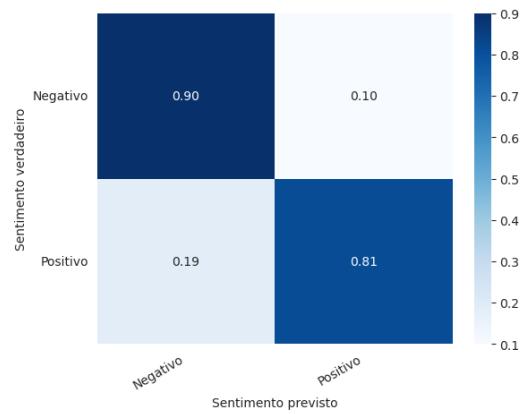


Figura 16: Matriz de confusão do conjunto de dados coletados com pré-processamento com remoção

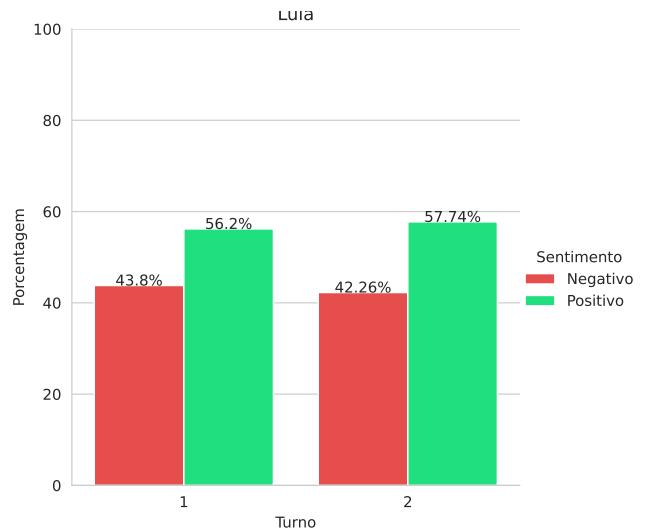


Figura 17: Resultados Lula - Melhor modelo *BERT* aplicando pré-processamento com remoção de caracteres especiais

enquanto o percentual de *tweets* negativos para o candidato Lula é menor. Essa análise é útil para prever o candidato com a maioria dos votos em cada turno, mas também revela a indecisão e o equilíbrio da polaridade entre os candidatos. A quantidade de *tweets* de cada sentimento não difere muito, variando menos de 10% em todos os casos. Além disso, a diferença percentual entre os *tweets* negativos e positivos para o candidato Lula variou pouco entre os dois turnos. Enquanto que a quantidade de *tweets* negativos cresceu para o candidato Bolsonaro.

Pelos resultados pode-se concluir que era um cenário acirrado, com os eleitores bastante divididos entre os candidatos analisados em ambos os turnos. Apesar disso, o candidato Lula obteve maior engajamento e mais apoiadores, enquanto para Bolsonaro, o percentual de apoiadores diminuiu no segundo turno, com *tweets* negativos se sobressaindo mais que positivos. Os resultados indicam um percentual maior de votos no primeiro turno e a vitória do candidato Lula em um cenário disputado.

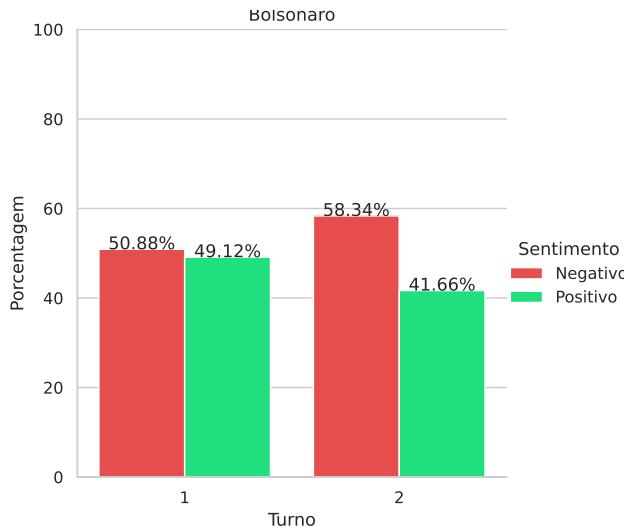


Figura 18: Resultados Bolsonaro - Melhor modelo *BERT* aplicando pré-processamento com remoção de caracteres especiais

| Lula | | | | Bolsonaro | | | |
|----------|----------|----------|----------|-----------|----------|----------|----------|
| 1º turno | | 2º turno | | 1º turno | | 2º turno | |
| Negativo | Positivo | Negativo | Positivo | Negativo | Positivo | Negativo | Positivo |
| 126.005 | 161.662 | 152.966 | 209.018 | 102.371 | 98.815 | 134.413 | 95.989 |

Tabela XIV: Resultados

B. Google Trends

Utilizando a ferramenta *Google Trends* observou-se a relevância das pesquisas relacionadas aos candidatos na internet com o resultado das eleições e comparou os resultados com o classificador. A Figura 19 apresenta uma comparação do interesse entre os candidatos obtida pela ferramenta. No gráfico, a busca por Lula é representada pela linha azul, enquanto as buscas por Bolsonaro são exibidas pela roxa, considerando o período de 05 de junho de 2022 a 25 de dezembro de 2022. O gráfico representa o interesse relativo ao ponto mais alto no gráfico, de modo que, o pico de pesquisas ocorreu para o candidato Lula no dia 30 de setembro, data do segundo turno das eleições. Além disso, é possível notar que houve um pico de interesse pelos candidatos na data do primeiro turno, no dia 02 de setembro.

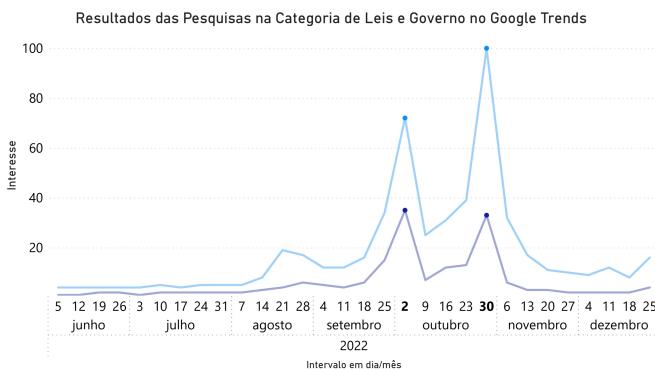


Figura 19: Gráfico de interesse de acordo com o Google Trends.

Durante todo o período da análise, as buscas por Lula superaram as buscas por Bolsonaro. Além disso, a análise de tendências de busca revelam que, em geral, as pesquisas por Lula tinham termos mais neutros e focadas em obter informações sobre os resultados da eleição e das pesquisas, enquanto as pesquisas por Bolsonaro continham também temas negativos com relação ao candidato. O gráfico não dá indícios do cenário concorrido perceptível nas pesquisas tradicionais e no classificador, apesar disso, pode-se concluir que o *Google Trends* pode ser uma ferramenta útil para a análise dos resultados das eleições. Os resultados demonstram maior interesse no candidato Lula e um sentimento negativo maior em relação à Bolsonaro, assim como os resultados encontrados pelo classificador.

C. Comparação com Institutos de Pesquisas

| Candidato | Lula | Bolsonaro | Demais candidatos |
|-------------------------|------|-----------|-------------------|
| Datafolha | 49,6 | 34,7 | 15,7 |
| Ipec | 51,1 | 35,6 | 13,3 |
| Resultados reais | 48,4 | 43,2 | 8,4 |

Tabela XV: Comparação pesquisas e resultados reais - 1º turno

| Candidato | Lula | Bolsonaro |
|-------------------------|------|-----------|
| Datafolha | 58,8 | 41,2 |
| Ipec | 59,0 | 41,0 |
| BERT | 53,4 | 46,6 |
| Resultados reais | 52,8 | 47,2 |

Tabela XVI: Comparação modelo, pesquisas e resultados reais considerando apenas Lula e Bolsonaro - 1º turno.

Foram analisadas pesquisas realizadas pelo Datafolha e pelo Ipec em cada turno, e os resultados foram comparados com o resultado do classificador para cada candidato.

Primeiro turno

Uma vez que, no primeiro turno constavam 11 candidatos à Presidência da República, realizou-se um ajuste nos percentuais das pesquisas e dos resultados encontrados para considerar apenas a porcentagem dos dois candidatos de interesse para este trabalho. Os resultados considerando os demais candidatos pode ser observado na Tabela XV. O resultado ajustado pode ser observado na Tabela XVI, juntamente com a comparação com o modelo do *BERT* e o resultado real.

A pesquisa do Datafolha [53] contém resultados entre 08 de julho de 2022 e 01 de outubro de 2022 e apresenta uma média de votos de 58,8% para Lula e 41,2% para Bolsonaro. A pesquisa do Ipec [54] apresenta resultados obtidos entre 15 de setembro de 2022 e 01 de outubro de 2022 e apresenta percentuais de intenção de voto de 60% para Lula e 40% para Bolsonaro. Dos *tweets* de apoio aos candidatos, o classificador obteve 53,4% de *tweets* em apoio a Lula e

46,6% dos *tweets* em apoio a Bolsonaro. Tanto as pesquisas quanto o classificador demonstram maior apoio ao Lula, em uma margem apertada de no máximo 10%. O classificador obteve uma diferença menor em relação aos 2 candidatos em comparação com as pesquisas tradicionais e se aproximou mais dos resultados reais.

Segundo turno

| Candidato | Lula | Bolsonaro |
|-------------------------|------|-----------|
| Datafolha | 52,6 | 47,4 |
| Ipec | 54,4 | 45,6 |
| BERT | 57,7 | 41,7 |
| Resultados reais | 50,9 | 49,1 |

Tabela XVII: Comparação modelo, pesquisas e resultados reais - 2º turno

A Tabela XVII apresenta os resultados das pesquisas, do classificador e o resultado real para o segundo turno das eleições. A pesquisa do Datafolha [55] contém resultados entre 07 de outubro de 2022 e 29 de outubro de 2022 e apresenta uma média de votos de 53% para Lula e 47% para Bolsonaro. A pesquisa do Ipec [56] apresenta resultados obtidos entre 05 de outubro de 2022 e 29 de outubro de 2022 e demonstra percentuais de intenção de voto de 54% para Lula e 46% para Bolsonaro.

Analisando os *tweets* positivos para cada candidato, como um sinal de apoio e possível voto, é possível comparar os resultados do classificador com os resultados apresentados na tabela. Dos *tweets* de apoio aos candidatos, o classificador obteve 57,7% de *tweets* em apoio a Lula e 41,7% dos tweets em apoio a Bolsonaro. A porcentagem de *tweets* positivos mostrou-se próxima dos resultados das pesquisas, resultando entre 4% e 5% de diferença entre os resultados para ambos os candidatos. Lula obteve uma média de 53% das intenções de voto durante o período das pesquisas, enquanto o classificador obteve 57.7% de *tweets* em apoio ao candidato. Bolsonaro obteve uma média de 47% das intenções de voto durante o período das pesquisas, enquanto o classificador identificou 41.7% de *tweets* favoráveis a ele. Os *tweets* apresentam uma diferença maior dos resultados reais do segundo turno em relação às pesquisas, entretanto, o percentual de *tweets* favoráveis para cada candidato se manteve próximo dos resultados das pesquisas e conseguiu prever de maneira correta o candidato que se sobressaiu nas eleições.

D. Comparação com trabalho de 2018

Rios 2018 [32] tinha o objetivo de predizer o resultado do primeiro turno das eleições presidenciais de 2018. Com esse propósito, ele utilizou classificadores clássicos (SVM, Regressão logística, KNN) e uma rede neural, em que os melhores resultados foram obtidos com a rede neural e a regressão logística, ambas com aproximadamente 84% de acurácia. Rios decidiu utilizar a regressão logística para classificação dos

tweets. O cenário político em 2018 era menos polarizado que as eleições de 2022 e Jair Bolsonaro liderava durante todo o ano, segundo as pesquisas. Rios apresenta os resultados em números absolutos e porcentagem de *tweets* positivos para cada um dos 8 principais candidatos.

A análise decisiva para o trabalho de Rios foi a análise da porcentagem de *tweets* positivos, que apresentou diferenças com relação à análise do montante de *tweets* positivos e se aproximou do resultado real e das pesquisas da época. O trabalho de Rios 2018 [32] previu corretamente os 3 primeiros candidatos, bem como a liderança de Jair Bolsonaro, todos os 3 candidatos possuíam alta porcentagem de apoio, com valores bem próximos em termos de porcentagem de *tweets* positivos. Na prática, Bolsonaro saiu na liderança com 46% dos votos, seguido por Fernando Haddad com 29% dos votos e Ciro Gomes com 12% dos votos.

Um ponto a se analisar é a quantidade de *tweets* para cada candidato. Em Rios, a maior parte dos *tweets* positivos fazia referência a Bolsonaro, demonstrando maior engajamento com o candidato. Para este trabalho, a quantidade de *tweets* fazendo referência à Lula era maior. Entretanto, no cenário político analisado, isso não é suficiente para uma conclusão como citado na Seção V-A. A principal diferença possível de se notar nos resultados é a porcentagem de *tweets* positivos para os candidatos. No cenário de 2018, os principais candidatos têm mais de 70% de *tweets* positivos, enquanto nos resultados deste trabalho, a porcentagem de positivos e negativos para os dois candidatos analisados é bem equilibrada nos dois turnos, inclusive com a quantidade de *tweets* negativos para Bolsonaro sendo maior que a de *tweets* positivos, demonstrando opiniões bastante divididas para os principais candidatos.

Este trabalho, lidou com um cenário bem polarizado, onde Lula e Jair Bolsonaro lideravam as pesquisas no primeiro turno e com resultados historicamente próximos. Foi possível concluir corretamente a vitória de Lula, com resultados relativamente próximos das pesquisas tradicionais, além de analisar sentimentos gerais para os candidatos com base nos sentimentos dos *tweets*.

VI. CONCLUSÃO

Neste artigo, apresentou-se uma abordagem para pré-processamento e classificação de *tweets* com o objetivo de analisar os sentimentos e obter previsões dos dois turnos das eleições presidenciais brasileiras de 2022, assim, observando e comparando com os resultados reais, do *Google Trends* e com o trabalho realizado nas eleições de 2018. Por fim, validando se dados provenientes da internet são boas fontes que podem ser consultadas para realizar pesquisas eleitorais.

O estudo permitiu observar o cenário político polarizado de 2022 nos dados coletados e classificados. Além disso, o modelo aproximou-se dos resultados reais das eleições no primeiro turno e conseguiu prever o candidato eleito no segundo. Demonstrando assim, que os *tweets* são bons indicadores para o resultado das eleições mesmo em um cenário disputado. Ao comparar o experimento realizado com o do Rios 2018 [32], percebeu-se o quanto a situação das eleições podem variar após quatro anos de governo. Enquanto, no trabalho

de 2022 dois candidatos se destacavam, no de 2018 um candidato estava disparado em relação aos demais. Nesse sentido, observa-se que o sentimento em relação aos candidatos mudam com o decorrer do tempo e gerências realizadas, tornando-se interessante e necessário prosseguir com pesquisas de análise de sentimento nas eleições presidenciais brasileiras.

Com relação às informações disponibilizadas pelo *Google Trends*, nem todas constam sentimentos direcionados aos candidatos, mas sim o interesse de buscar informações sobre, por exemplo, horário da posse, julgamentos, entrevistas e outros. Porém, ainda foi possível observar um maior interesse no candidato Lula e pesquisas negativas em relação ao candidato Bolsonaro, portanto, é uma boa fonte para obter análises em períodos curtos e próximos às datas das eleições, de acordo com a categoria a ser analisada na ferramenta.

Para trabalhos futuros, recomenda-se a criação de uma quantidade maior de dados rotulados para o treinamento, no intuito de gerar resultados mais precisos por proporcionar um contexto de aprendizado maior ao modelo. Os *tweets* contém uma linguagem mais informal, então um pré-processamento que trabalhasse com esse tratamento de gírias e abreviações poderia ser analisado e aplicado em busca de melhorar o desempenho do modelo. Por fim, se o cenário político permitir, a análise sobre sentimentos neutros observando a influência e a importância para a previsão das eleições.

AGRADECIMENTOS

Agradecemos a nossa professora e orientadora Deborah Fernandes pela confiança e a oportunidade de realizar este trabalho, assim como, o corpo docente da Universidade Federal de Goiás por nos guiar no decorrer do curso.

Eu, Giulia Borges, sou grata aos meus ilustres pais, Lola e Domingos, pela vida, educação, amor, ensinamentos, paciência e apoio durante todos esses anos de vida. A minha amada irmã, Isabela, por me fazer rir nas horas tristes, por me ensinar cada dia mais sobre o mundo, por me amar e *"And if we've only got this life in this adventure, oh, then I want to share it with you"* - Coldplay. As minhas duas amigas de alma Izabelli e Ana Luiza, por escutarem minhas reclamações durante todo esse processo, por me ajudarem com conselhos de vida, por criarem novas expectativas de vida e por me amarem durante todos esses anos de amizade. A minha família, aos meus queridos amigos da faculdade e da vida por me auxiliarem, apoiarem e festejarem esta fase final de um capítulo da minha vida. Agradeço também, ao meu nobre amigo Moacir por ter percorrido e trabalhado comigo durante o nosso período na universidade e no projeto final de curso. Agradeço a Deus pela oportunidade de viver, aprender e compartilhar.

REFERÊNCIAS BIBLIOGRÁFICAS

- %20Introduction%20%28Jurafsky%20e%20Martin%2C%202008%29.pdf.
- [18] T. Winograd, *SHRDLU*. endereço: <https://hci.stanford.edu/~winograd/shrdlu/>.
- [19] A. Kumar, *Complete Natural Language Processing Guide For Beginners In 2021 - Buggy Programmer*, 2021. endereço: <https://buggyprogrammer.com/what-is-natural-language-processing/>.
- [20] D. Khurana, A. Koli, K. Khatter e S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, v. 82, jul. de 2022. DOI: <https://doi.org/10.1007/s11042-022-13428-4>.
- [21] B. Felipe e E. Lins, "Claudionor Rocha* Consultor Legislativo da Área de Segurança Pública e Defesa Nacional 11 A evolução da Internet: uma perspectiva histórica," 2013. endereço: http://www.belins.eng.br/ac01/papers/aslegis48_art01_hist_internet.pdf.
- [22] B. Liu, *Sentiment Analysis and Opinion Mining*. 2012. endereço: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>.
- [23] IBM, *What is Machine Learning?* 2019. endereço: <https://www.ibm.com/topics/machine-learning>.
- [24] Endereço: <https://sites.icmc.usp.br/andre/research/neural/>.
- [25] 2020. endereço: <https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-deep-learning/>.
- [26] I. Machado e L. Driemeier, *PMR5251 - Avaliação do Comportamento Mecânico de Materiais Utilizando uma Abordagem de ML MACHINE LEARNING: PROBLEMAS DE CLASSIFICAÇÃO*. 2020. endereço: https://edisciplinas.usp.br/pluginfile.php/5809148/mod_resource/content/3/Aula04_Classification.pdf.
- [27] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, v. 55, n. 10, p. 78, out. de 2012. DOI: <https://doi.org/10.1145/2347736.2347755>.
- [28] É. Fontana, "Introdução aos Algoritmos de Aprendizagem Supervisionada," 2020. endereço: https://fontana.pginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf.
- [29] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," 2017.
- [30] J. Devlin, M.-W. Chang, K. Lee e K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," mai. de 2019. endereço: <https://arxiv.org/pdf/1810.04805.pdf>.
- [31] H. Face, *BERT*. endereço: https://huggingface.co/docs/transformers/model_doc/bert.
- [32] G. Rios, "PREDIÇÃO DO RESULTADO DAS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS UTILIZANDO ANÁLISE DE SENTIMENTOS EM TWETS," 2018, pp. 1–57.
- [33] T. Brasil, #Eleições2022: veja dados sobre essas conversas no Twitter, set. de 2022. endereço: https://blog.twitter.com/pt_br/topics/company/2022/-eleicoes2022--veja-dados-sobre-essas-conversas-no-twitter.
- [34] A. Tumasjan, T. Sprenger, P. Sandner e I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," vol. 10, jan. de 2010.
- [35] W. Budiharto e M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *Journal of Big Data*, v. 5, 1 dez. de 2018, ISSN: 21961115. DOI: [10.1186/s40537-018-0164-1](https://doi.org/10.1186/s40537-018-0164-1).
- [36] Twitter, *Twitter API Documentation*, 2023. endereço: <https://developer.twitter.com/en/docs/twitter-api>.
- [37] S. Premevida, *Guia de NLP - conceitos e técnicas*, nov. de 2021. endereço: <https://www.alura.com.br/artigos/guia-nlp-conceitos-tecnicas>.
- [38] P. Pincer, *Número de eleitores brasileiros no exterior bate recorde em 2022*, set. de 2022. endereço: <https://www12.senado.leg.br/radio/1/noticia/2022/09/30/numero-de-eleitores-brasileiros-no-exterior-bate-recorde-em-2022>.
- [39] V. Lendave, *A Guide to Text Preprocessing Using BERT*, set. de 2021. endereço: <https://analyticsindiamag.com/a-guide-to-text-preprocessing-using-bert/>.
- [40] IBM, *O que é rotulagem de dados? — IBM*. endereço: <https://www.ibm.com/br-pt/topics/data-labeling>.
- [41] R. Felice, *Eleições: Polarização entre Lula e Bolsonaro é cada vez mais provável*, abr. de 2022. endereço: <https://www.correiobrasiliense.com.br/politica/2022/04/4999677-polarizacao-cada-vez-mais-provavel.html>.
- [42] C. Cerqueira e D. Moliterno, *Disputa entre Lula e Bolsonaro é a eleição para presidente mais acirrada da história*, out. de 2022. endereço: <https://www.cnnbrasil.com.br/politica/disputa-entre-lula-e-bolsonaro-e-a-eleicao-para-presidente-mais-acirrada-da-historia/>.
- [43] G. Cloud, *Primeiros passos com o algoritmo BERT integrado — AI Platform Training*, fev. de 2023. endereço: <https://cloud.google.com/ai-platform/training/docs/algorithms/bert-start?hl=pt-br>.
- [44] A. von Wangenheim, *Deep Learning::Aprendizado por Transferência e Ajuste Fino*, 2018. endereço: <https://lapix.ufsc.br/ensino/visao/visao-computacionaldeep-learning/deep-learningaprendizado-por-transferencia-e-ajuste-fino/>.
- [45] A. Ly, B. Uthayasooriyar e T. Wang, *A SURVEY ON NATURAL LANGUAGE PROCESSING (NLP) & APPLICATIONS IN INSURANCE PREPRINT*. 2020. endereço: <https://arxiv.org/pdf/2010.00462.pdf>.
- [46] H. Face, *Transformers*. endereço: <https://huggingface.co/docs/transformers/index>.
- [47] PyTorch, *PyTorch*. endereço: <https://pytorch.org/>.
- [48] SciKit-Learn, *3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation*, 2009. endereço: https://scikit-learn.org/stable/modules/cross_validation.html.
- [49] D. Berrar, "Cross-validation," 2019. endereço: https://www.researchgate.net/profile/Daniel-Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf.

- [50] A. Chiorrini, C. Diamantini, A. Mircoli e D. Potena, “Emotion and sentiment analysis of tweets using BERT,” em *EDBT/ICDT Workshops*, vol. 3, 2021.
- [51] H. Zanini, *Análise de sentimentos em português utilizando Pytorch e Python*, nov. de 2021. endereço: <https://medium.com/data-hackers/an%C3%A1lise-de-sentimentos-em-portugu%C3%A9s-utilizando-pytorch-e-python-91a232165ec0>.
- [52] F. Benevenuto, F. Ribeiro e M. Araújo, *Capítulo 1 Métodos para Análise de Sentimentos em mídias sociais*. endereço: <https://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>.
- [53] G1, *Datafolha, votos válidos: Lula 50%; Bolsonaro 36%*, out. de 2022. endereço: <https://g1.globo.com/politica/eleicoes/2022/pesquisa-eleitoral/noticia/2022/10/01/datafolha-votos-validos-lula-50percent-bolsonaro-36percent.ghtml>.
- [54] G1, *Ipec, votos válidos: Lula, 51%; Bolsonaro, 37%*, out. de 2022. endereço: <https://g1.globo.com/politica/eleicoes/2022/pesquisa-eleitoral/noticia/2022/10/01/ipec-votos-validos-lula-51percent-bolsonaro-37percent.ghtml>.
- [55] G1, *Datafolha: Lula tem 52% dos votos válidos no 2º turno, e Bolsonaro, 48%*, out. de 2022. endereço: <https://g1.globo.com/politica/eleicoes/2022/pesquisa-eleitoral/noticia/2022/10/29/datafolha-lula-tem-52percent-dos-votos-validos-no-2o-turno-e-bolsonaro-48percent.ghtml>.
- [56] G1, *Ipec: Lula tem 54% dos votos válidos no 2º turno, e Bolsonaro, 46%*, out. de 2022. endereço: <https://g1.globo.com/politica/eleicoes/2022/pesquisa-eleitoral/noticia/2022/10/29/ipec-lula-tem-54percent-dos-votos-validos-2o-turno-e-bolsonaro-46percent.ghtml>.