Análise de polaridade e de tópicos em *tweets* no domínio da política no Brasil

Leonardo Capellaro[1], **Helena de Medeiros Caseli**[1]

[1]Universidade Federal de São Carlos (UFSCar)
Departamento de Computação – LALIC
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

leonardo.capellaro@gmail.com, helenacaseli@ufscar.br

Abstract. *The field of politics in Brazil is one of the busiest and most controversial in the last decade. With the advent of social networks, a new communication channel between voters and politicians was created, with users having a space to publish their opinions and beliefs. In this context, this work shows the performance of BERT models in the polarity (positive, negative or neutral) and topic analysis of tweets related to Brazilian politics. Experiments were caried out with tweets related to the 2018 elections achieving results ($F1 = 96\%$) better than the ones obtained by previous works for the polarity classification. The qualitative analysis of the topics was also promising.*

Resumo. *O campo da política no Brasil é um dos mais movimentados e mais polêmicos da última década. Com o advento das redes sociais, um novo canal de comunicação entre os eleitores e os políticos foi criado, com os usuários tendo um espaço para publicar suas opiniões e crenças. Neste contexto, este trabalho mostra como modelos BERT se saem na análise de polaridade (positiva, negativa ou neutra) e de tópicos de grandes quantidades de tweets relacionados à política do Brasil. Experimentos foram realizados com tweets relacionados às eleições de 2018, e melhores resultados ($F1 = 96\%$) foram obtidos para a classificação de polaridade em comparação com trabalhos anteriores. A avaliação qualitativa dos tópicos também mostrou resultados promissores.*

1. Introdução

Um estudo realizado pelo IBGE[1], em 2019, demonstrou que 82,7% dos domicílios brasileiros possuía conexão com a internet. Outro estudo realizado no mesmo ano pela Comscore[2] apontou que aproximadamente 88% da população brasileira tinha acesso a algum tipo de rede social, com o Twitter possuindo 40,2 milhões de usuários ativos no país. Por ser uma rede social que limita o número de caracteres nos textos postados na rede, o Twitter se tornou um canal de comunicação rápida e simplificada, sendo uma boa ferramenta para saber o que as pessoas estão comentando espontaneamente sobre os assuntos do momento [Sales e Barbosa 2019].

[1]https://educa.ibge.gov.br/jovens/materias-especiais/20787-uso-de-internet-televisao-e-celular-no-brasil.html

[2]https://olhardigital.com.br/2019/07/05/noticias/brasil-e-o-pais-que-mais-usa-redes-sociais-na-america-latina/

No domínio da política, as redes sociais também passaram a ser uma importante vitrine para os candidatos divulgarem seus trabalhos, suas propostas e opiniões e atingirem mais rapidamente seus eleitores. De acordo com um estudo divulgado pelo próprio Twitter em 2018[3], no período de 16 de agosto a 28 de outubro de 2018, foram contabilizados 165 milhões de *tweets* relacionados às eleições, um volume quatro vezes maior que o total das eleições anteriores, em 2014.

Saber "o que sentem" e "sobre o que falam" os usuários do Twitter, no domínio da política brasileira, é a grande motivação deste trabalho. Para tanto, este trabalho apresenta resultados da investigação de análise de polaridade e de tópicos em *tweets*, escritos em português, no domínio da política brasileira coletados em 2018. A **análise de polaridade** consiste em determinar qual a polaridade (ou valência) da opinião do autor de um texto com relação à entidade ou assunto em discussão. Desse modo é possível, por exemplo, verificar se a opinião do autor é positiva, negativa ou neutra em relação à entidade ou assunto alvo do texto. A **análise de tópicos**, por sua vez, visa lidar com grandes quantidades de textos, análises e *feedbacks*, de forma a retornar os principais assuntos/tópicos dos textos de acordo com sua importância e recorrência.

Alguns trabalhos anteriores foram realizados analisando sentimentos em *tweets* do campo da política, como [Moreira et al. 2020] e [Cristiani et al. 2020]. Em [Moreira et al. 2020], foi proposta uma análise da polarização da elite em comparação com a massa no procedimento de impeachment da presidente do Brasil em 2016. Em [Cristiani et al. 2020], a análise de sentimentos é aplicada a *tweets* das eleições presidenciais de 2018 no Brasil, onde foi estudada a relação entre as opiniões dos usuários que publicaram sobre os candidatos durante o período das eleições e seu resultado final. Entretanto, devido à ausência de um modelo de linguagem contextualizado (considerado o estado-da-arte para diversas aplicações de PLN) treinado/disponível para o português do Brasil naquela época, como o BERTimbau [Souza et al. 2020], outros métodos foram utilizados, como o SVM e Naive Bayes.

Assim, as principais contribuições deste trabalho estão relacionadas à análise de desempenho de modelos BERT no *corpus* de [Cristiani et al. 2020] para: (i) a análise de polaridade usando o BERTimbau [Souza et al. 2020], dimensionando o avanço em performance com o uso desta nova técnica; e (ii) a análise de tópicos utilizando o BERTopic [Grootendorst 2020], algo ainda inédito neste *corpus* e domínio.

2. Trabalhos relacionados

Esta seção traz uma visão geral de alguns trabalhos realizados para análise de sentimentos e de tópicos em *tweets*, escritos em português do Brasil, no domínio da política.

2.1. Análise de sentimentos

Podemos encontrar diversos trabalhos que realizaram a análise de sentimentos em *tweets* na língua portuguesa. Em [Christhie et al. 2018], os autores utilizaram Naive Bayes, SVM e Random Forest para realizar uma análise de posicionamento em *tweets* de diversos candidatos às eleições de 2018, classificando-os em: contra, a favor ou neutro. O algoritmo SVM conseguiu uma $F1$ de 99% em um dos conjuntos utilizados na análise.

[3] https://blog.twitter.com/pt_br/topics/company/2018/como-foram-as-eleicoes-2018-no-twitter

Em [Pereira 2019], o autor realizou uma análise dos sentimentos dos *tweets* para cada um dos candidatos das eleições de 2018, separados por evento, comparando o sentimento em todos os debates realizados no período e analisando a popularidade de cada candidato com base nos sentimentos positivos dos textos da rede social. Foram utilizados os métodos SVM e Naive Bayes, onde o método SVM atingiu 86% de acurácia no melhor caso, enquanto o Naive Bayes ficou em 85%.

Já em [Cristiani et al. 2020], os autores utilizaram Naive Bayes e SVM em um *corpus* com aproximadamente 369 mil *tweets* sobre as eleições presidenciais de 2018. O trabalho buscava relacionar os sentimentos identificados nos *tweets* a respeito de cada candidato com o resultado final das eleições, e os valores de $F1$ para os métodos Naive Bayes e SVM ficaram em $54,2\%$ e $66,1\%$ respectivamente.

No presente trabalho, a análise de polaridade dos *tweets* do *corpus* de [Cristiani et al. 2020] foi realizada com o *fine-tuning* de um modelo BERT pré-treinado em português do Brasil, o BERTimbau [Souza et al. 2020], lançando-se mão de uma série de técnicas de pré-processamento com o intuito de obter a melhor $F1$ para o modelo. O modelo foi treinado com 600 *tweets* anotados manualmente, e posteriormente aplicado em um *corpus* contendo aproximadamente 370 mil *tweets* para classificá-los.

2.2. Análise de tópicos

Diversas abordagens para a análise de tópicos foram empregadas em trabalhos relacionados. Em [Pinto et al. 2020], os autores utilizaram modelagem de tópicos e análise de sentimentos em *tweets* da língua inglesa durante a pandemia da COVID-19. Os pesquisadores utilizaram três algoritmos de aprendizagem de máquina: o *Latent Dirichlet Allocation* (LDA), o *Non-Negative Matrix Factorization* (NMF) e o *Latent Semantic Analysis* (LSA). Foi realizada uma análise qualitativa dos resultados, onde o NMF foi o algoritmo que gerou a melhor relação entre os termos e a maior coerência com o tema da pesquisa.

O BERTopic [Grootendorst 2020] foi utilizado em [Silveira et al. 2021] para fazer uma análise de tópicos em documentos do meio jurídico. Na avaliação qualitativa, realizada por especialistas, constatou-se que $84,6\%$ dos tópicos gerados pelo modelo correspondiam aos temas principais dos documentos.

No presente trabalho, foi utilizado o BERTopic para realizar a modelagem de tópicos com o intuito de obter os principais assuntos mencionados nos *tweets* previamente anotados automaticamente como positivos e negativos no domínio da política.

3. Metodologia

Esta seção descreve os métodos empregados para análise de polaridade e de tópicos em *tweets* do domínio da política, escritos em português do Brasil.

3.1. Análise de polaridade com BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) é um algoritmo que emprega treinamentos bidirecionais de *Transformers*, e foi proposto em [Devlin et al. 2019]. O algoritmo possui como característica fundamental o fato de ser bidirecional, o que permite uma maior compreensão do contexto de uma palavra e do texto como um todo, analisando as palavras adjacentes em ambas as direções.

Na maioria das aplicações práticas são utilizados modelos pré-treinados do BERT, que podem ser treinados em uma única língua ou em diversas línguas diferentes. O modelo pré-treinado passa, então, por uma etapa de ajuste ou sintonia fina (*fine-tuning*), que modela a última camada da rede neural BERT para a especificidade do problema em questão, como ilustrado na Figura 1. Desta forma, o modelo pré-treinado é criado utilizando grandes *corpora* e máquinas com grandes capacidades de processamento, deixando para o usuário apenas a necessidade de realizar o *fine-tuning* para um problema específico.

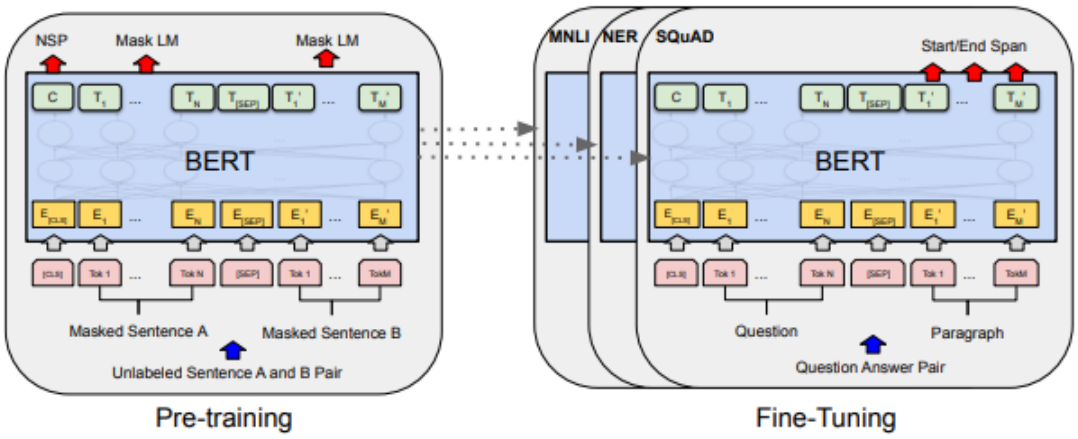

Figura 1. Processo de pré-treinamento e *fine-tuning* do BERT, de acordo com [Devlin et al. 2019]

A análise de polaridade de textos é tratada como uma tarefa de classificação, onde cada texto possui um rótulo indicando o tipo de sentimento empregado na escrita. Nesse tipo de tarefa, o BERT funciona adicionando uma camada de classificação na parte superior da saída do *Transformer*. A maioria dos hiperparâmetros da rede neural se mantém os mesmos do modelo pré-treinado previamente, com exceção do tamanho de lote (*batch*), da taxa de aprendizagem (*learning rate*) e do número de períodos de treinamento (*epochs*).

Para o presente trabalho, foi utilizado o BERTimbau [Souza et al. 2020], modelo pré-treinado para o português do Brasil e disponibilizado pela neuralmind[4]. Para o *fine-tuning* do modelo, foi utilizado o *corpus* disponibilizado por [Cristiani et al. 2020][5].

3.2. Análise de tópicos usando BERT e TF-IDF

Para a análise de tópicos, foi utilizado o algoritmo BERTopic[6] [Grootendorst 2020]. Este algoritmo utiliza as *embeddings* de palavras presentes em um modelo BERT para gerar agrupamentos hierárquicos por densidade de palavras utilizando o HDBSCAN e, por fim, gera tópicos com base na importância das palavras utilizando uma variação do TF-IDF, chamada de c-TF-IDF.

A primeira etapa do BERTopic converte os documentos em dados numéricos com base nas *embeddings* do modelo BERT. A segunda etapa é a de clusterização, que é realizada através do HDBSCAN, um algoritmo de agrupamento hierárquico por densidade

[4]https://github.com/neuralmind-ai/portuguese-bert
[5]https://github.com/andrecristiani/analise-de-sentimentos-eleicoes-2018
[6]https://maartengr.github.io/BERTopic/index.html

proposto em [Campello et al. 2013]. Neste algoritmo, os documentos que possuem maior similaridade entre si são agrupados em *clusters* baseados na estabilidade do *cluster*. Uma das importantes características do HDBSCAN é o fato de ele não forçar a seleção de um dado para um determinado *cluster*. Caso o dado não se encaixe em nenhum grupo por similaridade, ele é considerado um *outlier*.

A última etapa do processo é a seleção dos tópicos com base na importância das palavras. Para isso, o autor desenvolveu uma técnica nomeada de *c-TF-IDF*. Esta técnica funciona de forma muito parecida com o *TF-IDF* original, que compara a importância das palavras analisando todo o *corpus*, entretanto, o *c-TF-IDF* utiliza os *clusters* gerados na etapa de agrupamento aplicando o *TF-IDF* em cada um deles. Esse processo classifica as palavras de acordo com a importância para cada grupo gerado no processo de agrupamento, gerando os principais tópicos de cada grupo.

4. Experimentos

Esta seção descreve as configurações dos experimentos, com a descrição do *corpus* e suas etapas de pré-processamento, e as questões de pesquisa.

4.1. *Corpus*

O *corpus* de treinamento utilizado neste trabalho é o mesmo de [Cristiani et al. 2020]. Esse *corpus* está composto por 600 *tweets* anotadas manualmente com suas respectivas polaridades: positivo, quando a mensagem publicada demonstra apoio ao candidato; negativo, quando a mensagem publicada demonstra rejeição ao candidato; e neutro, quando a mensagem publicada não demonstra opinião de polaridade clara sobre o candidato.

Em [Cristiani et al. 2020], todas as letras dos textos do *corpus* foram transformadas em minúsculas, todos os *retweets* dos usuários foram removidos e qualquer tipo de *hiperlink* incluído na mensagem foi apagado. No presente trabalho, utilizou-se a ferramenta Enelvo[7] [Bertaglia e Nunes 2016], que visa a padronização (normalização) dos textos obtidos da Internet. Em testes preliminares, a configuração de pré-processamento que apresentou melhores resultados foi a que denominamos "Enelvo Raw sem emoji e retweet", onde os textos foram normalizados pelo Enelvo com o parâmetro *tokenizer* configurado como *readable*, mantendo as entidades dos textos inalteradas[8] e, em seguida, foram removidos emojis e *retweets*. Assim, nos experimentos apresentados neste artigo, dois pré-processamentos foram comparados:

- Original: pré-processamento utilizado em [Cristiani et al. 2020], onde são removidos *hiperlinks*, *retweets* e o texto é convertido para minúsculas.
- Enelvo Raw sem emoji e retweet: normalização feita no modo Enelvo Raw, mas removendo emojis e *retweets*.

A Tabela 1 traz exemplos de *tweets* originais do *corpus* de [Cristiani et al. 2020].

Além do *corpus* com 600 *tweets* anotados com polaridade, neste trabalho também foi utilizado o *corpus* de 369 mil *tweets* de [Cristiani et al. 2020], sem anotação de polaridade, que foram obtidos durante o segundo turno das eleições. Esse *corpus* foi usado para

[7] `https://thalesbertaglia.com/enelvo/`

[8] Por padrão, o Enelvo troca algumas entidades do texto por *tags*, como o nome do usuário por "username", *hashtags* por "hashtag" e números por "number".

Tabela 1. Exemplos de *tweets* anotados manualmente com polaridade

Texto do *tweet*	Classe
a justiça deveria ser sempre ágil, como é para tudo relacionado ao lula. só acho.	neutro
fui de #ciro no primeiro turno. parte não é minha preferência. mas agora sou #haddadpresidente desde criancinha. #eleições2018	positivo
realmente bolsonaro não tem raciocínio lógico. não conseguiu sair da primeira pergunta. #bolsonaronojornalnacional	negativo

Fonte: *Tweets* extraídos do *coprus* de [Cristiani et al. 2020]

a análise de tópicos relativos aos dois candidatos do segundo turno. Devido ao grande tamanho do *corpus*, não foi possível processá-lo com o Enelvo.[9] Assim, um processamento mais simples foi realizado para: (i) remover *hashtags* e links através de expressões regulares; (ii) converter para minúsculas; (iii) remover os inícios de parágrafos representados por \n; (iv) remover vogais repetidas nas palavras (por exemplo, 'corruptooo' foi transformado em 'corrupto'); (v) remover as pontuações do texto (vírgula, interrogação, exclamação, pontos finais, dentre outras); (vi) remover as *stopwords* (como "o", "a", "com", "para", entre outras); e (vii) remover as palavras com apenas uma letra.

4.2. Configurações dos experimentos

Para a classificação de polaridade dos *tweets*, os algoritmos de classificação selecionados foram: BERT, Naive Bayes e SVM, sendo os dois últimos utilizados em [Cristiani et al. 2020]. Os *corpora* utilizados foram o original de [Cristiani et al. 2020] e o pré-processado Enelvo Raw sem emoji e *retweet*.

No treinamento, o *corpus* foi dividido em: 90% para treinamento e 10% para teste, utilizando a mesma partição dos dados para todos os algoritmos investigados. O BERT foi refinado por 15 épocas, sendo considerado o melhor valor de $F1$ entre as épocas para a comparação com os demais algoritmos. A taxa de aprendizagem para o BERT foi fixada em 2e-5 e foi utilizado o otimizador AdamW. Buscando replicar os mesmos parâmetros utilizados por [Cristiani et al. 2020], para a classificação com o SVM foi utilizado o kernel linear, parâmetro de regularização C igual a 1, grau da função polinomial do kernel em 3 e valor de gamma em "auto". Para o Naive Bayes foram utilizados os parâmetros *default* do modelo multinomial do `scikit-learn`[10].

Na análise de tópicos, utilizou-se o BERTopic com o modelo BERTimbau limitando o número de tópicos a 20, quantidade que foi definida empiricamente analisando os resultados gerados pelo modelo completo, que mostrou ter aproximadamente este número de tópicos. Por ser um método de classificação não supervisionado, não precisa de nenhum treinamento e utiliza apenas dados não rotulados.

5. Resultados

Os experimentos foram divididos em três etapas para responder as seguintes questões de pesquisa:

Q1 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. 2020]?

[9] O Enelvo ficou em execução durante 4 dias e não concluiu o processamento do *corpus*.

[10] `https://scikit-learn.org/`

Q2 O pré-processamento refinado, com o auxílio da ferramenta Enelvo, traz ganho de desempenho de BERT, Naive Bayes e SVM na análise de polaridade?

Q3 O BERTopic pode ser considerado uma boa ferramenta para a extração de tópicos, em *tweets* classificados como positivos e negativos referentes a cada candidato, com base em uma análise qualitativa dos dados?

5.1. Análise de polaridade

A Tabela 2 resume os resultados dos experimentos realizados com a análise de polaridade, para responder as questões de pesquisa Q1 e Q2.

Tabela 2. Resultados da avaliação de análise de polaridade

Algoritmo	***Corpus* pré-processado**	$F1$	
Naive Bayes	Original	47,5%	Q1
Naive Bayes	Enelvo Raw sem emoji e *retweet*	53,1%	Q2
SVM	Original	55,5%	Q1
SVM	Enelvo Raw sem emoji e *retweet*	57,8%	Q2
BERT	Original	92,4%	Q1
BERT	Enelvo Raw sem emoji e *retweet*	96,6%	Q2

De acordo com os valores de $F1$ apresentados nesta tabela, é possível notar que o BERT obteve um resultado superior comparado aos algoritmos até então investigados para esse *corpus* em [Cristiani et al. 2020]. BERT alcançou uma $F1$ de 96,6% enquanto o melhor desempenho relatado em [Cristiani et al. 2020] tinha sido de 66,2%, utilizando SVM. No experimento deste trabalho, o algoritmo SVM conseguiu uma F1 de 57,8%. Assim, esses resultados respondem a primeira questão de pesquisa **Q1** mostrando que o BERT, refinado para o problema de análise de polaridade com o *corpus* de [Cristiani et al. 2020], apresenta o novo melhor desempenho obtido neste *corpus*.[11]

Quanto à **Q2**, nota-se que os resultados de $F1$ obtidos para os diferentes processamentos do *corpus* ficaram próximos para os diferentes modelos. Contudo, em todos os casos, o *corpus* Enelvo Raw sem emoji e *retweet* se saiu um pouco melhor, especialmente no Naive Bayes. Assim, respondendo à **Q2**, conclui-se que o pré-processamento proposto e adotado neste trabalho traz ganhos para a tarefa de análise de polaridade.

5.2. Análise qualitativa de tópicos

Para a análise qualitativa de tópicos usando o BERTopic, o *corpus* com 369.800 *tweets* descrito na seção 4.1 foi processado pelo modelo de análise de polaridade treinado com o BERT. Dos 369.800 *tweets*, 346.243 obtiveram rótulos gerados pelo modelo (um aproveitamento de $93,6\%$). Destes, $54,7\%$ são de *tweets* citando o candidato Jair Bolsonaro e $45,3\%$ citando o candidato Fernando Haddad. Cada um dos *tweets* foi classificado em positivo, negativo, neutro ou em mais de um rótulo.

[11] Os valores de SVM e Naive Bayes apresentados na Tabela 2 diferem dos apresentados em [Cristiani et al. 2020] porque não foi possível replicar a partição de treino e teste utilizada no experimento original. Entretanto, mesmo tendo resultados um pouco abaixo nos métodos SVM e Naive Bayes na experimentação deste trabalho, é possível afirmar que o uso do BERT gerou um salto significante de $F1$, indo para os 96,6%, enquanto a $F1$ máxima obtida pelo método SVM no trabalho de [Cristiani et al. 2020] foi de 66,1%.

A quantidade de *tweets* rotulados como positivos ($59,4\%$) é maior do que a de $tweets$ negativos ($21,6\%$) e neutros ($18,1\%$).[12] O candidato Fernando Haddad ficou com uma maior quantidade dos *tweets* positivos ($31,3\%$ contra $28,1\%$) e uma menor quantidade de *tweets* negativos ($7,7\%$ contra $13,9\%$) comparado ao candidato Jair Bolsonaro.

O *corpus* anotado automaticamente com polaridade foi, então, dividido em quatro conjuntos distintos de *tweets* para realizar a análise de tópicos: `Bolsonaro+`, `Bolsonaro-`, `Haddad+` e `Haddad-`. Para cada um desses conjuntos foram gerados 20 tópicos com 5 palavras cada. Destes 20 tópicos, foram selecionados os 9 primeiros, numerados de 0 a 8, uma vez que quanto maior o número do tópico, menos frequente ele é. As palavras dentro de cada tópico são ordenadas pela relevância gerada pelo *c-TF-IDF*. A Tabela 3 resume os principais resultados da análise de tópicos nestes quatro conjuntos.

Tabela 3. Análise de tópicos extraídos pelo BERTopic

Conjunto	Tópicos	Análise
`Bolsonaro+`	carvalho, hasselmann, paschoal	Partes dos nomes de alguns dos principais aliados de Bolsonaro em 2018: Olavo de Carvalho, Joyce Hasselmann e Janaína Paschoal
	acabouapiranhagempt, vitória, imparcial	Partes de *slogans* ou frases feitas de campanha, além de características mencionadas por possíveis apoiadores de Bolsonaro
`Bolsonaro-`	homofóbicos, fascistas, desrespeitarem, ódio	Termos encontrados em *tweets* com polaridade (argumentos) negativos em relação a Bolsonaro
`Haddad+`	haddadpresidente, trabalhando, vença, democracia, respeito, imparcial	Partes de *slogans* ou frases feitas de campanha, além de características mencionadas por possíveis apoiadores de Haddad
	indecisos	A palavra "indecisos" também apareceu, uma vez que naquela época uma das chances de virada de Haddad era na conversão dos votos de pessoas indecisas
`Haddad-`	xingando, meme	Poucas palavras que levam à conotação negativa apareceram relacionadas a Haddad, enquanto o restante foram palavras neutras, como "campanha", "votando" e "falando".

Sobre a **Q3**, pode-se concluir que o BERTopic se mostrou uma ferramenta bastante promissora para a extração de tópicos nos diferentes conjuntos, mostrando tópicos que faziam sentido em todos os cenários. Porém, vale ressaltar que a análise de tópicos foi a ponta final do experimento, carregando erros tanto de associação do *tweet* ao candidato correto quanto de falsos positivos e falsos negativos na classificação pelo modelo BERT.

6. Trabalhos futuros

Considerando os trabalhos futuros, há uma margem para melhorias nos métodos de pré-processamento, visto que o uso do Enelvo se tornou computacionalmente inviável para grandes quantidades de texto. Pode-se estudar métodos de melhoria de desempenho deste algoritmo utilizando técnicas de paralelização, por exemplo. Também existe uma margem para melhorias na extração de tópicos, buscando métodos de identificação de entidades e removendo as entidades neutras ou com baixo significado. Vale ressaltar que a metodologia e as ferramentas empregadas neste trabalho podem ser utilizadas em outros tipos de pesquisa relacionadas a conteúdos textuais em língua portuguesa.

[12] $0,9\%$ dos *tweets* foram rotulados em duas classes.

Referências

Bertaglia, T. F. C. e Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

Campello, R. J. G. B., Moulavi, D., e Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., e Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christhie, W., Reis, J. C. S., Moro, F. B. M. M., e Almeida, V. (2018). Detecção de posicionamento em tweets sobre política no contexto brasileiro. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil. SBC.

Cristiani, A., Lieira, D., e Camargo, H. (2020). A sentiment analysis of brazilian elections tweets. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 153–160, Porto Alegre, RS, Brasil. SBC.

Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, page 4171–4186.

Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

Moreira, R., Vaz de Melo, P., e Pappa, G. (2020). Elite versus mass polarization on the brazilian impeachment proceedings of 2016. *Social Network Analysis and Mining*, 10(92).

Pereira, J. G. (2019). Análise de sentimentos da população brasileira em relação a eleição presidencial de 2018 através da rede social twitter. Trabalho de Conclusão de Curso (Bacharelado) - Universidade Federal do Rio Grande do Norte. Centro de Ensino Superior do Seridó. Departamento de Computação e Tecnologia.

Pinto, M. A. S., Junior, A. F. L. J., Busson, A. J. G., e Colcher, S. (2020). Relacionando modelagem de tópicos e classificação de sentimentos para análise de mensagens do twitter durante a pandemia da covid-19. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 61–64, Porto Alegre, RS, Brasil. SBC.

Sales, M. L. e Barbosa, M. W. (2019). Uma avaliação do potencial de uso dos dados do Twitter para a predição do resultado de eleições: O caso das eleições presidenciais brasileiras de 2018. *Revista de Informática Aplicada - RIA*, 15(2):30–43.

Silveira, R., Fernandes, C. G., Neto, J. A. M., Furtado, V., e Filho, J. E. P. (2021). Topic modelling of legal documents via legal-bert1. In *RELATED - Relations in the Legal Domain Workshop, in conjunction with ICAIL 2021*, São Paulo, Brazil.

Souza, F., Nogueira, R., e Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Lecture Notes in Computer Science – Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, volume 12319, pages 403–417.