Análise de Sentimentos em Tweets

Nome	E-mail	Registro do Aluno
Adriano Mamoru Takeshita	10415887@mackenzista.com.br	10415887
Hideki Nakamura	hidekinakamura@gmail.com	10415167
Pedro Henrique Gonçalves Machado	pedrohmachado@gmail.com	10414885
Vitor Baldoino	vdbaldoino@gmail.com	10414498

1. Objetivo

O aumento do uso das redes sociais e a disponibilização de quantidades massivas de textos online fez surgir o interesse em se analisar e compreender o conteúdo disponível nos mais diversos meios. Uma das frentes de pesquisa que se encarrega de tal tarefa é a Mineração de Argumentos, cujo objetivo é identificar, extrair e compreender a estrutura argumentativa de textos online (Sousa et al. 2021). Como toda grande tarefa, a mineração de argumentos é dividida em conjuntos menores de trabalhos que podem ser realizados individualmente. Um deles — talvez o mais popular — é a análise de sentimento, que consiste em identificar se um dado documento ou pedaço de texto carrega uma conotação positiva, negativa ou neutra.

Partindo das definições acima, o objetivo do presente trabalho é executar[1] a análise de sentimentos contidos em Tweets coletados entre 01/08/2018 e 20/10/2018. Cabe ressaltar que a análise de sentimento é uma tarefa pertinente para contextos acadêmicos e não acadêmicos, possuindo grandes aplicações no setor privado (Abirami e Askarunisa 2017; Shukri et al. 2015; Souza et al. 2015).

2. Dados

A base de dados utilizada neste trabalho contém o texto dos tweets coletados no período entre 01/08/2018 e 20/10/2018, bem como uma coluna informativa da classe — positiva, negativa ou neutra — do tweet. Segundo as informações do texto de apresentação da base[2], a mesma foi rotulada utilizando um método de anotação distante, tomando como

[1] Todo o trabalho foi feito em Python. Os *scripts* criados podem ser encontrados neste repositório do GitHub.

[2] A base pode ser consultada aqui.

inspiração o apresentado em Go, Bhayani, e Huang (2009) e Kouloumpis, Wilson, e Moore (2011).

A base original é dividida em diversas partes menores e agregadas por tema, com o total de 900.688 tweets. Entretanto, no presente trabalho, utilizaremos a base completa e sem distinção de tópico do tweet. Nessa configuração, os dados são desbalanceados, como é possível ver na Figura 1.

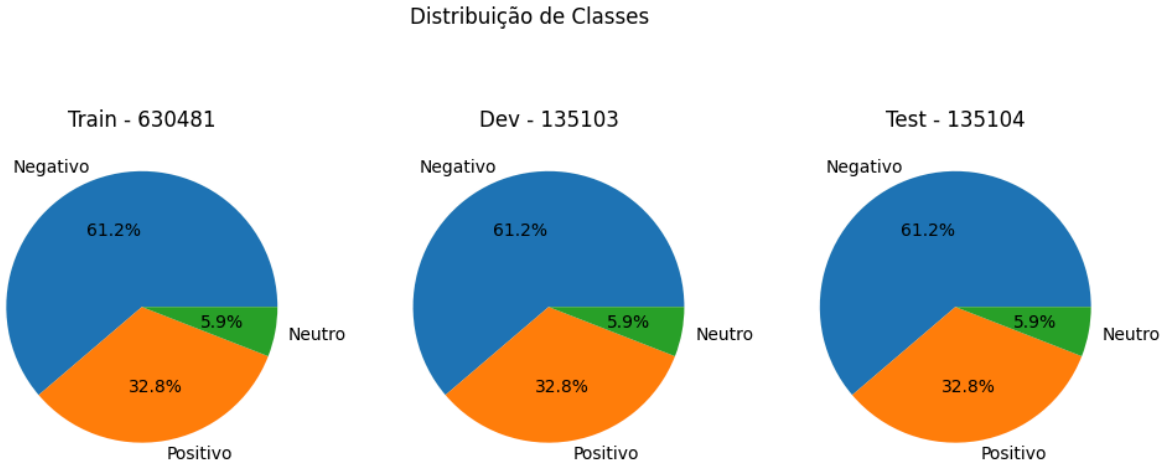

Figura 1. Descrição dos conjuntos de treino, validação e teste do *dataset*. Os números ao lado do título indicam a quantidade de observações em cada conjunto da base de dados.

Além do desafio que o desbalanceamento de classes oferece, é importante ressaltar a natureza do conteúdo das redes sociais. Os textos obtidos frequentemente: (i) são de baixa qualidade e/ou informais, caracterizados por erros gramaticais, abreviações e linguagem coloquial; e (ii) podem não carregar sentido semântico, devido à falta de contexto ou à ambiguidade na comunicação; e (iii) possuem um tamanho reduzido, limitando a expressão completa de ideias e nuances. A Figura 2 e a Figura 3 fornecem exemplos concretos desses desafios, ilustrando a frequência de termos como "https" e o tamanho médio dos tweets, respectivamente, evidenciando as limitações impostas pela natureza dos dados das redes sociais.

Figura 2. Nuvem de palavras construída a partir dos cem mil primeiros tweets do dataset de treino. Quanto maior o termo, mais frequente ele é na base de dados.

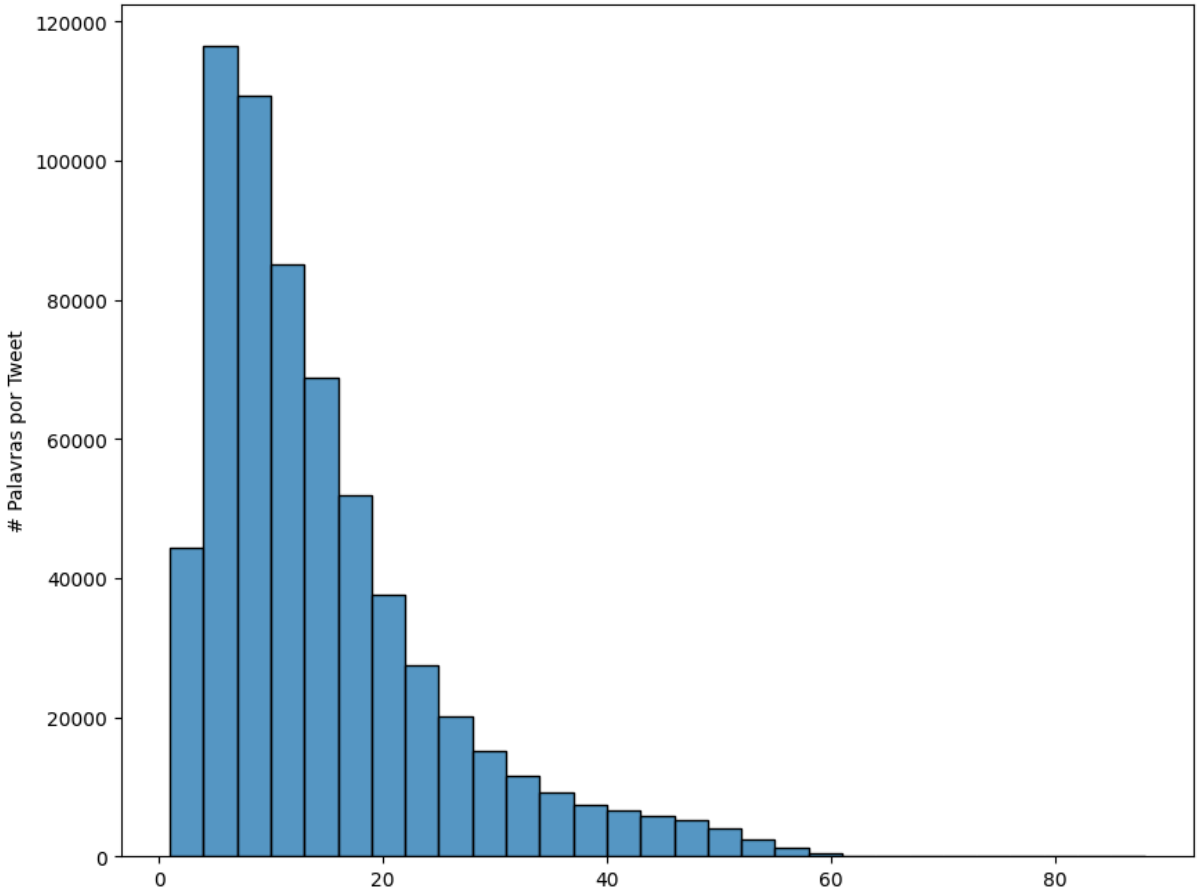

Figura 3. Histograma da quantidade de palavras por tweet do dataset de treino.

Método Analítico

A estratégia atual do projeto é realizar o *fine-tuning* do BERTimbau (Souza, Nogueira, e Lotufo 2020), um modelo que segue a mesma estratégia de treinamento do BERT (Devlin et al. 2019), mas com dados em português. Os modelos supracitados são baseados na arquitetura Transformer (Vaswani et al. 2017), que, até o momento, constitui o estado da arte para o processamento de linguagem natural. O principal fundamento do Transformer é a **atenção**, um mecanismo que permite que a rede neural focalize nas partes mais importantes do texto de entrada durante o processamento. Dentro dessa arquitetura, as camadas de ***self-attention*** realizam uma ponderação com todas as palavras da mesma sequência, permitindo a detecção de dependências ainda que as palavras estejam distantes umas das outras. A Figura 4 apresenta a representação geral do Transformer.

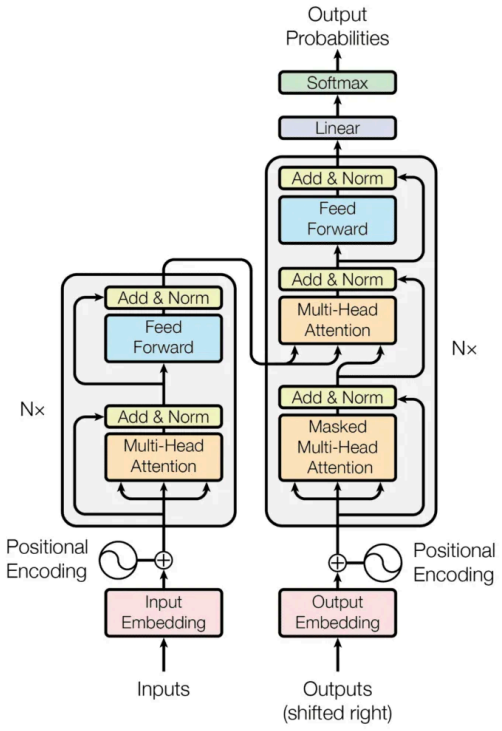

Figura 4. Arquitetura do Transformer. Retirado de Vaswani et al (2017).

3. Cronograma

	Fevereiro	Março	Abril	Maio
Formulação dos objetivos do trabalho	X			
Tratamento dos dados e treinamento do modelo		X		

Avaliação dos resultados e verificação de qualidade da modelagem	X	
Apresentação do trabalho e envio do documento final		X

4. Referências

Abirami, A. M., e A. Askarunisa. 2017. “Sentiment analysis model to emphasize the impact of online reviews in healthcare industry”. *Online Information Review* 41(4):471–86. doi: 10.1108/OIR-08-2015-0289.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. P. 4171–86 em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, organizado por J. Burstein, C. Doran, e T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics.

Go, Alec, Richa Bhayani, e Lei Huang. 2009. “Twitter sentiment classification using distant supervision”. *CS224N project report, Stanford* 1(12):2009.

Kouloumpis, Efthymios, Theresa Wilson, e Johanna Moore. 2011. “Twitter Sentiment Analysis: The Good the Bad and the OMG!” *Proceedings of the International AAAI Conference on Web and Social Media* 5(1):538–41. doi: 10.1609/icwsm.v5i1.14185.

Shukri, Sarah E., Rawan I. Yaghi, Ibrahim Aljarah, e Hamad Alsawalqah. 2015. “Twitter sentiment analysis: A case study in the automotive industry”. P. 1–5 em *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

Sousa, João Pedro da Silva, Rodrigo Costa Uchoa do Nascimento, Renata Mendes de Araujo, e Orlando Bisacchi Coelho. 2021. “Não se perca no debate! Mineração de Argumentação em Redes Sociais”. P. 139–50 em *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. SBC.

Souza, Fábio, Rodrigo Nogueira, e Roberto Lotufo. 2020. “BERTimbau: Pretrained BERT Models for Brazilian Portuguese”. P. 403–17 em.

Souza, Thársis Tuani Pinto, Olga Kolchyna, Philip C. Treleaven, e Tomaso Aste. 2015. “Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry”.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, e Illia Polosukhin. 2017. “Attention is All you Need”. em *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.