Análise de sentimentos de conteúdo compartilhado em comunidades brasileiras do Reddit: Avaliação de um conjunto de dados rotulados por humanos

Giovana Piorino
Universidade Federal de Minas Gerais
giovana.piorino@dcc.ufmg.br

Vitor Moreira
Universidade Federal de Minas Gerais
vitormoreira@dcc.ufmg.br

Luiz Henrique Quevedo Lima
Universidade Federal de Minas Gerais
luiz.quevedo@dcc.ufmg.br

Adriana Silvina Pagano
Universidade Federal de Minas Gerais
apagano@ufmg.br

Ana Paula Couto da Silva
Universidade Federal de Minas Gerais
ana.coutosilva@dcc.ufmg.br

ABSTRACT

The soaring use of social media and its impact on society have been raising ethical issues about the content disseminated by these platforms, particularly from the perspective of responsible AI given the need to mitigate the propagation of bias and the spread of toxic language. Sentiment Analysis of the language of these communities poses big challenges, since it requires quality datasets that can be used in supervised training of models. The social network Reddit comprises smaller, sub-communities centered on specific topics, called Subreddits. Through manual annotation of posts in Subreddits related to Brazilian content and communities, we have developed a dataset for Sentiment Analysis in Brazilian Portuguese. We report the results of our annotation process and characterize the language of the posts. Our dataset is meant to support Sentiment Analysis tasks for social media language in Brazilian Portuguese.

KEYWORDS

Análise de Sentimentos, Comunidades do Reddit, Tarefa de Anotação, Português Brasileiro

1 INTRODUÇÃO

As redes sociais têm rompido barreiras de comunicação, proporcionando às pessoas a oportunidade de interagir com amigos e familiares ao redor do mundo, participar de discussões e tomar conhecimento dos mais variados assuntos [2]. Além disso, têm permitido a rápida divulgação de questões atuais [33]. A cada ano, aumenta o número de usuários, com taxas de crescimento maiores que 5% ao ano, alcançando 5,07 bilhões de usuários no início de abril de 2024 [18]. Contudo, essa expansão também trouxe um número crescente de pessoas vulneráveis que veem suas emoções serem afetadas negativamente devido às interações nessas plataformas [19]. Nessa perspectiva, uma pesquisa da Universidade de Torino, com o objetivo de compreender o efeito da agressão cibernética em adultos na Itália, aponta que, dos 341 participantes, 43% disseram ter sido vítimas de cyber agressão. Segundo 95,1% dos participantes, esses episódios ocorreram em redes sociais e foram considerados potencialmente danosos à saúde mental [22].

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

Dado o crescimento do conteúdo gerado nas redes sociais diariamente (mais de 4,4 bilhões de postagens no segundo semestre de 2023 no caso da rede Reddit [32]), sua moderação tem se tornado um problema desafiador e dispendioso. Para lidar com isso, grandes empresas como X, antigo Twitter, adotaram modelos de aprendizado de máquina e revisão humana [39]. Esses modelos ajudam as plataformas a tomar medidas que julgarem adequadas em relação a conteúdos identificados como violadores de suas diretrizes. Contudo, ainda que existam modelos aptos para realizar essas tarefas, há limitações no seu desempenho em idiomas com menor disponibilidade de dados, como o português brasileiro.

Algumas iniciativas têm desenvolvido pesquisas e conjuntos de dados em português brasileiro com anotação manual de sentimentos em textos de redes sociais. A maior parte dos trabalhos têm privilegiado a rede social Twitter, cujos textos possuem padrões específicos de linguagem, razão pela qual modelos treinados com eles mostram limitações quando aplicados a textos de outras redes sociais. No caso do Reddit, não encontramos nenhum conjunto de dados de textos originalmente redigidos em português brasileiro, extraídos do Reddit e anotados com rótulos de sentimentos.

Buscando expandir os recursos de PLN para o português brasileiro, este artigo apresenta um novo conjunto de dados anotados para análise de sentimentos. Os textos anotados foram retirados do Reddit.[1] Reddit é uma comunidade que permite aos usuários interagirem por meio de postagens (submissões) e comentários anônimos. Os usuários são organizados em comunidades (subreddits) e se inscrevem nas comunidades mais alinhadas com seus tópicos de interesse. Os dados anotados com um dos sentimentos (*positivo*, *negativo* ou *neutro*) e a caracterização linguística dos textos categorizados em cada uma das classes de sentimento pode auxiliar na proposta de novos modelos de classificação de sentimentos e na melhora de modelos existentes de tal modo que estes sejam mais adequados para as características específicas da língua portuguesa.

Adicionalmente, nosso trabalho explora formas de análise que lidam com a complexidade da tarefa de classificação de sentimentos, a partir do tratamento e caracterização dos casos em que os anotadores não conseguem atribuir nenhum dos três sentimentos ao comentário em análise. Por fim, para permitir a reprodutibilidade e incentivar estudos subsequentes, o conjunto de dados anotados será disponibilizado em endereço a ser divulgado na versão para publicação.

[1] http://reddit.com/

2 TRABALHOS RELACIONADOS

Estudos exploraram a análise de sentimentos em textos de redes sociais em português brasileiro, tendo alguns deles disponibilizados publicamente os conjuntos de dados utilizados nos trabalhos.

Dentre eles, os autores [36] realizaram análises textuais e de sentimento com base em textos em português brasileiro da rede social Twitter. O trabalho mostra diferentes metodologias para auxiliar análises textuais com enfoque nessa rede social, como o Twittómetro e o *Amazon Mechanical Turk*[2].

Outras análises, como em [14], também exploraram a detecção de tópicos e a análise de sentimentos em textos do Twitter no contexto brasileiro, com ênfase em temas relacionados à COVID-19. A abordagem de extração de tópicos utilizada foi a LDA (*Latent Dirichlet Allocation*), e a análise de sentimentos para os textos em português obteve auxílio do mUSE(*Multilingual Universal Sentence Encoder for Semantic Retrieval*), e do SemEval 2018. Ainda relativo ao Twitter, os autores em [40] coletaram tweets em português brasileiro de forma a compor um conjunto de dados [3] de 15.000 tweets, extraídos entre janeiro e julho de 2017. Para esse estudo, os tweets também foram classificados com os rótulos *positivo, negativo e neutro*, por anotadores cuja anotação obteve métricas de alfa de krippendorf de 0,529, considerada uma concordância moderada.

Dentre os modelos direcionados à língua portuguesa e à análise de sentimentos, tem-se o VADER[16](*Valence Aware Dictionary for Sentiment Reasoning*), que apresenta uma extensão para a língua portuguesa chamada LeIA [1] (*Léxico para Inferência Adaptada*), a qual rotula textos entre categorias *positivo, negativo e neutro*, podendo se adaptar a diferentes contextos, sem se restringir ao escopo de textos de uma rede social específica.

No que diz respeito a trabalhos relacionados à rede social Reddit, os autores [6] utilizaram o modelo *GoEmotions* baseado em um conjunto de dados com aproximadamente 58.000 comentários rotulados manualmente com categorias de emoções, redigidos em inglês e traduzidos para o português. O estudo também realizou correlações linguísticas entre as emoções identificadas e os comentários rotulados, e obteve métricas de avaliação das anotações e do modelo. No entanto, devido à grande quantidade de categorias de emoções presentes na rotulação, as métricas geraram valores, em geral, moderados ou fracos para a tarefa de anotação.

Em [17], os autores utilizaram conjunto de dados de textos extraídos do Twitter e do Reddit para avaliar distintas configurações de pipelines de pré-processamento dos textos em português brasileiro, passíveis de serem implementadas antes da aplicação de métodos de modelagem de tópicos. As adaptações avaliadas evidenciaram melhoras em todas as métricas.

Apesar do recente crescimento da rede social Reddit, ainda há poucas referências na literatura sobre análises textuais e de tarefas de anotação de sentimentos em textos dessa rede, principalmente em português brasileiro. Assim, nosso estudo busca expandir os recursos de PLN em português brasileiro, fornecendo um conjunto de dados anotado com sentimentos, juntamente com os resultados das métricas centrais de avaliação da anotação humana e a caracterização da linguagem dos textos no conjunto de dados.

[2]https://www.mturk.com/

[3]https://bitbucket.org/HBrum/tweetsentbr/src/master/

Tabela 1: Subreddits selecionados e total de postagens e comentários (2022).

Subreddit	Postagens	Comentários
r/brasil	115,876	2,382,928
r/desabafos	115,876	1,487,076
r/futebol	35,826	1,272,009
r/saopaulo	7,308	88,894
r/eu_nvr	12,631	221,348
r/botecodoreddit	7,059	62,999
r/conversas	21,967	355,761
r/investimentos	9,756	156,695
r/tiodopave	2,371	12,106
r/brasilivre	67,301	1,308,441
Total	390,924	7,348,257

3 METODOLOGIA

Nesta seção primeiramente apresentamos a metodologia utilizada para coletar o conjunto original de dados. A seguir, descrevemos a processo de anotação manual de um conjunto selecionado dos dados. Por fim, apresentamos os métodos usados para a análise linguística dos comentários anotados.

3.1 Conjunto de Dados

Reddit é uma mídia social online organizada em subcomunidades por áreas de interesse ou subreddits, nas quais usuários discutem diferentes assuntos, através de interações do tipo postagem-comentários, chamadas de *threads*. Nossa base original de dados consiste em atividades de usuários (postagens e comentários)[4] entre os meses de janeiro e dezembro de 2022 realizadas nas 10 comunidades brasileiras com maior número de usuários ativos. A Tabela 1 apresenta as principais estatísticas das 10 comunidades selecionadas. Os dados foram coletados a partir da plataforma Pushshift, que coleta, analisa e arquiva conteúdos do Reddit desde 2015 [4]. Estes dados foram previamente apresentados em [21] e utilizados para a tarefa de classificação de toxicidade dos comentários compartilhados nestes subreddits.

3.2 Anotação dos dados

Para a classificação manual do sentimento associado a cada comentário, foram selecionados 2,000 comentários da base original coletada, seguindo uma amostra estratificada do total de comentários em cada comunidade analisada. Estes comentários foram divididos em 4 grupos, com 500 comentários cada, denominados *Grupo1, Grupo2, Grupo3, Grupo4*. Cada grupo foi anotado por 3 anotadores distintos.

Os anotadores são estudantes universitários convidados a participar de forma anônima e instruídos a ler e classificar cada comentário como *Positivo, Negativo, Neutro ou Não sei dizer*, levando em consideração o sentimento predominante em cada texto. Caso não fosse possível determinar um sentimento, a opção a ser escolhida deveria ser *Não sei dizer*. Para auxiliar a identificação do sentimento predominante, foi sugerido aos anotadores ter atenção especial a dois pontos: (i) os comentários *negativos* geralmente manifestam

[4]O termo 'comentários' será utilizado abrangendo comentários e postagens.

emoções de medo, culpa, mágoa, tristeza, raiva, angústia, ansiedade e depressão; e (ii) os comentários *neutros* não apresentam nenhuma característica que possa levar a sua classificação como negativos ou positivos.

Ao final do processo de anotação, cada comentário recebeu o rótulo com o sentimento atribuído pela maioria dos anotadores e a concordância entre os avaliadores foi medida por três métricas comumente usadas: *Kappa de Fleiss* [7], *Alpha de Krippendorf* [20] e *Concordância Observada* [8].

3.3 Classificação Automática de Sentimentos

Para medir a correlação entre a classificação automática e manual de sentimentos nos comentários amostrados do Reddit, escolhemos o modelo XLM-RoBERTa (*Cross Lingual Language Model - Robustly Optimized BERT-Pretraining Approach*) [5] como nosso *baseline*, que está disponível na biblioteca *Hugging Face*.

O modelo utilizado já havia passado por um ajuste fino baseado em textos da rede social *Twitter* em português. A escolha desse modelo se deve à sua grande base treinada em aproximadamente 10 milhões de *tweets* na língua portuguesa [3] e a o modelo ser direcionado à tarefa de análise de sentimentos.

3.4 Análise Textual

Antes de iniciar a análise textual, os 2.000 comentários anotados foram submetidos a filtros, utilizando-se expressões regulares [13], com o objetivo de detectar o conteúdo dos comentários a serem excluídos da análise: endereços de sites, menções a outros usuários, hashtags, textos citados, datas ou emojis. Risadas expressas em texto foram removidas, assim como comentários contendo palavras gramaticais que ocorriam isoladamente e sem valor de informação para nossa análise, com base na lista de (*stopwords*) da biblioteca NLTK[26] e em um modelo do spaCy[37]. Assim, 19 comentários foram removidos das análises após os filtros.

3.4.1 Razão Type-Token Type-Token Ratio (TTR). Com a tokenização dos comentários feita pela biblioteca [25], determinamos a diversidade lexical usando a medida TTR. O resultado do TTR advém do número de tokens distintos dividido pelo número total de tokens existentes no comentário. Complementamos a análise avaliando o tamanho (em número de tokens) dos comentários de cada grupo.

3.4.2 Etiquetagem de classe de palavra (Pos Tagging). Para examinar as classes de palavra predominantes nos comentários rotulados, fizemos o POS tagging [28] com um modelo pré-treinado [37], baseado em um treebank anotado de acordo com o padrão das Universal Dependencies [11]. Esse treebank tem como principal base o trabalho de [30].

3.4.3 Reconhecimento de Entidades Nomeadas (REN). Exploramos as entidades nomeadas através do uso de um modelo pré-treinado do spaCy. Empregamos novamente o modelo utilizado no Pos Tagging, sendo o conjunto de dados utilizado para treinar esse modelo o WikiNER [27]. Essa técnica classifica as entidades em 3 categorias: PESSOA (PER), LOCALIZAÇÃO (LOC) e ORGANIZAÇÃO (ORG). Entidades que não se enquadram nessas categorias são classificadas como DIVERSAS (MISC).

[5]https://huggingface.co/docs/transformers/model_doc/xlm-roberta

3.4.4 Análise de n-gramas. . Para complementar as análises linguísticas, realizamos a análise de *n*-gramas. Um n-grama é uma sequência contígua de *n* itens de uma determinada amostra de texto.

3.4.5 Classificação de tópicos (BERTopic). Para a extração de tópicos dos comentários utilizamos o modelo BERTopic [15], a fim de caracterizar os conteúdos mais frequentes dos textos, e como eles se relacionam com os sentimentos rotulados pelos anotadores e pelo modelo automático RobERTa. Os comentários foram convertidos em vetores de representação com auxílio do modelo BERTimbau [35], no qual há um estágio adicional de ajuste fino direcionado à similaridade de semântica textual [9] [23] [31].

Para garantir uma modelagem mais consistente dos tópicos foi realizada a redução da dimensionalidade dos vetores por meio do UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*), técnica que melhora agrupamentos subsequentes. Também foi utilizado o algoritmo HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*), obtendo-se um agrupamento dos vetores de representação a partir de similaridades semânticas. Por fim, c-TF-IDF (*Class-based Term Frequency-Inverse Document Frequency*) e MMR (*Maximal Marginal Relevance*) foram aplicados e ajustados para melhorar a definição de palavras-chave para os tópicos e para diversificar seu conteúdo semântico, respectivamente. Foram utilizadas recomendações presentes na documentação do modelo[6] para o ajuste de tais parâmetros, sendo que para o UMAP, o número de vizinhos foi ajustado para 10, e o número de componentes, para 8. Para o HDBSCAN, o número mínimo do tamanho de agrupamentos é de 10, e número mínimo de amostras, 8. O parâmetro MMR foi atualizado para uma taxa de 0.8.

3.4.6 Rotulações semânticas (PyMUSAS). Para a análise semântica dos comentários, foi utilizada a ferramenta pyMUSAS, baseada na estrutura USAS[7] (*UCREL Semantic Analysis System*) adaptada à linguagem Python. Essa classificação apresenta modelos em diferentes línguas, incluindo o português [29], um dos motivos para seu uso neste trabalho.

Resumidamente, ela apresenta uma estrutura organizada em códigos, que representam categorias semânticas distintas [34]. Cada comentário pode ser enquadrado em uma ou mais categorias semânticas, fornecendo uma visão ampla e abstrata dos conteúdos presentes nos comentários e como eles estão relacionados aos sentimentos rotulados.

4 RESULTADOS

Nesta seção, apresentamos os principais resultados obtidos na avaliação e caracterização do conjunto de dados anotados.

4.1 Concordância entre anotadores

As métricas para analisar os resultados de concordância entre os anotadores foram aplicadas aos subconjuntos *Todas as anotações*, abrangendo todos os comentários rotulados com as quatro categorias disponíveis e *Apenas sentimentos*, abrangendo comentários rotulados desconsiderando o rótulo *Não sei dizer*, de forma a verificar o impacto desse rótulo de incerteza nos resultados. A Tabela 2 apresenta os resultados. O Alfa de Krippendorf e o Kappa de Fleiss

[6]https://maartengr.github.io/BERTopic/index.html
[7]https://ucrel.lancs.ac.uk/usas/

apresentaram valores que podem ser interpretados como concordância moderada entre os anotadores. Já a concordância observada apresenta valores consideravelmente maiores que as outras métricas, porém não apresenta tanta robustez, por não considerar que a concordância entre anotadores possa ter acontecido ao acaso. Em geral, observa-se que a qualidade das métricas melhora consideravelmente ao incluir apenas os comentários rotulados com os sentimentos e desconsiderar a categoria *Não sei dizer*.

No que diz respeito à concordância total entre anotadores, isto é, todos os anotadores indicando o mesmo rótulo, o percentual de comentários que obtiveram concordância total foi 44.65% no subconjunto que considera todos os rótulos de anotação. Já no subconjunto de comentários apenas com rótulos de sentimentos, a concordância total aumenta para 57%. Alguns exemplos desses comentários se encontram na Tabela 3.

A fim de estabelecer classificações de sentimentos para análises posteriores de comparação com modelos automáticos e caracterização dos textos, atribuímos às ocorrências de concordância parcial o sentimento anotado predominante, isto é, a concordância de dois ou mais anotadores sobre um mesmo rótulo. A Tabela 4 mostra que quase metade do conjunto de comentários foi majoritariamente rotulado como *negativo*, indicando um desbalanceamento de classes considerável. Já os rótulos *positivo* e *neutro* obtiveram proporções semelhantes. 10,55% dos comentários obtiveram discordância total, ou seja, cada anotador apontou um rótulo diferente. Este valor de discordância pode ser o resultado de diferentes perspectivas que cada anotador pode ter do que é algo positivo ou negativo [24], ou a presença de conteúdo com teor de sarcasmo ou falta de contexto adicional para facilitar a atribuição de um sentimento.

A Tabela 5 indica o desempenho de cada trio de anotadores e suas respectivas métricas, sendo possível observar que a anotação dos comentários pertencentes ao grupo 4 obteve o melhor e aqueles do grupo 3 o pior desempenho. No entanto, em geral, as anotações obtiveram métricas de concordância razoavelmente próximas, apontando para concordâncias médias e moderadas entre seus respectivos anotadores. De forma complementar, a Tabela 6 apresenta a rotulação de sentimentos para cada anotador, dentro de cada grupo de comentários.

Uma análise interessante a ser realizada está relacionada ao grau de incerteza presente na tarefa de rotulação. No total dos 2,000 comentários rotulados, a opção *Não sei dizer* foi selecionada por pelo menos um dos três anotadores em 23,05% do conjunto total. No entanto, apenas em 4,1% dos comentários, dois ou mais anotadores rotularam o mesmo comentário com *Não sei dizer*, uma queda significativa que pode indicar que há uma dificuldade maior em 2 ou mais anotadores caracterizarem incerteza de sentimento para um mesmo texto. Um exemplo de texto em que 2 ou mais anotadores apresentaram incerteza na rotulação é: *"Curti muito sua dupla personalidade, hehe."*, o que parece ser uma frase de sarcasmo, dificultando ainda mais a tarefa de rotulação, mesmo para humanos.

Observamos também grande variação nas categorias escolhidas pelos anotadores. A Tabela 7 apresenta os resultados das métricas de avaliação de concordância entre anotadores em cada grupo de comentários. Vemos que para um mesmo grupo de textos, o anotador 1 assinalou 13,60% dos comentários com *Não sei dizer*, enquanto que o anotador 3 atribuiu esse rótulo a apenas 0,40% dos comentários. Tais resultados reafirmam o apontado na literatura sobre a subjetividade nas avaliações de sentimento e a dificuldade dessa tarefa. Adicionalmente, analisando os dados correspondentes na Tabela 6, observa-se que o grupo 1 apresentou, em geral, maior classificação de comentários *positivos*, e o anotador 3 desse grupo foi o que mais rotulou positivamente, por uma grande margem de diferença em relação aos demais. Por outro lado, o anotador 2 do grupo 4 foi o que mais rotulou negativamente, ainda que, em geral, obteve um proporção de anotações consideravelmente constante em relação aos outros anotadores de seu grupo, que é o que apresentou as melhores métricas de concordância. O grupo 3, grupo com as mais baixas métricas de concordância, apresentou disparidades consideráveis entre as proporções de rotulações, tendo o anotador 1 desse grupo demonstrado discrepância razoável de proporções de rotulações em relação aos anotadores 2 e 3.

Tabela 2: Concordância entre anotadores.

Métrica	Todas as anotações	Apenas sentimentos
Kappa de Fleiss	0,40	0,51
Alfa de Krippendorf	0,47	0,53
Concordância observada	0,60	0,70

Tabela 3: Exemplos de comentários que obtiveram concordância total entre os anotadores

Sentimento	Exemplo de comentário
Positivo	Ahh para, eu curto cidadezinha, as vezes eu vou pra uns lugares desse, fico uns 2 ou 4 dias, acho super legal.
Negativo	Intervencionismo externo visando ganho próprio e sem estudar a situação complexa e possiveis consequencias. Um clássico dos estados unidos de m*rda
Neutro	Subsidio para quem vender preferencialmente para o mercado interno ou o contrario, cobrar mais imposto sobre o produto exportado.

Tabela 4: Porcentagem de comentários para cada agrupamento de rotulação e para a discordância total.

Classificação	Porcentagem
Negativo	48,05%
Neutro	20,95%
Positivo	16,30%
Discordância total	10,55%
Não sei dizer	4,15%

Tabela 5: Métricas de avaliação para concordância entre anotadores para cada grupo desconsiderando o rótulo *Não sei dizer*.

Métrica	Grupo1	Grupo2	Grupo3	Grupo4
Kappa de Fleiss	0,41	0,39	0,34	0,44
Alfa de Krippendorf	0,48	0,48	0,40	0,50
Concordância observada	0,60	0,58	0,56	0,64

Tabela 6: Distribuição de rotulação de sentimentos para cada anotador.

	Grupo1			Grupo2			Grupo3			Grupo4		
	Anotador 1	**Anotador 2**	**Anotador 3**	**Anotador 1**	**Anotador 2**	**Anotador 3**	**Anotador 1**	**Anotador 2**	**Anotador 3**	**Anotador 1**	**Anotador 2**	**Anotador 3**
Positivo	22,8%	16,6%	35,4%	19,6%	18,8%	19,8%	17,6%	24,0%	10,0%	15,2%	15,4%	14,8%
Negativo	47,6%	46,8%	38,4%	45,4%	50,2%	44,0%	55,6%	43,6%	48,0%	42,0%	57,2%	46,6%
Neutro	16,0%	28,0%	25,8%	24,4%	21,4%	14,0%	19,4%	27,2%	25,2%	25,8%	27,2%	38,4%
Não sei dizer	13,6%	0,86%	0,4%	10,6%	9,6%	22,2%	7,4%	5,2%	16,8%	17,0%	0,2%	0,2%

Tabela 7: Porcentagem de comentários categorizados como *Não sei dizer* por anotador e grupo.

Anotadores	Grupo1	Grupo2	Grupo3	Grupo4
Anotador 1	13,60%	10,60%	7,40%	17,00%
Anotador 2	8,60%	9,60%	5,20%	0,20%
Anotador 3	0,40%	22,20%	16,80%	0,20%

4.2 Comparação com XLM-RoBERTa

O rótulo final de cada comentário foi atribuído com base na concordância de 2 ou mais anotadores. Feita essa atribuição, a fim de ajustar os rótulos aos que estão presentes no modelo XLM-RoBERTa (*Positivo*, *Negativo* e *Neutro*), foram removidos comentários em que houve discordância total entre anotadores e comentários em que a maioria dos anotadores selecionou *Não sei dizer*, resultando num total de 1.706 comentários utilizados para comparação com o modelo treinado.

Após essa etapa, realizamos a comparação entre os anotadores e o modelo, que obteve acurácia de 62,37% (porcentagem de comentários em que o rótulo indicado pelo modelo coincidiu com a anotação humana) e Kappa de Cohen de 0,34, considerado razoavelmente fraco [5]. O modelo apresentou as seguintes porcentagens de sentimentos: 12,49% de comentários *positivos*, 60,90% de comentários *negativos* e 26,61% de comentários *neutros*. A taxa de rótulos *negativos* foi consideravelmente superior à da anotação humana, que é de 48,05%.

A distribuição da concordância entre os grupos foi bem similar, com variações entre 60% - 65% de concordância do modelo em relação aos anotadores. O grupo 3 apresentou a menor taxa de concordância com modelo, com 60,66%, e o grupo 4 apresentou a melhor taxa, com 65,27%. Tais resultados são análogos aos obtidos com as métricas de concordância entre anotadores apresentadas anteriormente, em que os grupos 3 e 4 apresentaram o pior e o melhor desempenho, respectivamente.

O modelo apresentou uma taxa de concordância para rótulos negativos razoavelmente alta entre os grupos, variando de 75,10%-83,33%, enquanto que as taxas de concordância para rótulos positivos foram as mais baixas entre os grupos, abrangendo uma porcentagem de rótulos indicados corretamente entre 33,01% e 38,24%. Isso contrasta com o fato de que sua taxa de rotulações positivas é similar à taxa dos anotadores, indício de que o modelo exibe grande dificuldade para identificar corretamente um comentário *positivo*. Tal observação se relaciona com os resultados das principais métricas apresentadas na Tabela 8, em que a classe *negativo* apresentou valores maiores e mais consistentes que as demais, além do desbalanceio da amostra citado anteriormente, havendo ocorrências significantemente maiores de rótulos negativos. Para rotulações

Tabela 8: Métricas de comparação entre anotadores e modelo XLM-RoBERTa.

Classe	Precisão	Recall	F1-Score
Positivo	0,54	0,35	0,42
Negativo	0,72	0,78	0,75
Neutro	0,44	0,48	0,46
Média Macro	0,57	0,54	0,54
Média Ponderada	0,62	0,62	0,62

Tabela 9: Quantidade de comentários para cada agrupamento de rotulação e para a discordância total.

Classificação	Quantidade de Comentários
Negativo	960
Neutro	413
Positivo	319
Discordância total	210
Não sei dizer	79

para *negativo*, o grupo 1 apresentou maior taxa de concordância em relação ao modelo, com 83,33%. Para os rótulos *positivo* e *neutro*, o grupo 4 obteve a maior taxa, com 38,24% e 56,64% respectivamente.

Também buscamos identificar características dos textos em que o modelo fez uma predição do rótulo errado. Um exemplo de comentário que obteve concordância total entre anotadores, mas que o modelo errou em sua predição de rótulo, é: *"Dai querem vir opinar no nosso jogo. Americano é f*da… bom que não entendem nada"*, que os anotadores indicaram como *negativo*, mas o modelo considerou como *positivo*. Em geral, para as ocorrências de concordância total dos anotadores, o modelo obteve uma taxa de erros de 38%.

4.3 Caracterização da Linguagem

A comparação dos padrões de linguagem foi realizada por meio do agrupamento de comentários baseado no rótulo predominante das 3 rotulações feitas pelos anotadores a cada comentário. Utilizou-se um p-valor < 0,5 em todas as análises para garantir a significância estatística. A Tabela 9 exibe os dados após a filtragem de textos e, consequentemente, de comentários que não continham informações úteis. Esses dados foram utilizados nas análises realizadas.

Razão Type-Token Type-Token Ratio (TTR): Quanto à análise do TTR, existem diferenças entre as médias de caracteres por comentário nos agrupamentos, em especial, entre o agrupamento *Não sei dizer* e os demais. O agrupamento *Não sei dizer* apresentou a menor média, com 43,13 [29,08, 60,15]. O agrupamento de comentários *neutros* teve uma média de 81,41 [69,65, 94,45], já os agrupamentos de comentários *negativos* e *positivos* apresentaram

Tabela 10: Porcentagem de etiquetas pos para cada classe rotulada.

Classificação	NOUN	VERB	ADJ	PROPN	ADV
Negativo	35,51%	30,54%	18,28%	6,28%	4,17%
Neutro	34,97%	27,83%	18,08%	10,07%	3,92%
Positivo	34,49%	32,19%	17,82%	6,31%	3,66%
Não sei dizer	29,84%	26,16%	14,15%	18,41%	2,71%

médias de 98,46 [90,57, 106,66] e 99,13 [80,11, 122,28]. Esses resultados podem indicar que comentários com sentimentos negativos e positivos tendem a ser mais longos do que comentários neutros e aqueles que necessitam de mais contexto para serem interpretados, categorizados como *Não sei dizer*. No entanto, ao aplicar o teste estatístico de Mann-Whitney[8][38], não foi encontrada uma diferença entre o agrupamento *neutro* e o agrupamento *positivo*. Por outro lado, ao realizarmos o teste estatístico tanto para compararmos o agrupamento *negativo* com o *neutro* quanto para compararmos o agrupamento *negativo* com o *positivo*, foi demonstrado que há diferenças significativas entre eles.

Em relação à média e ao intervalo de confiança TTR, os agrupamentos apresentaram os seguintes valores: o agrupamento *Não sei dizer* apresenta 0,98 [0,96, 0,99], o agrupamento *neutro*, 0,97 [0,96, 0,98], o agrupamento *negativo*, 0,97 [0,96, 0,97] e o agrupamento *positivo*, 0,97 [0,97, 0,98]. A análise com o teste de Mann-Whitney indicou que o único agrupamento com diferença significativa em relação aos outros foi o agrupamento *Não sei dizer*. Nos demais resultados, o mesmo teste estatístico será utilizado e, portanto, iremos omitir o seu nome.

Etiquetagem de classe de palavra (Pos Tagging): A média e o intervalo de confiança da diversidade de etiquetas POS para cada agrupamento são os seguintes: o agrupamento *Não sei dizer* apresenta 0,73 [0,66, 0,79], o agrupamento *neutro*, 0,57 [0,55, 0,60], o agrupamento *negativo*, 0,50 [0,48, 0,51] e o agrupamento *positivo*, 0,56 [0,52, 0,59]. Esses resultados corroboram os obtidos no TTR, especialmente na diferença entre o agrupamento *Não sei dizer* e os demais em termos de diversidade. Vale ressaltar que o agrupamento *negativo* possui a menor média.

A Tabela 10 apresenta a representatividade das principais etiquetas pos em relação ao total de palavras etiquetadas de cada agrupamento de comentários. Para um maior aprofundamento, analisamos a média de palavras classificadas com etiquetas específicas por comentário, começando pelos adjetivos (ADJ). A média e o intervalo de confiança para cada agrupamento são os seguintes: o agrupamento *Não sei dizer* apresenta 0,93 [0,62, 1,29], o agrupamento *neutro*, 1,94 [1,58, 2,36], o agrupamento *negativo*, 2,41 [2,21, 2,62] e o agrupamento *positivo*, 2,38 [1,90, 2,98]. O agrupamento *Não sei dizer* possui a menor média. Ao aplicar o teste estatístico nos demais agrupamentos, observamos que há diferenças significativas entre todos eles, a partir de comparações entre *negativos* e *neutros*, *negativos* e *positivos*, e *positivos* e *neutros*.

Para os substantivos (NOUN), a média e o intervalo de confiança para cada agrupamento são os seguintes: o agrupamento *Não sei dizer* apresenta 1,96 [1,38, 2,65], o agrupamento *neutro*, 3,76 [3,24, 4,33], o agrupamento *negativo*, 4,681 [4,31, 5,06] e o agrupamento *positivo*, 4,61 [3,74, 5,66]. Como no caso dos adjetivos, o agrupamento *Não sei dizer* apresenta a menor média. O teste estatístico revela diferenças significativas ao confrontarmos o agrupamento *negativo* com o *positivo* e o agrupamento *negativo* com o *neutro*, porém, isso não se prova verdade ao confrontarmos o agrupamento *neutro* com o *positivo*.

A média e intervalo de confiança dos agrupamento para os verbos (VERB) são os seguintes: o agrupamento *Não sei dizer* apresenta 1,71 [1,09, 2,52], o agrupamento *neutro*, 2,99 [2,58, 3,43], o agrupamento *negativo*, 4,03 [3,68, 4,38] e o agrupamento *positivo*, 4,30 [3,45, 5,33]. Observa-se o mesmo padrão para o agrupamento *Não sei dizer* nas 3 etiquetas. O teste estatístico indica diferenças significativas ao compararmos o agrupamento *negativo* com o *positivo* e o agrupamento *negativo* com o *neutro*, mas não mostrou diferenças significativas ao compararmos o agrupamento *neutro* com o *positivo*.

Por fim, o agrupamento de comentários classificados como *Não sei dizer* é o único que apresenta mais etiquetas POS de nome próprio (PROPN) do que de adjetivos (ADJ), como pode ser visto na Tabela 10. Isso também mostra que esse agrupamento é o que contém mais nomes próprios.

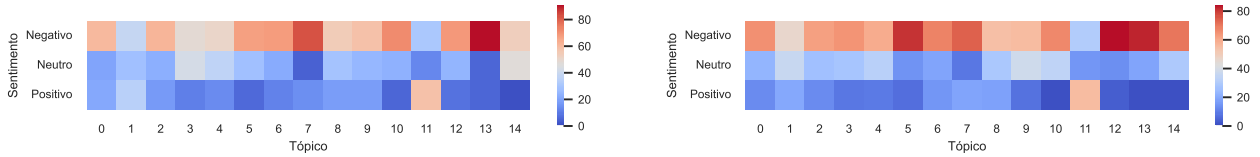
Reconhecimento de Entidades Nomeadas (REN): Os comentários classificados como *Não sei dizer* apresentam uma predominância de entidades do tipo PER, representando 51% das entidades identificadas, seguido por 19% de entidades do tipo LOC, 16% de ORG e 14% de MISC. Os comentários *neutros* exibem uma distribuição de 43% de entidades PER, 25% de LOC, 16% de ORG e 16% de MISC. Já os comentários *positivos* mostram 40% de entidades PER, 24% de LOC, 11% de ORG e 24% de MISC. Por fim, nos comentários *negativos*, 44% de entidades PER, 35% de LOC, 12% de ORG e 10% de MISC. Esses dados destacam a predominância de entidades PER no grupo *Não sei dizer*, a quantidade de entidades LOC nos comentários *negativos* e a presença significativa de entidades MISC nos comentários *positivos*.

Adicionalmente, considerando os 2000 comentários, nossas análises mostraram um crescimento no número de entidades mencionadas de janeiro para fevereiro e de fevereiro para março, em especial, possivelmente em decorrência da guerra entre Rússia e Ucrânia. Além disso, existem picos próximos de outubro, coincidindo com o período eleitoral no Brasil, com exceção do grupo *Não sei dizer*, provavelmente por ter poucos comentários, como demonstrado na Tabela 9.

Análise de n-gramas: Na análise dos n-gramas, podemos destacar os resultados de bigramas dos comentários classificados como *positivo*, que frequentemente abordam temas relacionados à vida. Já para os comentários *negativos*, evidencia-se *lula, bolsonaro*. Nos trigramas de sentimento positivo, temos palavras relacionadas à conselhos sobre relacionamento (por exemplo *sociedade, vê, casais*). os trigramas dos comentários *negativos*, há a ocorrência da combinação *bandido, bandido, morto*, possivelmente relacionada a debates políticos e posicionamentos ideológicos.

Extração de Tópicos (BERTopic): Realizamos a extração de tópicos, obtendo 15 tópicos correspondentes, ordenados por sua frequência de ocorrência entre os comentários, apresentados na Tabela 11. Analisando comentários em que se obteve discordância total entre os anotadores, os tópicos proporcionalmente mais relacionados são, em ordem decrescente: 14, 3, 1, 9 e 13. Enquanto os tópicos 3 e 1 são mais genéricos, relacionados a rotina, família e situações

[8]O teste Mann-Whitney é um teste não paramétrico utilizado para verificar se dois grupos de amostras independentes pertencem ou não à mesma população.



(a) Sentimentos rotulados por anotadores

(b) Sentimentos rotulados pelo modelo

Figura 1: Comparação de frequência de sentimentos para cada tópico entre anotadores e modelo.

do cotidiano, os tópicos 14, 9 e 13 são relacionados a política em diferentes escopos: o tópico 14 é mais direcionado a ideologias políticas, sobretudo o nazismo; o tópico 9 relaciona-se a ideia de *fake news*, resultado de eleições e partidos políticos, e o tópico 13 relata temas acerca de problemáticas e temas do governo durante a presidência de Jair Bolsonaro.

Já em relação aos comentários em que houve concordância total entre anotadores, destacam-se os tópicos 11, 12, 7 e 2. Pelo conjunto de palavras, em geral, trata-se de tópicos polarizados e que transmitem ideias positivas (tópico 11) ou negativas (tópicos 7,2), além do tópico 12, que apresenta críticas a governos brasileiros.

Em relação à análise de tópicos, realizamos uma comparação entre os dados anotados por humanos e pelo modelo XLM-RoBERTA, como pode ser visualizado na Figura 1. Identificamos que os tópicos 5, 12 e 14, relacionados a questões políticas, foram rotulados como *negativos* mais pelo modelo do que pela anotação humana. Já o tópico 7, composto por palavrões e palavras que expressam conceitos negativos no geral, como *odeio*, *pena*, *horrível*, apresentou maior proporção desse rótulo entre anotadores, comparado ao modelo. Os tópicos 3 (relacionados a conversas gerais sobre família e rotina) e 4 (sobre questões financeiras e mercado de trabalho), são razoavelmente polarizados entre os anotadores, mas o modelo rotulou mais como *negativos*.

Em relação aos comentários em que ao menos um anotador rotulou como *Não sei dizer*, destacam-se os tópicos 10, 14 e 2. O tópico 10 apresenta 32,6% de seus comentários em que ao menos um anotador rotulou como *Não sei dizer*, e apresenta conteúdos relacionados à crimes, xingamentos e conteúdos sexuais. Já o tópico 14, em que 30,4% de seus comentários apresenta ao menos um rótulo para *Não sei dizer*, relaciona-se à questões ideológicas e políticas. Por fim, o tópico 2, que apresenta 28% de seus comentários com ao menos um anotador indicando o rótulo, apresenta conteúdos genéricos relacionados a gírias e relatos do cotidiano. Em relação à comentários em que todos os anotadores assinalaram como *Não sei dizer*, destacam-se 3 comentários nos tópicos 2 e 3, geralmente relacionados a relatos e fatos do cotidiano e gírias. Possivelmente por apresentarem contextos muito específicos dentro de uma postagem, são considerados mais difíceis de rotular.

Rotulações semânticas (PyMUSAS): Para os resultados obtidos a partir da categorização semântica dos comentários, obteve-se, no total, 163.704 rotulações para níveis semânticos gerais(principais categorias semânticas, que excluem pontuações, por exemplo), lembrando que cada palavra de cada comentário apresenta uma ou mais rotulações possíveis dentro do domínio das categorias do USAS[9].

[9]https://ucrel.lancs.ac.uk/usas/Lancaster_visual/Frames_Lancaster.htm

Tabela 11: Tópicos e termos mais frequentes.

Tópico	Termos mais Frequentes
0	pessoa, pessoas, ficar, nada, fazer, aí, coisa, ainda, vida, porque
1	carro, acho, nunca, vou, uso, desse, lembro, sei, ver, achei
2	burro, bozo, ai, and, of, p*ca, vem, comida, pode, comentário
3	nome, filho, criança, banho, banheiro, tomar, p*ta, durante, lembro, deve
4	dinheiro, pagar, salário, fazer, trabalho, mercado, todos, ganhar, história, sobre
5	brasil, país, estado, eua, direita, países, rússia, china, nuclear, esquerda
6	time, goleiro, jogo, gol, futebol, palmeiras, jogador, vasco, paulo, passado
7	f*da, odeio, mano, tô, p*rra, pqp, tomara, gosto, pena, horrível
8	palavras, entender, dia, 11, pois, pessoas, falando, palavra, países, comecei
9	falou, entendi, falei, resultado, fake, hoje, disse, pt, pesquisa, dia
10	bandido, quer, bunda, p*u, pq, mãos, matar, passou, cima, bola
11	obrigado, sorte, entendi, comentários, man, respeito, espero, deus, feliz, boa
12	lula, bolsonaro, bolsonarista, governo, auxílio, gastos, mal, época, presidente, contra
13	população, política, direito, popular, governo, político, saúde, economia, bolsonaro, passar
14	socialismo, amp, x200b, hitler, nacional, comunismo, alemães, contrário, dizem, igreja

Considerando todos os comentários, as categorias de maior ocorrência são *nomes próprios, gírias e palavrões*, que compõe 29,64% das ocorrências totais, *termos abstratos, que abrangem ações gerais, afeto, classificação, avaliação, comparação, posse, importância, facilidade/dificuldade, grau, exclusividade e segurança*, que apresenta 17,6%, e *termos sociais, que abrangem ações, estados e processos, reciprocidade, participação, merecimento, traços de personalidade, pessoas, relacionamentos , família, grupos, obrigação, poder*, que apresenta 9,17% do total das ocorrências categorizadas.

Considerando apenas as anotações de sentimentos, observa-se grande relevância das categorias *termos numéricos* e *julgamentos de aparência e atributos físicos, como aparência, cor, forma, textura e temperatura* na composição de comentários rotulados como *positivos* pelos anotadores. Nesse caso, a segunda categoria compõe 6,7% de todas as rotulações de classes para comentários *positivos*, contra 1,9% em *negativos*, e 2,8% em *neutros*.

Em contrapartida, para os comentários rotulados como *negativos*, destacam-se as categorias *conceitos de movimento, localização, viagem e transporte*, bem como *conceitos de clima e questões ambientais*. A primeira categoria constitui 9,5% de todas as ocorrências categorizadas em comentários *negativos*, contra 3,3% para *positivos* e 5,0% para *neutros*. Para comentários rotulados como *neutros*, destacam-se, em relação às proporções de comentários *negativos* e *positivos*, as categorias *conceitos de ciência e tecnologia*, *conceitos de dinheiro, negócios, trabalho e indústria*, assim como *termos abstratos, que abrangem ações gerais, afeto, classificação, avaliação, comparação, posse, importância, facilidade/dificuldade, grau, exclusividade e segurança*.

Além disso, a categoria composta por *nomes próprios, gírias e palavrões* constitui parte considerável tanto de comentários *positivos*(28,65% do total de comentários positivos) quanto para *negativos*(29,9%). Assim, conceitos como gírias, palavrões e nomes próprios podem não ser considerados características predominantes para determinar o sentimento de um comentário, visto que para ambos sentimentos, tais conceitos apresentam presença similar. Essa questão é abordada ao comparar a anotação humana com o modelo, que classifica mais tópicos negativamente se forem constituídos por alguns palavrões, acima da média da rotulação humana para negativos. Logo, tais padrões semânticos podem levar o modelo a rotular comentários como *negativos* excessivamente, devido à dificuldade de tratar tais padrões no texto.

Para comentários em que houve discordância total entre anotadores, temos as categorias *arquitetura, tipos de edifícios e casas, construções, residência, móveis e acessórios domésticos*, *conceitos de dinheiro, negócios, trabalho e indústria* e *entretenimento em geral, música, teatro, esportes e jogos*. Considerando o total de ocorrências da categoria *arquitetura, tipos de edifícios e casas,construções, residência, móveis e acessórios domésticos* para todos os comentários, 12,38% deles participam da discordância total. Para as categorias *conceitos de dinheiro, negócios, trabalho e indústria* e *entretenimento em geral, música, teatro, esportes e jogos*, as porcentagens são 10,37% e 10,10%, respectivamente. Tais categorias compõem as proporções mais altas de discordância total entre todas as categorias. Esses resultados indicam certa dificuldade de concordância de anotações em relação a assuntos específicos que envolvem conhecimento de mundo do anotador, como arquitetura, entretenimento e mercado financeiro, por exemplo.

Por fim, a análise de categorias predominantes nos comentários que os anotadores rotularam com *Não sei dizer*mostra predominância das categorias *conceitos artísticos, artes, artesanato, alimentos, bebidas, tabaco e drogas, agricultura e horticultura* e *educação e estudos*.

5 CONCLUSÕES E TRABALHOS FUTUROS

Os achados da nossa pesquisa corroboram apontamentos na literatura sobre desenvolvimento de conjunto de dados por meio de anotação humana em tarefas que envolvem grande subjetividade, como é a análise de sentimentos. Um deles diz respeito à concordância entre anotadores, que, em nosso estudo, se mostrou moderada de acordo com os resultados do Alfa de Krippendorf e do Kappa de Fleiss.

Também em relação à composição do conjunto de dados, nossos resultados mostram que quase metade do conjunto de comentários foi majoritariamente rotulado como *negativo*, indicando desbalanceamento de classes considerável, podendo evidenciar um ambiente mais nocivo de interações.

Ainda em relação aos resultados das métricas de concordância, as anotações obtiveram valores próximos, apontando para concordâncias médias e moderadas entre seus respectivos anotadores. Em relação à incerteza, apenas em 4,1% dos comentários, dois ou mais anotadores rotularam o mesmo comentário com *Não sei dizer*, o que revelou dificuldade maior em 2 ou mais anotadores caracterizarem incerteza de sentimento para um mesmo texto.

Na comparação com os anotadores humanos, o modelo obteve acurácia de 62,37% e Kappa de Cohen de 0,34, valores considerados fracos. O modelo rotulou 60,90% dos comentários como *negativos*, taxa consideravelmente superior à da anotação humana, que foi de 48,05%. De fato, determinados tópicos relacionados a questões políticas foram rotulados como *negativos* mais pelo modelo do que pela anotação humana. O modelo também mostrou dificuldade para identificar corretamente comentários *positivos*.

A caracterização da linguagem dos comentários revelou que o tamanho dos comentários categorizados como *negativos* e *positivos* tendeu a ser maior do que o tamanho dos comentários categorizados como *neutros* e aqueles categorizados como *Não sei dizer*. O tamanho do texto pode impactar a rotulação, uma vez que quanto maior o contexto, maior a chance de os rotuladores conseguirem fazer uma interpretação e atribuir um sentimento.

No que diz respeito às classes de palavra mais frequentes em cada tipo de sentimento, destacam-se os comentários classificados como *Não sei dizer*, que apresentaram, tanto predominância de entidades do tipo PER, bem como maior número de etiquetas da classe nome próprio (PROPN), o que pode sugerir que esses comentários demandam reconhecimento dessas entidades e, por consequência, conhecimento de mundo, para poder atribuir um sentimento, problema que parece ter sido enfrentado pelos anotadores.

A análise de tópicos revelou que para os comentários em que se obteve discordância total entre os anotadores, figura, em primeiro lugar, o tópico 14, que é mais direcionado a ideologias políticas. Este resultado pode ser interpretado em relação aos achados sobre o desempenho do modelo, que rotulou comentários de questões políticas como *negativos* em maior número do que os anotadores humanos. No que diz respeito aos comentários em que ao menos um anotador rotulou como *Não sei dizer*, sobressaíram tópicos relacionados a crimes, xingamentos e conteúdos sexuais e a questões ideológicas e políticas.

Em termos metodológicos, nosso estudo evidenciou que a qualidade das métricas melhorou consideravelmente ao se separar o conjunto de dados em dois subconjuntos e incluir apenas os comentários rotulados com sentimentos, desconsiderando a categoria *Não sei dizer*. O mesmo aconteceu com o cálculo do percentual de concordância total dos anotadores sobre um mesmo rótulo, que foi superior quando desconsiderada a categoria *Não sei dizer*.

Em consonância com a literatura, nosso estudo corrobora a complexidade da tarefa de criação de conjunto de dados, dado o desafio de se lidar com níveis de concordância moderados entre anotadores. Para a consolidação do conjunto de dados, o voto da maioria, ou a agregação das distintas respostas, é decisório para o rótulo único de referência que será adjudicado. Em tarefas que envolvem alto grau de subjetividade, como é o caso da análise de sentimentos, a decisão pela maioria reduz a representatividade das diversas opiniões passíveis de existir em uma população ainda maior. Nesse sentido, estudos recentes[10, 12] propõem uma mudança em direção a uma abordagem mais inclusiva de todas as perspectivas dos anotadores como alternativa à maioria enquanto referência ou *ground truth*. Em trabalhos futuros, pretendemos explorar a perspectivização de forma a mitigar o problema do nível de concordância entre anotadores.

Agradecimentos: Este trabalho foi parcialmente financiado pela FAPEMIG, CAPES e CNPq.

REFERÊNCIAS

[1] Rafael J. A. Almeida. 2018. LeIA - Léxico para Inferência Adaptada. https://github.com/rafjaa/LeIA.

[2] Jacob Amedie. 2015. The Impact of Social Media on Society. *Advanced Writing: Pop Culture Intersections* (2015). https://scholarcommons.scu.edu/engl_176/2/

[3] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 258–266.

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[5] Victoria Bobicev and Marina Sokolova. 2017. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In *Recent Advances in Natural Language Processing*. 97–102. https://doi.org/10.26615/978-954-452-049-6_015

[6] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. arXiv:2005.00547

[7] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[8] Joseph L. Fleiss. 1975. Measuring Agreement between Two Judges on the Presence or Absence of a Trait. *Biometrics* 31, 3 (1975), 651–659. http://www.jstor.org/stable/2529549

[9] E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. ASSIN: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*. 13–15.

[10] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 2591–2597. https://doi.org/10.18653/v1/2021.naacl-main.204

[11] Claudia Freitas, Paulo Rocha, and Eckhard Bick. 2008. A new world in Floresta Sintá(c)tica – the Portuguese treebank. *Calidoscópio* 6, 3 (2008), 142–148. https://doi.org/10.4013/cld.20083.03

[12] Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-Perspective Annotation of a Corpus of Irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13844–13857. https://doi.org/10.18653/v1/2023.acl-long.774

[13] Jeffrey E. F. Friedl. 2006. *Mastering regular expressions* (3 ed.). O'Reilly Media.

[14] Klaifer Garcia and Lilian Berton. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* 101 (2021), 107057. https://doi.org/10.1016/j.asoc.2020.107057

[15] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[16] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[17] Antônio Pereira De Souza Júnior, Pablo Cecilio, Felipe Viegas, Washington Cunha, Elisa Tuler De Albergaria, and Leonardo Chaves Dutra Da Rocha. 2022. Evaluating Topic Modeling Pre-processing Pipelines for Portuguese Texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (Curitiba, Brazil) *(WebMedia '22)*. Association for Computing Machinery, New York, NY, USA, 191–201. https://doi.org/10.1145/3539637.3557052

[18] Simon Kemp. 2024. Digital 2024 April Global Statshot Report. https://datareportal.com/reports/digital-2024-april-global-statshot Acessado em: 13/06/2024.

[19] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014). https://doi.org/10.1073/pnas.1320040111

[20] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

[21] Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic Content Detection in online social networks: a new dataset from Brazilian Reddit Communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. 472–482.

[22] M Martella, F Bert, G Colli, G Lo Moro, A Pagani, R Tatti, G Scaioli, and R Siliquini. 2021. Consequences of cyberaggression on Social Network on mental health of Italian adults. *European Journal of Public Health* 31 (2021). https://doi.org/10.1093/eurpub/ckab165.589

[23] Philip May. 2021. Machine translated multilingual STS benchmark dataset. https://github.com/PhilipMay/stsb-multi-mt

[24] Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743* (2023).

[25] NLTK. 2023. NLTK - Sample usage for tokenize. https://www.nltk.org/howto/tokenize.html Acessado em: 22/06/2024.

[26] NLTK. 2023. NLTK - stopwords. https://www.nltk.org/search.html?q=stopwords Acessado em: 24/06/2024.

[27] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2013), 151–175.

[28] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086* (2011).

[29] Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio, and Paul Rayson. 2015. Development of the Multilingual Semantic Annotation System. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Rada Mihalcea, Joyce Chai, and Anoop Sarkar (Eds.). Association for Computational Linguistics, Denver, Colorado, 1268–1274. https://doi.org/10.3115/v1/N15-1137

[30] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria De Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*. Pisa, Italy, 197–206. http://aclweb.org/anthology/W17-6523

[31] Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 406–412.

[32] Reddit. 2023. Transparency Report: July to December 2023. https://www.redditinc.com/policies/transparency-report-july-to-december-2023 Acessado em: 13/06/2024.

[33] Shabnoor Siddiqui and Tajinder Singh. 2016. Social Media its Impact with Positive and Negative Aspects. *International Journal of Computer Applications Technology and Research* 5 (2016), 71–75. https://jogamayadevicollege.ac.in/uploads/1586197536.pdf

[34] Scott Songlin Piao, Paul Edward Rayson, Dawn Archer, Francesca Bianchi, Carmen Dayrell, Mahmoud El-Haj, Ricardo-María Jiménez-Yáñez, Dawn Knight, Michal Křen, Laura Lofberg, et al. 2016. Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (Portorož, Slovenia, 23-28). European Language Resources Association (ELRA), Paris, France.

[35] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[36] Marlo Souza and Renata Vieira. 2012. Sentiment Analysis on Twitter Data for Portuguese Language. In *Computational Processing of the Portuguese Language*, Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 241–247.

[37] spaCy. 2023. Portuguese Models. https://spacy.io/models/pt. Acessado em: 22/06/2024.

[38] Ronald J. Tallarida and Rodney B. Murray. 1987. *Mann-Whitney Test*. Springer New York, New York, NY, 149–153. https://doi.org/10.1007/978-1-4612-4974-0_46

[39] X. 2024. DSA Transparency Report - April 2024. https://transparency.x.com/dsa-transparency-report.html Acessado em: 14/06/2024.

[40] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:1907.04307