Proposta recebida em Outubro 2022 e aceite para publicação em Março 2023.

Classificação da qualidade da argumentação em *tweets* no domínio da política brasileira

Argument Quality Assessment in Brazilian political tweets

Cássio Faria da Silva
Rede Gonzaga de Ensino Superior / Universidade Federal de São Carlos

Vânia Paula de Almeida Neris
Universidade Federal de São Carlos

Helena de Medeiros Caseli
Universidade Federal de São Carlos

Resumo

A argumentação é uma habilidade inerente à comunicação humana, tanto em situações orais quanto escritas. Argumentos bem fundamentados são importantes para amparar a tomada de decisões e aprendizado, assim como para a obtenção de conclusões amplamente aceitas. Como área de pesquisa, a argumentação é um campo multidisciplinar que estuda os processos de debate e raciocínio. Em linguística computacional, investigações têm sido realizadas para (i) identificar argumentos e suas unidades e (ii) gerar ou (iii) avaliar a qualidade dos argumentos. No entanto, a maioria dos trabalhos atuais se concentra na mineração de argumentos em textos formais em inglês. Neste artigo, foi avaliada a qualidade da argumentação em *tweets* de domínio político, escritos em português do Brasil, usando algoritmos tradicionais de aprendizado de máquina – como Regressão Logística, *K-Nearest Neighbors*, Árvores de Decisão, Máquinas de Vetores Suporte (SVM), Floresta Aleatória e *Naive Bayes* – e também um ajuste fino de dois modelos neurais (BERTimbau e RobertaTwitterBR). Além de trazer resultados práticos para a avaliação da qualidade da argumentação em um gênero textual desafiador, como o Twitter, e em um domínio controverso, como a política brasileira, este artigo também visa suprir a carência de trabalhos que avaliem automaticamente a qualidade dos argumentos em português. Dentre os algoritmos de classificação avaliados, o modelo obtido a partir do ajuste fino do BERTimbau apresentou os melhores resultados com uma precisão de 69,65% quando foram consideradas todas as classes e de 100,00% para as mensagens de alta qualidade de argumentação.

Palavras chave

avaliação da qualidade da argumentação, *tweet*, BERT, política brasileira

Abstract

Argumentation is an inherent skill in human communication, both in oral and written situations. Well-founded arguments are important to support decision-making and learning, as well as to reach widely accepted conclusions. As a research area, argumentation is a multidisciplinary field that studies the processes of debate and reasoning. In computational linguistics, investigations have been carried out to (i) identify arguments and their units and (ii) generate or (iii) evaluate the quality of arguments. However, most current work focuses on argument mining in formal English texts. In this article, we evaluated the quality of argumentation in political domain tweets, written in Brazilian Portuguese, using traditional machine learning algorithms – such as Logistic Regression, KNearest Neighbor, Decision Trees, Support Vector Machines (SVM), Random Forest and Naive Bayes – and also a fine-tuning of two neural models (BERTimbau and RobertaTwitterBR). In addition to bringing practical results for the assessment of argumentation quality in a challenging textual genre, such as Twitter, and in a controversial domain, such as Brazilian politics, this article also aims to fill in the lack of works that automatically assess the quality of arguments in Portuguese. Among the evaluated classification algorithms, the model obtained from the fine-tuning of BERTimbau presented the best results, with an accuracy of 69.65% when all classes were considered and 100.00% for messages with high quality of argumentation.

Keywords

argument quality assessment, tweet, BERT, Brazilian politics

DOI: 10.21814/lm.15.1.387

1. Introdução

Um argumento é uma afirmação (ou conclusão) acompanhada por um número arbitrário de premissas que justificam, fundamentam, apoiam, defendem ou explicam a afirmação (Potthast et al., 2019). Argumentos bem fundamentados são importantes para amparar a tomada de decisões e aprendizado, assim como para a obtenção de conclusões amplamente aceitas. A argumentação (capacidade de produzir argumentos) é uma habilidade inerente à comunicação humana tanto em situações orais quanto escritas. Como área de pesquisa, a argumentação é um campo multidisciplinar que estuda os processos de debate e raciocínio (Habernal & Gurevych, 2017). Para Eemeren & Grootendorst (2003), a argumentação consiste em uma ou mais sentenças nas quais várias premissas são apresentadas para sustentar uma conclusão. As sentenças que fazem parte da argumentação constituem uma expressão completa que visa convencer um interlocutor.

Em linguística, estuda-se a argumentação em textos em linguagem natural (Stab & Gurevych, 2017a). Na ciência da computação, a identificação ou avaliação automática da argumentação é estudada no campo da inteligência artificial (Bench-Capon & Dunne, 2007). Ao combinar essas duas áreas de pesquisa, em linguística computacional ou no processamento de linguagem natural (PLN), investigações têm sido realizadas para (i) identificar argumentos e suas unidades e (ii) gerar ou (iii) avaliar a qualidade de tais argumentos. Mais especificamente, as tarefas mais comumente investigadas são: a mineração de unidades de argumentação (Al-Khatib et al., 2016; Habernal & Gurevych, 2015, 2017), a detecção de evidências que apoiam reivindicações[1] (Rinott et al., 2015) e a identificação de relações argumentativas (Peldszus & Stede, 2015). Outros trabalhos classificaram esquemas de argumentação (Feng et al., 2014), realizam a análise de estruturas gerais de argumentação (Wachsmuth et al., 2015; Stab & Gurevych, 2017a) e geram reivindicações (Bilu & Slonim, 2016).

Algumas teorias ou dimensões da qualidade da argumentação foram avaliadas computacionalmente (Stab & Gurevych, 2017b; Wachsmuth et al., 2017d; Zhang et al., 2016). No entanto, de acordo com Wachsmuth et al. (2017b), ainda não se constituiu um conceito geral para a qualidade da argumentação ou uma definição clara de suas dimensões. Apesar da falta do conceito geral, tarefas relacionadas à argumentação computacional — como mineração, geração, identificação de argumentos e sua avaliação — têm se mostrado relevantes em atividades como apoio à escrita e assistência à discussão (Stab & Gurevych, 2017b; García-Gorrostieta et al., 2018).

No campo da comunicação, como bem pontuado por Lytos et al. (2019), a internet e as redes sociais são, hoje, o meio de comunicação mais utilizado. Consistem em espaços que permitem a emissão de opiniões sobre qualquer assunto e são fonte para a produção de um grande volume de textos com potencial argumentativo. De especial interesse para este trabalho é a rede social Twitter[2], bastante utilizada para troca de informações e opiniões sobre política no Brasil.

A argumentação no Twitter, além de envolver características específicas do gênero textual, também é permeada pelas necessidades de comunicação, pelo contexto histórico e pelo assunto da mensagem (Marcuschi et al., 2002), como exemplificado no exemplo (1).[3]

(1) @gleisi Prezada **coxa** vc não tem vergonha nessa cara reformada com dinheiro público**??** **A merda do seu partido** que se diz do povo não fez nada disso e ainda **enfiou dinheiro nosso no rabo** do **joesley** do **eike** e do **Marcelo**. **Esse discursinho vagabundo não cola mais**.

Entre os indícios linguísticos que impactam a qualidade da argumentação neste exemplo (1), podem ser citados: (i) referências pejorativas como "coxa"; (ii) a repetição de sinais de pontuação ("??") que indicam indignação; (iii) a presença de discurso de ódio contra um partido político em "a merda do seu partido" e contra a autora do *tweet* original (ao qual este é uma resposta) em "esse discursinho vagabundo"; (iv) expressões coloquiais como a expressão idiomática em "enfiou [...] no rabo" e gíria em "não cola mais"; (v) uma tentativa de tornar o argumento mais pessoal em "dinheiro nosso"; e (vi) contexto histórico citando "joesley" (Joesley Batista), "eike" (Eike Batista) e "Marcelo" (Marcelo Odebrecht), três empresários brasileiros que ficaram nacionalmente conhecidos em um dos escândalos políticos ocorridos no Brasil.

Ao lidar com textos do Twitter, escritos em português, esse trabalho enfrenta desafios não

[1] Reivindicações, no contexto dos trabalhos relacionados à mineração de unidades de argumentação, se referem a quaisquer declarações ou afirmações que possuam propósito argumentativo.

[2] https://twitter.com/

[3] Todos os exemplos de *tweets* apresentados neste artigo são transcritos exatamente como foram publicados pelos seus autores. O destaque (em negrito) de alguns trechos foi inserido pelos autores deste artigo para enfatizar.

presentes na maioria dos trabalhos da literatura, que focam na mineração de argumentos em textos formais em inglês. Em relação ao idioma, vale destacar que a língua portuguesa é uma das mais faladas no mundo, com mais de 280 milhões de falantes. Nesse sentido, são notáveis os avanços e a importância que tem sido dada atualmente ao desenvolvimento e aprimoramento dos recursos de PLN para o português. Isso pode estar relacionado, em grande parte, aos aspectos da globalização e da popularização do acesso ao ambiente online, que são fatores facilitadores do intercâmbio sociolinguístico dessa língua.[4]

Em relação ao gênero textual, é importante salientar que embora hajam recursos linguísticos valiosos para o português, a maioria foi desenvolvida para lidar com textos formais, bem escritos e autocontidos, como artigos de jornal. No entanto, o cenário atual da comunicação nas mídias sociais exige que os recursos de PLN sejam capazes de lidar com textos cheios de desafios, como a presença de gírias, linguagem figurada, erros de português e uso do "internetês" (vocabulário próprio composto por termos e abreviações usualmente encontrados em textos de mídias sociais). Nesse sentido, embora fenômenos linguísticos que trazem indícios de uma qualidade da argumentação boa ou ruim possam ser identificados nos *tweets*, não há garantia de que os recursos e as ferramentas disponíveis hoje para o processamento automático também sejam capazes de identificá-los.

No que diz respeito à avaliação da qualidade da argumentação, que é o foco deste artigo, desde o esquema argumentativo de Toulmin (2003), estudos têm sido realizados para simplificar a compreensão da estrutura e determinar a importância dos elementos argumentativos do texto. Wachsmuth et al. (2017b) propuseram uma taxonomia composta por três dimensões para avaliar a qualidade da argumentação: retórica, lógica e dialética. No entanto, desde então, poucos estudos (Potthast et al., 2019; Wachsmuth & Werner, 2020; Skitalinskaya et al., 2021) se dedicaram à avaliação automática da qualidade da argumentação, muito menos em gêneros textuais cujos textos apresentam conteúdos distantes da norma linguística padrão e longe da própria noção convencional de argumentação.

Entre os trabalhos mais recentes que exploraram essa temática pode-se citar o de Gretz et al. (2019), que aplicaram o modelo de dimensões de qualidade da argumentação proposto por Wachsmuth et al. (2017b) para avaliar a qualidade da argumentação no corpus IBM-Rank-30k, contendo 30.497 argumentos provenientes de clubes de debates, no idioma inglês, onde os participantes são incentivados a redigirem textos bem escritos e com argumentos de alta qualidade. Foram apresentadas avaliações com três métodos baseados em BERT: BERT-Vanilla, BERT-Finetune e BERT-FT$_{\text{TOPIC}}$.[5] Os experimentos evidenciaram um melhor desempenho com o BERT-FT$_{\text{TOPIC}}$, com um coeficiente de correlação de Pearson de 0,53, calculado em relação a uma fração superior e inferior dos argumentos que estão nos extremos da escala de qualidade da argumentação. De acordo com os autores, as dimensões de Relevância e Eficácia Globais são as mais indicativas para os índices gerais de qualidade. Outros trabalhos que usaram o BERT são os de (Fromm et al., 2019) e (Reimers et al., 2019), mas ambos para mineração de argumentos e não para a avaliação da qualidade da argumentação.

Considerando as lacunas identificadas nos estudos que focam na avaliação automática da qualidade da argumentação, esse trabalho é inédito no sentido de que é o primeiro a investigar métodos para avaliar automaticamente a qualidade da argumentação: (i) em postagens de redes sociais e (ii) em português. Assim, neste artigo investigou-se como utilizar o aprendizado de máquina para classificar a qualidade da argumentação em *tweets* de domínio político escritos em português.

As questões de pesquisa que se busca responder com este estudo são:

(Q1) É possível predizer automaticamente a qualidade da argumentação em *tweets* no domínio da política brasileira?

(Q2) Como identificar automaticamente os *tweets* no campo da política brasileira com bons argumentos?

Para tanto, foram usados algoritmos de Aprendizado de Máquina (AM) baseados em *features*, como regressão logística (LR), *K-Nearest Neighbor* (KNN), árvore de decisão (DT), máquinas de vetor de suporte (SVM), Flo-

[4]Para se ter uma ideia desse crescente interesse, o Centro Regional de Estudos para a Desenvolvimento da Sociedade da Informação (CETIC) observou o comportamento de brasileiros maiores de 16 anos na internet entre 23 de junho e 8 de julho de 2020. Seus achados mostraram que 49% dos internautas realizavam atividades laborais e 72% buscavam informações relacionadas à saúde na internet. Para mais informações, consulte: https://cetic.br/media/docs/publicacoes/2/20200817133735/painel_tic_covid19_1edicao_livro%20eletr%C3%B4nico.pdf

[5]Consiste em concatenar o tópico da discussão ao argumento, separando-os por um delimitador.

resta Aleatória (RF) e *Naive Bayes* (NB) — e também um ajuste fino de um modelo BERT e um RoBERTa. Os experimentos realizados com o BERT alcançaram uma precisão de 100% para as mensagens consideradas com Alta qualidade de argumentação. Enquanto os algoritmos baseados em *features* obtiveram precisões médias que variaram de 32% a 54%.

Este artigo está organizado em cinco seções, além desta introdução. Na seção 2, são apresentados os trabalhos mais relacionados a esta pesquisa. A seção 3 descreve brevemente o processo de construção do corpus e as pistas linguísticas propostas por Silva et al. (2021) e adotadas neste artigo para definir uma boa qualidade de argumentação em um *tweet* do domínio da política brasileira. Na seção 4, apresentam-se os experimentos realizados para avaliar a qualidade da argumentação. Os resultados de tais experimentos são descritos e analisados na seção 5. Por fim, na seção 6, são feitas algumas considerações finais, com a apresentação de limitações e apontamentos sobre trabalhos futuros.

2. Trabalhos relacionados

Avaliar a validade, a qualidade e a força dos argumentos representa um desafio inerente ao discurso argumentativo. Vale destacar que existem fundamentos teóricos e diversas teorias normativas para embasar a tarefa, tais como (i) o modelo argumentativo de Toulmin (2003); (ii) os esquemas e questões críticas de Walton & Walton (1989); (iii) o modelo ideal de argumentação crítica na abordagem pragma-dialética (Eemeren & Grootendorst, 1987), em que as falácias são consideradas movimentos incorretos em uma discussão cujo objetivo é a resolução bem sucedida de uma disputa; e (iv) o estudo das falácias (Boudry et al., 2015). No entanto, julgar os critérios qualitativos da argumentação cotidiana ainda representa um desafio para os estudiosos e profissionais da argumentação (Weltzer-Ward et al., 2009; Swanson et al., 2015; Rosenfeld & Kraus, 2016).

Apesar desses fundamentos teóricos, os métodos e técnicas já propostos para avaliar a qualidade dos argumentos não concordam sobre quais critérios devem ser considerados, nem mesmo sobre se a qualidade deve ser avaliada do ponto de vista teórico ou prático. Wachsmuth et al. (2017a) tentaram elucidar a questão de quão diferentes são as visões teóricas e práticas da qualidade da argumentação. Do ponto de vista teórico, apontam que a convicção é entendida como a principal qualidade lógica, e sustentam o fato de que a avaliação teórica da qualidade da argumentação permanece complexa. Eles também apontaram que as abordagens práticas indicam o que focar para simplificar a teoria, enquanto a teoria parece benéfica para orientar a avaliação da qualidade na prática.

Na mesma direção, outros estudos têm buscado avaliar a relevância dos argumentos por meio da identificação de sentenças argumentativas com a posterior avaliação da importância/relevância delas. Potthast et al. (2019) avaliaram o grau de relevância de um conjunto de argumentos no corpus args.me[6] (Wachsmuth et al., 2017c), constituído por textos escritos em inglês, provenientes de cinco portais de debates, com o objetivo de construir um motor de busca de argumentos na web. Quarenta anotadores avaliaram a relevância de cada um dos 437 argumentos relacionados a 40 tópicos selecionados, além de sua qualidade retórica, lógica e dialética. Dos 437 argumentos anotados, 208 foram marcados a favor e 195 foram marcados como contrários ao tópico em questão, além de 34 que foram anotados como não argumentativos. Pontuações de 1 (baixa) a 4 (alta) foram atribuídas às dimensões de qualidade de argumentação — retórica (*rethoric*), lógica (*logic*) e dialética (*dialectic*) — e à sua relevância (*relevance*). A distribuição das pontuações (de 1 a 4) produzidas nesse trabalho pode ser vista na Figura 1.

As pontuações de relevância indicam que muitos argumentos relevantes (classificados como 4) foram recuperados do corpus args.me. Outros trabalhos também investigaram o aspecto da relevância dos textos argumentativos (Wachsmuth et al., 2017d; Gleize et al., 2019).

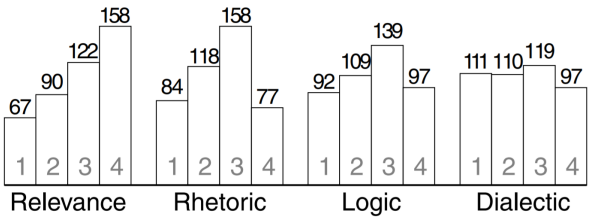

Figura 1: Distribuições de pontuação por dimensões de relevância e qualidade (Potthast et al., 2019).

Habernal & Gurevych (2016) sugeriram que a avaliação da qualidade da argumentação deve ser feita comparando argumentos, enquanto outros trabalhos (Persing & Ng, 2015; Wachsmuth et al., 2017b) relataram avaliações da qualidade dos argumentos individuais com resultados satisfatórios.

Vale mencionar que, neste trabalho, a quali-

[6]Disponível em: `https://www.args.me`

dade de argumentação é avaliada para um *tweet* isoladamente e não com base na comparação entre *tweets*.

Trabalhos mais recentes (Wachsmuth et al., 2017b; Lauscher et al., 2020; Wachsmuth & Werner, 2020; Skitalinskaya et al., 2021) têm utilizado uma taxonomia que visa avaliar aspectos individuais com base nas características da estrutura argumentativa, como o apelo emocional empregado, a organização da sentença e a credibilidade do autor da mensagem. A taxonomia escolhida para ser utilizada neste trabalho é apresentada na seção 2.3.

Embora passos importantes em AM tenham sido dados rumo ao processamento de *tweets*, como a detecção de argumentos, alegações e evidências, Schaefer & Stede (2021) demonstraram que várias áreas ainda estão em desenvolvimento, como o emprego de abordagens de AM utilizando aprendizado profundo e técnicas de redes neurais, além de pesquisas em outros idiomas diferentes do inglês.

Antes de detalhar a taxonomia adotada nesta pesquisa, as duas próximas subseções citam alguns trabalhos que investigaram a qualidade da argumentação em vários domínios, mais usuais do que o da política, aqui investigado.

2.1. Avaliação da qualidade da argumentação em redações de alunos

Fornecer aos estudantes *feedback* útil a respeito da persuasividade de seus argumentos tem sido objeto de estudos recentes em mineração de argumentos. Alguns corpora foram construídos, principalmente em inglês (Stab & Gurevych, 2014; Persing & Ng, 2015; Stab & Gurevych, 2017a; Carlile et al., 2018; Putra et al., 2021), e estudos têm sido realizados para avaliar a qualidade da argumentação em redações escritas por estudantes.

Stab & Gurevych (2017b) apontaram que as premissas de um argumento bem fundamentado devem fornecer evidências suficientes para aceitar ou rejeitar sua afirmação. Para chegar a essa conclusão, um corpus composto por 402 redações (Stab & Gurevych, 2017a) foi anotado com estruturas de argumentação. Os autores relataram uma pontuação de concordância de Fleiss entre os anotadores de 0,877, o que indicou que os anotadores foram capazes de identificar de forma confiável as principais reivindicações presentes em redações persuasivas. Para a construção do modelo computacional, os autores experimentaram máquinas de vetores de suporte (SVMs) e redes neurais convolucionais (CNNs) e alcançaram 84% de precisão na tarefa de identificar argumentos insuficientemente suportados. O corpus final e as diretrizes de anotação estão disponíveis.[7]

Carlile et al. (2018), com o objetivo de realizar futuras avaliações no nível de persuasão dos argumentos em redações de alunos, construíram um corpus composto por 102 redações selecionadas aleatoriamente do *Argument Annotated Essays* (Stab & Gurevych, 2014). O corpus foi anotado com árvores de argumentos, pontuações de persuasão e atributos de componentes de argumentos que, segundo os autores, impactam essas pontuações. Os autores relatam uma concordância que variou de 0,549 a 1,000 (valores de α de Krippendorff (2011)), de acordo com o atributo. O corpus anotado e as diretrizes de anotação estão disponíveis.[8] Na mesma linha de pesquisa, outros trabalhos avaliaram a qualidade da argumentação em teses e trabalhos acadêmicos (García-Gorrostieta et al., 2018; García-Gorrostieta & López-López, 2018).

Putra et al. (2021) criaram um corpus com anotação estrutural argumentativa para redações escritas em inglês como língua estrangeira e também definiram um esquema de anotações. O corpus anotado produzido como resultado desse esforço, o ICNALE-AS, inclui 434 redações enviadas por estudantes de inglês de várias nações asiáticas. A análise de concordância inter-anotador mostrou que o esquema de anotação proposto é estável, alcançando um coeficiente de Cohen de 0,66.

Apesar dos trabalhos apresentados nesta seção terem avaliado a qualidade da argumentação, eles o fizeram para um gênero textual formal e menos desafiador do que o Twitter. Os textos formais admitem forma e estilo e, como consequência, podemos presumir que apresentam uma boa argumentação com argumentos encadeados baseados em fatos reais e frases estruturadas. Todos esses aspectos podem não ocorrer em argumentos feitos em postagens no Twitter; e, mesmo quando ocorrem, devem ser ressignificados de acordo com esse gênero e com a maneira como os usuários das mídias sociais utilizam esses aspectos, o que resulta, por exemplo, em uma possível argumentação baseada em *fake news*.

[7] Disponível em: https://www.ukp.tu-darmstadt.de/data/

[8] Disponível em http://www.hlt.utdallas.edu/~zixuan/EssayScoring

2.2. Avaliação da qualidade da argumentação em mensagens de fóruns e portais de discussão

Para investigar a identificação de postagens persuasivas em fóruns de discussão (como Change My View[9]), Wei et al. (2016) criaram um corpus de mensagens argumentativas selecionando tópicos com mais de 100 comentários publicados entre janeiro de 2014 e janeiro de 2015, totalizando 1.785 tópicos com 374.472 comentários.

Experimentos foram realizados para encontrar quais *features* seriam as mais adequadas para predizer comentários persuasivos. Entre as *features* investigadas estavam: (i) o número de palavras e frases, (ii) a presença/ausência de pontuação, (iii) as *part-of-speech tags* (POS tags), entre outras. Depois disso, os autores calcularam a correlação entre a pontuação humana de um comentário argumentativo e o conjunto de atributos da reputação do autor da postagem. Então, para a tarefa de classificação de comentários, três conjuntos de *features* foram avaliados, incluindo as que consideravam apenas as características superficiais do texto, aquelas que consideravam a interação social e aquelas focadas na argumentação propriamente dita. Os resultados experimentais mostraram que as *features* baseadas em argumentação são mais informativas no estágio inicial da discussão e que a eficácia das *features* de interação social aumenta com o número de comentários na discussão.

Habernal & Gurevych (2016) investigaram a comparação qualitativa entre pares de argumentos: dados dois argumentos (como mostrado na Figura 2), um deles deve ser selecionado como o mais convincente. A pesquisa produziu os seguintes resultados: (i) um corpus anotado composto por 16.000 pares de argumentos, escritos em inglês, (ii) a análise dos dados anotados em relação às propriedades definidas como convincentes, e (iii) modelos computacionais gerados com SVM e uma arquitetura neural bidirecional de memória de longo prazo (BLSTM). O modelo SVM superou o modelo BLSTM (78% versus 76% de precisão, respectivamente) com uma diferença sutil, mas significativa, segundo os autores. Os—dados anotados e os códigos estão disponíveis publicamente[10].

Wachsmuth & Werner (2020) investigaram a avaliação automática de argumentos extraídos de portais de debate usando a taxonomia proposta por Wachsmuth et al. (2017b). Para tanto, com-

Argument 1	Argument 2
physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!	YES, because some children don't understand anything excect physical education especially rich children of rich parents.

Figura 2: Exemplo de argumentos sobre a obrigatoriedade da educação física (Habernal & Gurevych, 2016).

binaram *features* textuais (como mostrado nos exemplos da Figura 3) e SVM. Os exemplos na Figura 3 demonstram algumas *features* textuais[11] que, segundo os autores, podem ser preditivas de certas dimensões. Os autores relataram que o tamanho limitado do corpus dificultou a adoção de recursos mais complexos para a avaliação da qualidade. No entanto, eles destacaram que modelar a subjetividade por meio de *features* textuais pode ser propício para avaliar as dimensões lógica e dialética. Quanto à dimensão retórica, apontaram três aspectos difíceis: clareza, credibilidade e apelo emocional.

Estudos mais recentes (Fromm et al., 2022; Skitalinskaya et al., 2021; Toledo et al., 2019) usaram modelos BERT (Devlin et al., 2019) com ajuste fino (*fine-tuning*) em diferentes conjuntos de dados para avaliar a qualidade da argumentação. Skitalinskaya et al. (2021) investigaram a avaliação da qualidade da argumentação, independentemente dos aspectos discutidos. Para tanto, um corpus de 377.000 pares de comentário e argumento foi gerado a partir do fórum de discussão `kialo.com`,[12] abrangendo diversos temas de política, ética, entretenimento e outros. Duas tarefas foram realizadas: (i) avaliar qual afirmação de um par de comentários é melhor e (ii) classificar todas as versões de uma afirmação por qualidade. Os experimentos com regressão logística baseada em *embeddings* e redes neurais baseadas em *transformers* mostraram resultados promissores, sugerindo que os indicadores aprendidos generalizam bem entre os tópicos.

Para a primeira tarefa, os experimentos conduzidos com Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) apresentaram uma precisão de até 77,7%. Na segunda tarefa, o modelo baseado em SBERT superou todas as abordagens

[9]Disponível em: `https://www.reddit.com/r/changemyview`

[10]Disponível em: `https://github.com/UKPLab/acl2016-convincing-arguments`

[11]As *features* textuais apresentadas na pesquisa de Wachsmuth & Werner (2020) são semelhantes às pistas linguísticas (seção 3.1) estudadas neste trabalho.

[12]Disponível em `https://kialo.com`

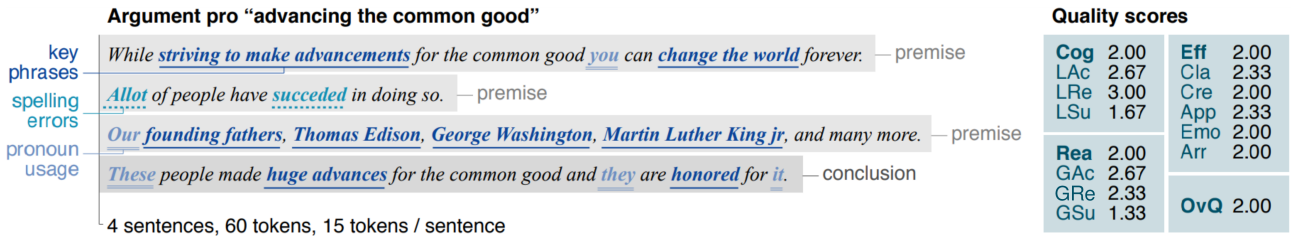

Figura 3: Exemplo de um argumento e os recursos linguísticos que afetam sua qualidade (Wachsmuth & Werner, 2020).

testadas alcançando até 0,73 em correlação de Pearson e 0,72 em correlação de Spearman.

Esses trabalhos investigaram a avaliação da argumentação em mensagens de fóruns de discussão e portais de debate (Wei et al., 2016; Habernal & Gurevych, 2016) e redações de alunos (Stab & Gurevych, 2017b; Carlile et al., 2018; Wachsmuth et al., 2016). No entanto, os *tweets* no domínio da política brasileira, nos dias atuais, possuem características mais desafiadoras do que as encontradas em fóruns e portais de debate, como: (i) um número muito limitado de caracteres, o que dificulta o uso de estratégias de argumentação linguística; e (ii) a presença de discurso incivil e intolerante (Rossini, 2019, 2022), decorrente da polarização e agressividade presentes no cenário atual da política brasileira, que traz a necessidade de estratégias para identificar discurso de ódio e polaridade, por exemplo.

2.3. Taxonomia de Wachsmuth et al.

Wachsmuth et al. (2017b) conduziram uma pesquisa sobre a qualidade da argumentação considerando tanto a teoria da argumentação quanto as perspectivas de mineração de argumentos. Com base nesse estudo, foi proposta a Taxonomia da Qualidade da Argumentação, cujas dimensões são utilizadas para definir a "qualidade". A Figura 4 ilustra esta taxonomia, com todas as dimensões dela.

De acordo com essa taxonomia, a qualidade da argumentação pode ser dividida nas dimensões lógica, retórica e dialética (Blair, 2012), descritas a seguir:

- A **dimensão lógica** refere-se à estrutura e composição de um argumento. Um argumento de alta qualidade lógica é baseado em premissas aceitáveis e as combina de forma convincente para apoiar a afirmação do argumento. Está relacionado com a irrefutabilidade lógica do argumento.
- A **dimensão retórica**, ao contrário, inclui noções de eficácia persuasiva, linguagem correta, precisão e estilo. Um argumento de alta qualidade retórica é bem escrito e atraente para o público e está relacionado à eficácia retórica do argumento. Especificamente, um argumento é retoricamente eficaz se for capaz de convencer o público-alvo (ou corroborar a concordância) que a posição do autor sobre o assunto é a correta.
- A **dimensão dialética** captura a contribuição de um argumento para o discurso. Um argumento de alta qualidade dialética é útil para apoiar a tomada de decisão cooperativa ou para resolver conflitos. O argumento é razoável se for capaz de contribuir para a resolução do problema de uma forma que seja suficientemente aceitável pelo público-alvo.

Wachsmuth et al. (2017b) testaram a taxonomia em um experimento de anotação com o Dagstuhl-15512-ArgQuality,[13] que contém 320 textos argumentativos com notas atribuídas por três anotadores que compõem os 15 aspectos da taxonomia. Nesse processo de anotação, cada texto foi primeiramente classificado como argumentativo ou não. Em seguida, para os textos argumentativos, todos os aspectos foram avaliados com notas 1 (baixo), 2 (médio) ou 3 (alto), além da opção "não posso julgar."

Na Figura 5, pode-se ver as pontuações atribuídas pelos três anotadores (A, B e C) em dois textos produzidos em resposta à pergunta "garrafas plásticas de água devem ser banidas?". O valor mais alto em cada coluna está marcado em negrito. A linha inferior representa a maioria dos votos dos três anotadores.[14]

A Figura 6 mostra os resultados deste experimento de anotação para os 304 textos do corpus

[13]Corpus Dagstuhl-15512-ArgQuality disponível em: http://arguana.com/

[14]A dimensão lógica mede a convicção (Co) e é composta por 3 aspectos: aceitabilidade local (LA), relevância local (LR) e suficiência local (LS). A dimensão retórica mede a eficácia (Ef) e é composta por 5 aspectos: credibilidade (Cr), apelo emocional (Em), clareza (Cl), adequação (Ap) e organização (Ar). Por fim, a dimensão dialética mede a razoabilidade (Re) e é composta por 3 aspectos: aceitabilidade global (GA), relevância global (GR) e suficiência global (GS).

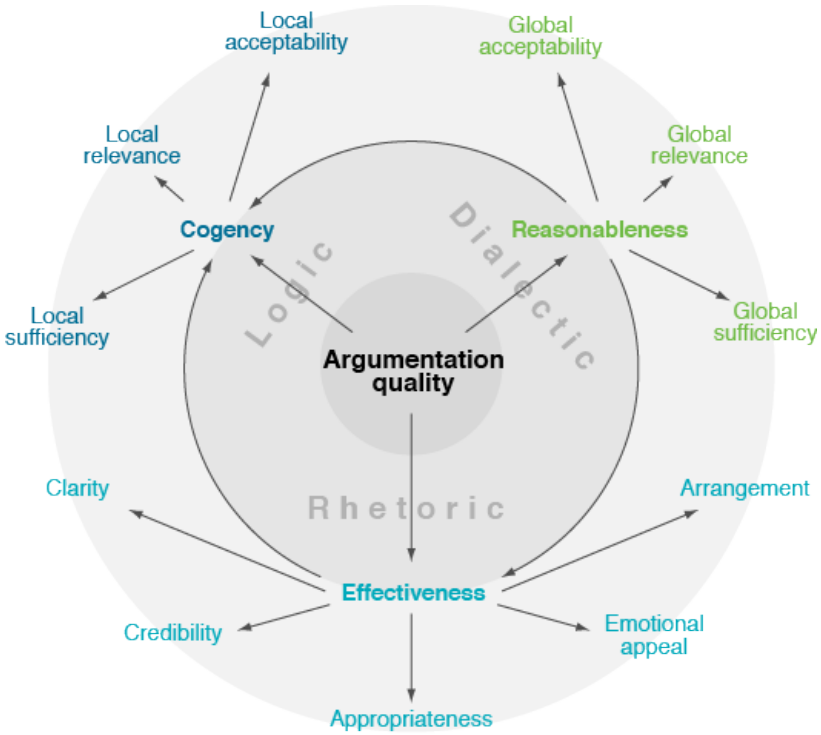

Figura 4: Taxonomia da Qualidade da Argumentação de Wachsmuth et al. (2017b).

Arguments	Pro Water bottles, good or bad? Many people believe plastic water bottles to be good. But the truth is water bottles are polluting land and unnecessary. Plastic water bottles should only be used in emergency purposes only. The water in those plastic are only filtered tap water. In an emergency situation like Katrina no one had access to tap water. In a situation like this water bottles are good because it provides the people in need. Other than that water bottles should not be legal because it pollutes the land and big companies get 1000% of the profit.														Con Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. (http://www.friendsjournal.org/bottled-water) (http://www.cdc.gov/healthywater/drinking/bottled/)													
Scores	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**
Annotator A	3	3	3	2	3	3	3	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	3	2	3	3	3	3	3	3	3	**3**
Annotator B	2	2	3	2	1	2	2	2	2	1	2	2	2	1	2	2	3	3	2	2	3	2	3	3	2	3	3	2	2	**3**
Annotator C	2	3	3	2	2	2	2	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	2	1	3	3	3	3	3	3	3	**3**
Majority score	2	3	3	2	2	2	2	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	3	2	3	3	3	3	3	3	3	**3**

Figura 5: Pontuações de cada anotador e pontuação majoritária para todas as dimensões de qualidade. Os argumentos são sobre o tópico "banir garrafas plásticas de água" (Wachsmuth et al., 2017b).

classificados como argumentativos por todos os anotadores: (a) a distribuição das pontuações majoritárias para cada dimensão; (b) o α de Krippendorff (2011) utilizado para medir a concordância entre anotadores; (c) a correlação para cada par de dimensões, calculada com base na média das correlações de todos os anotadores. O valor mais alto em cada coluna é destacado em negrito.

Nesta pesquisa foi escolhida a dimensão retórica da taxonomia de Wachsmuth et al. (2017b) para a avaliação da qualidade da argumentação em postagens do Twitter no domínio da política no Brasil. Assim como Gretz et al. (2019), foram selecionados os aspectos da taxonomia da qualidade da argumentação que possuíam características semelhantes às *features* linguísticas disponíveis. Desse modo, a escolha pela dimensão retórica se deu com base nas pistas linguísticas apontadas como relevantes pelos anotadores. Segundo Wachsmuth et al. (2017b), os aspectos que constituem a dimensão retórica estão relacionados ao apelo emocional aplicado na argumentação, ambiguidade, imprecisão, estilo de linguagem e organização da estrutura do texto. Portanto, entende-se que essas características podem ser, em certa medida, identificadas por meio de recursos linguísticos superficiais.

A dimensão retórica, segundo Wachsmuth et al. (2017b), possui cinco aspectos:

		(a) Maj. Scores			(b) Agreement			(c) Pearson Correlation Coefficients											
	Quality Dimension	**1**	**2**	**3**	α	**full**	**maj.**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**
Co	**Cogency**	150	131	23	.44	40.1%	91.8%		.64	.61	**.84**	**.81**	.46	.27	.41	.32	.55	.78	.64	**.71**	.70
LA	Local acceptability	84	169	51	.46	27.0%	90.8%	.64		.51	.53	.60	**.54**	.30	.40	.54	.46	.68	.75	.46	.45
LR	Local relevance	25	155	**124**	.47	32.6%	92.4%	.61	.51		.56	.56	.39	.27	.46	.35	.50	.62	.58	.68	.45
LS	Local sufficiency	172	119	13	.44	37.2%	92.8%	**.84**	.53	.56		.73	.39	.25	.37	.23	.51	.67	.51	.68	**.74**
Ef	**Effectiveness**	184	111	9	.45	42.1%	94.4%	.81	.60	.56	.73		.48	.31	.35	.34	.54	.75	.58	.66	.71
Cr	Credibility	99	199	6	.37	37.8%	95.7%	.46	.54	.39	.39	.48		**.37**	.32	.49	.37	.52	.52	.36	.40
Em	Emotional appeal	48	**235**	21	.26	42.8%	94.4%	.27	.30	.27	.25	.31	.37		.14	.30	.20	.30	.26	.26	.22
Cl	Clarity	42	191	71	.35	29.3%	89.8%	.41	.40	.46	.37	.35	.32	.14		.45	.56	.44	.45	.38	.27
Ap	Appropriateness	43	196	65	.36	17.4%	87.5%	.32	.54	.35	.23	.34	.49	.30	.45		.48	.47	.59	.20	.20
Ar	Arrangement	91	189	24	.39	26.6%	93.4%	.55	.46	.50	.51	.54	.37	.20	**.56**	.48		.55	.51	.49	.48
Re	**Reasonableness**	126	159	19	.50	41.4%	95.7%	.78	.68	.62	.67	.75	.52	.30	.44	.47	.55		**.78**	.65	.61
GA	Global acceptability	88	161	55	.44	31.6%	95.4%	.64	**.75**	.58	.51	.58	.52	.26	.45	**.59**	.51	.78		.46	.43
GR	Global relevance	69	167	68	.42	21.7%	90.1%	.71	.46	**.68**	.68	.66	.36	.26	.38	.20	.49	.65	.46		.61
GS	Global sufficiency	**231**	72	1	.27	**44.7%**	**98.0%**	.70	.45	.45	.74	.71	.40	.22	.27	.20	.48	.61	.43	.61	
Ov	**Overall quality**	152	128	24	**.51**	44.1%	94.4%	**.84**	.66	.61	.74	**.81**	.52	.30	.45	.42	**.59**	**.86**	.71	.70	.68

Figura 6: Resultados para os 304 textos do corpus classificados como argumentativos por todos os anotadores (Wachsmuth et al., 2017b).

1. **Credibilidade** (Cr): Credibilidade refere-se a como o autor transmite seus argumentos e os torna críveis. Segundo Wachsmuth et al. (2017b), um estilo apropriado em termos de escolha de palavras suporta a credibilidade. Além disso, de acordo com esses autores, aspectos que podem ser considerados para avaliar a credibilidade são: a honestidade do autor da mensagem, a polidez da linguagem utilizada ou o conhecimento e experiência do autor sobre os assuntos discutidos.
2. **Apelo emocional** (Em): O apelo emocional é considerado bem-sucedido em um argumento se ele cria emoções de tal forma que torna o público-alvo mais receptivo aos argumentos do autor.
3. **Clareza** (Cl): Clareza refere-se ao uso de uma linguagem gramaticalmente correta e amplamente inequívoca e que evita complexidade desnecessária e desvio do assunto discutido. A linguagem utilizada deve facilitar a compreensão e não deixar dúvidas sobre a posição do autor e a forma como ele a defende.
4. **Adequação** (Ap): A adequação de um argumento refere-se à linguagem (forma e conteúdo) utilizada para apoiar a criação de credibilidade e emoções, bem como a adequação ao assunto discutido.
5. **Organização** (Ar): Uma argumentação é considerada adequadamente organizada se apresentar a pergunta, os argumentos e a conclusão na ordem correta.

É importante destacar que o corpus utilizado no estudo de Wachsmuth et al. (2017b) é composto por mensagens que têm características diferentes do corpus adotado nesta pesquisa. Primeiramente, as mensagens de fóruns de discussão se caracterizam por serem mais longas do que as mensagens do Twitter, que têm um limite de 280 caracteres. O tamanho limitado das mensagens do Twitter pode impactar aspectos como Clareza e Organização. Outra diferença está relacionada ao contexto no qual as mensagens foram produzidas. No corpus de Wachsmuth et al. (2017b), as mensagens eram sobre tópicos gerais, às vezes controversos, mas sem a polarização política existente atualmente no Brasil, a qual pode impactar aspectos como Credibilidade, Apelo emocional e Adequação. Assim, embora a mesma taxonomia de Wachsmuth et al. (2017b) tenha sido adotada neste trabalho para embasar as medidas de qualidade da argumentação, o gênero textual e o domínio do corpus utilizado nesta pesquisa podem levar a resultados diferentes daqueles encontrados em (Wachsmuth et al., 2017b).

3. Corpus e anotação

Embora existam corpora abrangendo vários aspectos da análise argumentativa, alguns deles descritos na seção 2, até onde se sabe o corpus desenvolvido no projeto Arg Q!,[15] e descrito em (Silva et al., 2021), é o primeiro construído especificamente para a análise da qualidade da argumentação no domínio da política brasileira. Portanto, neste trabalho, foi utilizado o corpus construído e anotado conforme descrito por Silva et al. (2021). Este corpus é composto por *tweets* coletados como respostas a mensagens de parlamentares brasileiros postadas de 6 de março a 6 de abril de 2021.[16]

[15] Disponível em: `https://argq.org/`

[16] Embora os *tweets* dos parlamentares tenham sido considerados como semente para recuperar as respostas dos

Após a coleta dos *tweets*, cerca de 400 deles foram anotados com base nas pistas linguísticas descritas na Seção 3.1, conforme detalhado na Seção 3.2. Ressalta-se que apenas dados públicos foram coletados do Twitter e, embora os usuários não sejam identificados nos *tweets* do corpus, não podemos garantir que não seja possível rastrear a identidade do autor da mensagem.

3.1. Pistas linguísticas

Para a anotação do corpus foram utilizadas 30 pistas linguísticas definidas em (Silva et al., 2021): 4 para o aspecto de Clareza, 7 para o aspecto de Organização, 6 para o aspecto Credibilidade e 13 para o aspecto de Apelo emocional (6 para polaridade e 7 para intensidade). A Adequação não foi considerada em Silva et al. (2021) uma vez que se mostrou não relevante para a qualidade da argumentação em *tweets*, e também porque os anotadores apontaram que os critérios referentes à Adequação já eram contemplados pelos outros quatro aspectos. Portanto, o aspecto Adequação também foi desconsiderado neste trabalho. Nas seções seguintes, citamos as pistas linguísticas definidas para cada aspecto.

3.1.1. Clareza

Wachsmuth et al. (2017b) consideram um argumento claro se ele usa uma linguagem gramaticalmente correta e amplamente inequívoca e evita complexidade e desvios desnecessários da questão discutida. Além disso, a linguagem utilizada deve facilitar a compreensão e não deixar dúvidas sobre a posição do autor e a forma como ele a defende.

Nessa perspectiva, Silva et al. (2021) assumiram que todo argumento escrito em português tem o potencial de ser naturalmente claro, a menos que haja certos fenômenos linguísticos que interfiram negativamente na clareza. Dessa forma, a pontuação para o aspecto Clareza diminui com a presença de uma ou mais pistas linguísticas que prejudicam a clareza da argumentação, a saber: (i) questão que leva à dúvida, (ii) linguagem complexa desnecessária, (iii) presença de erros de língua portuguesa e (iv) desvio desnecessário do assunto principal. O aspecto Clareza foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (1).

$$\text{Clareza} = \begin{cases} \text{Baixa}, & \text{se } |Pistas| \geq 3 \\ \text{Média}, & \text{se } |Pistas| = 2 \\ \text{Alta}, & \text{caso contrário} \end{cases} \quad (1)$$

3.1.2. Organização

A definição de organização proposta por Wachsmuth et al. (2017b) considera que um texto deve ser composto de três partes, na seguinte ordem: (i) introdução do assunto, (ii) alguns argumentos para sustentar a conclusão e (iii) a conclusão. Essa definição, no entanto, não se aplica quando analisam-se textos do Twitter, pois os usuários não têm espaço suficiente para desenvolver essas 3 partes devido à limitação de caracteres máximos que uma mensagem pode conter (que é de 280). Assim, Silva et al. (2021) redefiniram a noção de organização para este gênero textual considerando a presença de pistas linguísticas que pontuam positivamente para a organização de um argumento.

As sete pistas linguísticas utilizadas para avaliar o aspecto Organização estão relacionadas à presença de relações linguísticas bem conhecidas: (i) relação condicional; (ii) relação de concessão; (iii) oposição ou contraste; (iv) comparação entre duas ideias; (v) relação de causa e efeito, explicação ou propósito; (vi) encadeamento cronológico ou enumeração; e (vii) exemplificação ou interligação lógica.

Diferentemente do aspecto da Clareza, que se ancora no pressuposto de que todo argumento no Twitter tem o potencial de ser inerentemente claro, no aspecto Organização, considera-se que esse gênero textual tem o potencial intrínseco de ser desorganizado, a menos que ocorram determinados fenômenos linguísticos que contribuam positivamente para a organização. Dessa forma, o aspecto Organização foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (2).

$$\text{Organização} = \begin{cases} \text{Alta}, & \text{se } |Pistas| \geq 2 \\ \text{Média}, & \text{se } |Pistas| = 1 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (2)$$

3.1.3. Credibilidade

De acordo com Wachsmuth et al. (2017b), um argumento deve ser avaliado como bem-sucedido em criar credibilidade se transmitir argumentos e outras informações de uma forma que torne o

seguidores, vale ressaltar que a avaliação da qualidade da argumentação foi feita apenas nas respostas dos seguidores.

autor crível, por exemplo, indicando a honestidade do autor, a polidez da linguagem utilizada ou revelando o conhecimento do autor ou a experiência em relação aos assuntos abordados.

Para a avaliação do aspecto Credibilidade, Silva et al. (2021) consideraram que um argumento escrito em português é verossímil se algumas pistas linguísticas estiverem presentes na superfície textual; ou seja, não levaram em consideração nenhum critério ou dado externo como a aceitação ou engajamento do autor nas redes sociais. Vale ressaltar que o Twitter é uma plataforma aberta e, portanto, qualquer usuário cadastrado pode postar mensagens diversas nesta plataforma de mídia social. Como não há pré-análise do perfil do usuário ou do conteúdo postado por ele, a dúvida sobre a credibilidade do que é postado é inerente à plataforma.

Assim, a pontuação do aspecto Credibilidade aumenta se as seguintes pistas estiverem presentes: (i) menção a uma data específica; (ii) menção a um fato midiático, histórico ou enciclopédico; (iii) menção a uma autoridade pública; (iv) presença de uma *hashtag* que reforça uma posição; (v) presença de um termo especializado; e (vi) um relato de alguma experiência pessoal ou individual. O aspecto Credibilidade foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (3).

$$\text{Credibilidade} = \begin{cases} \text{Alta}, & \text{se } |Pistas| \geq 3 \\ \text{Média}, & \text{se } |Pistas| = 2 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (3)$$

3.1.4. Apelo emocional

Segundo Wachsmuth et al. (2017b), um argumento tem apelo emocional quando cria emoções no interlocutor. Silva et al. (2021) decidiram dividir as pistas linguísticas para o aspecto apelo emocional em polaridade e intensidade.

Polaridade A polaridade de um argumento é considerada positiva ou negativa se contribui para criar emoções boas ou ruins no leitor, respectivamente. Se o seu conteúdo não causar nenhuma emoção ou se o apelo emocional estiver bem equilibrado entre positivo e negativo, então a polaridade do argumento é considerada neutra.

Assim, as pistas linguísticas de Silva et al. (2021) para uma polaridade positiva são: (i) a presença de uma referência cordial a uma pessoa/organização e (ii) o uso de linguagem polida. Por outro lado, as pistas linguísticas para uma polaridade negativa são: (i) a presença de uma referência pejorativa a uma pessoa/organização; (ii) o uso de xingamento ou palavra de baixo calão; (iii) a presença de discurso de ódio ou ameaça; e (iv) o uso de expressões que denotam especulação. Não há nenhuma pista linguística específica para uma polaridade neutra, pois é o meio termo entre os dois extremos. A Polaridade do aspecto Apelo emocional foi classificada com base na diferença entre a quantidade de pistas positivas (Pistas$^+$) e negativas (Pistas$^-$), como apresentado na equação (4).

$$\text{AE}_{pol} = \begin{cases} \text{Negativa}, & \text{se } \left|Pistas^+\right| - \left|Pistas^-\right| < 0 \\ \text{Positiva}, & \text{se } \left|Pistas^+\right| - \left|Pistas^-\right| > 0 \\ \text{Neutra}, & \text{caso contrário} \end{cases} \quad (4)$$

Intensidade O apelo emocional também é avaliado de acordo com sua intensidade, que é determinada pela quantidade de pistas linguísticas que potencializam a criação de emoções. Quanto maior o número de pistas linguísticas, maior a intensidade do apelo emocional. A intensidade é avaliada de acordo com a presença de (i) um pronome ou verbo na primeira pessoa; (ii) a repetição de sinais de pontuação; (iii) estrutura enfática (por exemplo, palavra em maiúscula); (iv) uma frase imperativa ou palavra de ordem; (v) uma expressão que denota exagero (por exemplo, "sempre", "nunca", "todo mundo"); (vi) linguagem não verbal (por exemplo, *emoticons*); e (vii) a presença de uma expressão idiomática, provérbio ou metáfora. A Intensidade do aspecto Apelo emocional foi classificada com base na quantidade de pistas de intensidade identificadas (Pistas), mas também em relação à quantidade de pistas positivas (Pistas$^+$) e negativas (Pistas$^-$) de polaridade identificadas, como apresentado na equação (5).

$$\text{AE}_{int} = \begin{cases} \text{Alta}, & \text{se } \left|Pistas^-\right| \geq 3 \\ & \text{ou } \left|Pistas^+\right| \geq 2 \\ & \text{ou } |Pistas| \geq 4 \\ \text{Média}, & \text{se } \left|Pistas^-\right| = 2 \\ & \text{ou } \left|Pistas^+\right| = 1 \\ & \text{ou } 2 \leq |Pistas| \leq 3 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (5)$$

3.1.5. Qualidade geral da argumentação

Após a anotação das pistas linguísticas associadas a cada aspecto, os anotadores em (Silva et al., 2021) definiram como eles seriam combinados para resultar em um valor para a Qualidade Geral da argumentação. Para tanto, o mapeamento de categorias para valores foi realizado como apresentado na tabela 1.

f		g	
Categoria	**Valor**	**Categoria**	**Valor**
Baixa	1	Negativa	-1
Média	2	Neutra	0
Alta	3	Positiva	1

Tabela 1: Funções f e g de mapeamento das categorias atribuídas pelos anotadores para valores.

A nota do Apelo emocional (N_{AE}) foi definida tal como se apresenta algoritmicamente de seguida.

if AE_{pol} =Neutra **then**
  $N_{AE} \leftarrow \frac{f(AE_{int})}{2}$
else
  $N_{AE} \leftarrow g(AE_{pol}) \times f(AE_{int})$
end if

E as notas dos demais aspectos foram definidas com base na aplicação da função f:

- $N_{Cla} = f(\text{Clareza})$
- $N_{Org} = f(\text{Organização})$
- $N_{Cred} = f(\text{Credibilidade})$

A partir das notas associadas a cada aspecto, a nota final foi dada pela soma das notas dos aspectos. Por fim, definiu-se que a nota para a qualidade geral da argumentação (N_{QG}) seria a média das notas dos anotadores e a categoria associada à Qualidade Geral foi definida como apresentado na equação (6).

$$QG = \begin{cases} \text{Alta}, & \text{se } N_{QG} \geq 7 \\ \text{Média}, & \text{se } 5 \leq N_{QG} < 7 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (6)$$

3.2. Anotação do corpus

Com base nas pistas linguísticas apresentadas anteriormente e seguindo as diretrizes estabelecidas em (Silva et al., 2021),[17] 400 *tweets* foram analisados por quatro pesquisadores em linguística computacional (3 com formação em linguística e 1 em computação). Desses, 48 *tweets* foram descartados por não terem sido considerados argumentativos[18] por todos os quatro anotadores. Assim, o corpus final é composto por 352 *tweets*.

Na Figura 7 é exibido um gráfico de distribuição de pontuação para cada aspecto dos 352 *tweets* argumentativos. A maioria deles tem alta Clareza e Organização, mas baixa Credibilidade. De fato, apenas 3% deles foram avaliados como de alta credibilidade. Com relação ao apelo emocional, 54% dos *tweets* foram avaliados como de polaridade negativa e 75% de média intensidade.

Na Tabela 2 (a) são exibidas as pontuações finais de cada aspecto para os 352 *tweets*. O grau de concordância entre anotadores foi calculado e é exibido na Tabela 2 (b) e expresso como o intervalo de α de Krippendorff (2011) (valor mais baixo - valor mais alto) que apresentaram os trios menos concordantes e mais concordantes de anotadores[19], e tanto os acordos totais quanto os majoritários.

A "Concordância total" é alcançada quando todos os anotadores concordam com a mesma pontuação, e "maioria" indica que pelo menos três anotadores concordaram. A concordância total encontrada ficou entre 27,84% e 57,67%, e a concordância majoritária dos anotadores ficou entre 69,89% e 86,93%. Com exceção do aspecto Clareza, todos os aspectos tiveram valores máximos de concordância acima de 0,40, diferentemente de Wachsmuth et al. (2017b) (veja Figura 6) onde os valores de concordância α para todos os aspectos ficaram abaixo de 0,40.

4. Experimentos

Para responder as questões de pesquisa definidas para este trabalho, foram realizados experimentos utilizando algoritmos de AM baseados em *features* e, também, uma abordagem neural baseada em transformers (BERT e RoBERTa).

[17] As diretrizes utilizadas para a anotação estão disponíveis no site do projeto Arg Q!: https://argq.org/

[18] Tradicionalmente, um texto é considerado argumentativo se contém argumentos organizados e estruturados em uma sequência lógica. No entanto, no que diz respeito ao gênero textual e domínio deste corpus, este conceito foi adaptado para abranger quaisquer *tweets* em que a posição/opinião do autor pudesse ser determinada.

[19] Decidimos relatar a concordância alcançada entre os trios para que nossos resultados pudessem ser comparados com os de Wachsmuth et al. (2017b) já que em seu trabalho havia três anotadores.

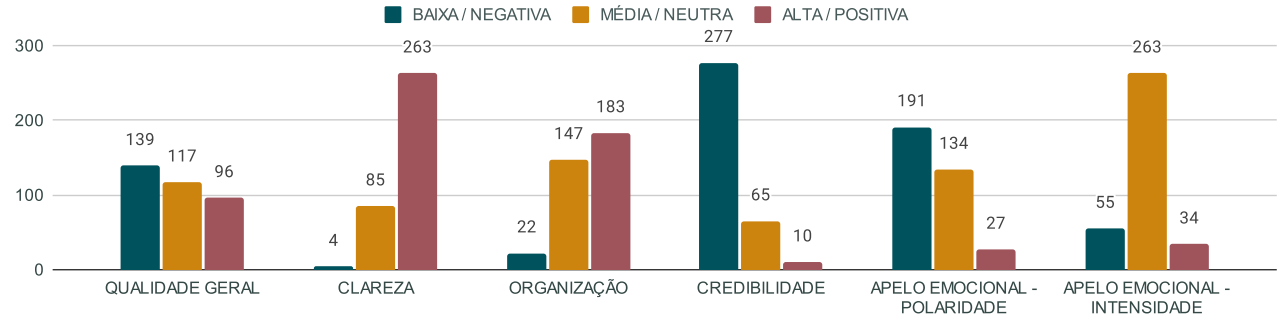

Figura 7: Distribuições de pontuação por aspectos de qualidade da argumentação no corpus de Silva et al. (2021) (tradução nossa).

Aspecto da qualidade	(a) Pontuação final			(b) Concordância		
	Baixo/ Negativo	Médio/ Neutro	Alto/ Positivo	α trios	total (4/4)	maioria (3-4/4)
Clareza	4	85	**263**	0,26 - 0,30	48,58%	79,26%
Organização	22	**147**	183	0,51 - 0,71	50,57%	82,67%
Credibilidade	**277**	65	10	0,36 - 0,48	57,67%	86,93%
Apelo emocional - Polaridade	**191**	134	27	0,60 - 0,66	51,99%	82,67%
Apelo emocional - Intensidade	55	**263**	34	0,48 - 0,55	40,63%	82,39%
Qualidade geral	139	117	96	0,50 - 0,54	27,84%	69,89%

Tabela 2: Pontuações para cada aspecto e a concordância entre os juízes humanos do corpus de Silva et al. (2021) (tradução nossa).

Para a abordagem neural, foram utilizadas apenas os *tweets* presentes no corpus, como descrito na seção 4.2. Entretanto, para a avaliação com os algoritmos de AM tradicionais, foi necessária a construção de um conjunto de *features* linguísticas para a tarefa de classificação, como detalhado na seção 4.1.

4.1. Experimentos com AM baseado em *features*

Para realizar os experimentos com AM tradicional foi necessário definir as *features* linguístico-computacionais (como descrito na seção 4.1.1) e treinar os modelos computacionais como descrito na seção 4.1.2.

4.1.1. Definição das features computacionais

Com o objetivo de encontrar as *features* linguístico-computacionais que melhor se correlacionam com as pistas linguísticas utilizadas para medir a qualidade da argumentação, foi definido e gerado um conjunto de 337 *features*.

Essas *features* foram categorizadas em 14 grupos, a maioria delas geradas pelo NILC-Metrix[20] (Leal et al., 2022):

1. **Medidas Psicolinguísticas** extraem características subjetivas do texto, tais como: imageabilidade, concretude, familiaridade e idade de aquisição. 44 *features* deste tipo foram geradas pelo NILC-Metrix.

   Imageabilidade envolve a facilidade e rapidez de evocar uma imagem mental associada a uma palavra;

   Concretude diz respeito ao grau em que uma palavra se refere a objetos, pessoas, lugares ou coisas que podem ser percebidas pelos sentidos;

   Familiaridade é o grau em que as pessoas conhecem e usam palavras em suas vidas cotidianas;

   Idade de Aquisição é uma estimativa da idade que a pessoa tinha pessoa quando uma palavra foi aprendida.

[20] NILC-Metrix (Leal et al., 2022) composto por 200 métricas linguísticas e psicolinguísticas utilizadas na avaliação de métodos de predição de complexidade textual. Disponível em: `http://fw.nilc.icmc.usp.br:23380/nilcmetrix`

2. **Informação morfossintática** engloba *features* relacionadas à classe gramatical de uma palavra e a sua função sintática no texto, como a proporção de adjetivos, advérbios, pronomes, substantivos, verbos e preposições, em relação à quantidade total de palavras no texto. Neste trabalho, para extrair essas *features* foram usados Apertium[21] (Armentano-Oller et al., 2006), NLPNet[22] (Fonseca & Rosa, 2013) e NILCMetrix. Foram geradas 94 *features* deste tipo.
3. **Complexidade Sintática** está relacionada à dificuldade de processamento de alguns tipos de estruturas de frases sofisticadas, por exemplo, a proporção de orações subordinadas reduzidas pela quantidade de orações do texto ou a proporção de orações na voz passiva analítica em relação à quantidade de orações do texto. 37 *features* deste tipo foram geradas pelo NILC-Metrix.
4. **Frequências de Palavras** mostram os valores das frequências absolutas e relativas das palavras no texto. 20 *features* deste tipo foram geradas pelo NILC-Metrix.
5. **Conectivos** incluem métricas relacionadas à quantidade de conectivos, operadores lógicos ou palavras que denotam negação em relação às palavras do texto. 18 *features* deste tipo foram geradas pelo NILC-Metrix.
6. **Simplicidade Textual** fornece métricas que medem o nível de complexidade em um texto com relação à dificuldade de compreensão de leitura. 9 *features* deste tipo foram geradas pelo NILC-Metrix.
7. **Índices de Legibilidade** incluem métricas que medem a legibilidade de um texto, como o índice Brunet e Flesch. 10 *features* deste tipo foram geradas pelo NILC-Metrix.
8. **Indicadores Temporais de um Léxico** incluem índices relacionados à diversidade de tempos verbais que ocorrem no texto. 14 *features* deste tipo foram geradas pelo NILC-Metrix.
9. **Informações Semânticas** referem-se a várias métricas que fornecem informações sobre o significado das palavras no texto, como a quantidade média de hiperônimos por verbo nas sentenças, ou a proporção de substantivos abstratos em relação à quantidade de palavras do texto. 18 *features* deste tipo foram geradas pelo NILC-Metrix.
10. **Medidas Descritivas** referem-se a métricas como a quantidade de parágrafos, frases e palavras em um texto e de sílabas por palavra. 16 *features* deste tipo foram geradas pelo NILC-Metrix.
11. **Coesão Semântica** é expressa por métricas que calculam as relações semânticas entre palavras em um texto, por exemplo a média de similaridade entre os pares de sentenças no texto, ou a média da entropia cruzadas das sentenças do texto. 19 *features* deste tipo foram geradas pelo NILC-Metrix.
12. **Polaridade do sentimento** mede a frequência de palavras com emoções positivas, negativas e neutras no texto. Neste trabalho, realizamos análise de polaridade com base em um modelo treinado usando o algoritmo de Floresta Aleatória e o corpus TweetSentBR[23] (Brum & Nunes, 2018) e também um modelo baseado em BERT de (Capellaro & Caseli, 2021). Foram geradas 4 *features* deste tipo.
13. **Linguagem Tóxica** indica se há ou não linguagem tóxica no texto. 7 *features* desse tipo foram geradas com um modelo baseado em BERT treinado no corpus ToLD-Br[24] (Leite et al., 2020), todas elas *features* binárias: não tóxico, LGBTQ+fobia, obsceno, insultuoso, racismo, misoginia e xenofobia.
14. A frequência de uso para **diferentes categorias de palavras** gerado com base na versão em português do LIWC[25] (Balage Filho et al., 2013). Foram geradas 62 *features* deste tipo.

4.1.2. Treinamento dos modelos de AM baseado em features

Este processo consiste em três etapas principais: **(i)** o pré-processamento, criação do conjunto de *features* e seleção das melhores *features*; **(ii)** ajuste de hiperparâmetros com validação cruzada aninhada; e **(iii)** melhor seleção de modelo e geração de métricas. Esta configuração experimental é ilustrada na Figura 8.

Para realizar o **primeiro passo** (i), cada *tweet* foi pré-processado pelo Enelvo[26] para remover informações desnecessárias (como o identificador do usuário do Twitter ao qual a men-

[21] Disponível em: https://www.apertium.org/
[22] Disponível em: http://nilc.icmc.usp.br/nlpnet/
[23] Disponível em: https://bitbucket.org/HBrum/tweetsentbr/src/master/
[24] Disponível em: https://github.com/JAugusto97/ToLD-Br
[25] Disponível em: http://143.107.183.175:21380/portlex/index.php/en/liwc
[26] Disponível em: https://github.com/thalesbertaglia/enelvo

sagem se referia como resposta) e normalizar palavras ruidosas em conteúdo gerado pelo usuário (por exemplo, substituir abreviações como "vc" por sua versão utilizada na norma culta "você"). Em seguida, os *tweets* normalizados foram processados pelas ferramentas linguístico-computacionais a fim de extrair as *features* descritas na seção 4.1.1.

Após a geração das 337 *features*, foi aplicado um método de seleção de *features* para determinar quais delas se correlacionam melhor com a qualidade geral da argumentação. O objetivo da seleção de *features* é encontrar aquelas que possivelmente são as melhores preditoras para tarefas de AM. Esse procedimento é importante, pois algumas *features* podem ser irrelevantes para a tarefa de AM e, desse modo, adicionar ruído ao modelo treinado. Adi et al. (2019) apontam que entre os problemas causados por conjuntos de dados de alta dimensão estão o alto custo computacional e a baixa precisão do modelo gerado. Para contornar esses problemas, sugere-se selecionar apenas as *features* mais relevantes que tenham uma alta correlação com a classe.

Alguns métodos de seleção de *features* podem ser aplicados para reduzir a alta dimensionalidade do conjunto de *features*, como Eliminação Recursiva de *Features* com Validação Cruzada (RFECV) (Misra & Yadav, 2020), Ganho de Informação (IG), Taxa de Ganho de Informação (Taxa de Ganho) (Adi et al., 2019) e Análise de Componentes Principais (PCA) (Maćkiewicz & Ratajczak, 1993). Neste trabalho, o método escolhido para a fase de seleção de *features* foi o RFECV dado que a literatura aponta a efetividade dele na eliminação de *features* irrelevantes por se tratar de um método que busca evitar o problema de *overfiting* com Eliminação Recursiva de *Features* (RFE), aplicando a validação cruzada estratificada (Misra & Yadav, 2020). O RFECV classifica as *features* com eliminação recursiva de *features* e validação cruzada de 10 vezes e, assim, seleciona o número ideal de *features* para construção de modelo.

Assim, as melhores *features* foram selecionadas com o auxílio da biblioteca scikit-learn usando RFE com Classificador de Floresta Aleatória e validação cruzada de 10 vezes. O conjunto final de *features* selecionadas para predizer a qualidade geral da argumentação contém 290 *features*. As 10 *features* mais significativas para a qualidade geral, de acordo com essa seleção automática, são apresentadas na Tabela 3. Contudo, como alguns dos algoritmos investigados nesta pesquisa são conhecidos por lidarem bem com um número elevado de *features*, como o SVM, também foram realizados experimentos sem esta etapa de seleção de *features*.

A **segunda etapa** (ii) consiste em refinar os parâmetros dos algoritmos de AM. Para isso, foi utilizado o GridSearchCV,[27] uma pesquisa exaustiva sobre valores de parâmetros especificados para um estimador. Conforme mostrado na Figura 8, as dobras de treinamento do laço externo são usadas no laço interno para ajustar os hiperparâmetros. O laço interno seleciona a melhor configuração de hiperparâmetros.

Cinco dobras estratificadas foram utilizadas em ambas os laços. O laço interno faz uma pesquisa de grade no espaço de hiperparâmetros que é validado de forma cruzada em relação aos conjuntos de treinamento e validação adquiridos pelo laço externo. A configuração do hiperparâmetro que maximiza a pontuação de precisão é retornada para cada pesquisa de grade. A generalização da configuração do modelo selecionado é então validada usando as métricas padrão de acurácia, precisão, cobertura e medida F nos conjuntos de teste criados pelo laço externo.

A **terceira etapa** (iii) consiste em gerar os melhores modelos de AM para cada algoritmo testado e classificar os *tweets* de teste usando os modelos treinados. Testamos seis modelos classificadores em nosso conjunto de dados: Regressão Logística (*Logistic Regression*, LR), *K-Nearest Neighbors* (KNN), Árvores de Decisão (*Decision Tree*, DT), Máquinas de Vetores de Suporte (*Support Vector Machines*, SVM), Floresta Aleatória (*Random Forest*, RF) e *Naive Bayes* (NB).[28]

4.2. Experimentos com BERTimbau e RoBERTaTwitterBR

Além dos experimentos descritos utilizando os algoritmos de AM baseado em *features*, foram realizados experimentos com o BERTimbau[29] (Souza et al., 2020) e o RobertaTwitterBR,[30], modelos

[27] Disponível em:https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[28] Os hiperparâmetros e respectivos valores explorados foram os seguintes: LR: *penalty*=l2, C=*np.power*(10., *np.arange*(-4, 4)), KNN: *n neighbors*=*list*(*range*(1, 10)), p=[1, 2], DT: *max depth*=*list*(*range*(1, 10), *criterion*=[*gini*, *entropy*], SVM: *kernel*=*rbf*, C=*np.power*(10., *np.arange*(-4, 4)), *gamma*=*np.power*(10., *np.arange*(-5, 0)); *kernel*=*linear*, C=*np.power*(10., *np.arange*(-4, 4)), RF: *n estimators*=[10, 100, 500, 1000, 10000], *criterion*=[*gini,entropy*], *max depth*=[10,11,12,13,14], NB: *var smoothing*= *np.logspace*(0,-9, num=100).

[29] Disponível em: https://github.com/neuralmind-ai/portuguese-bert

[30] https://huggingface.co/verissimomanoel/RobertaTwitterBR

ID	Grupo	Métrica	Descrição
1	Linguagem tóxica	toxiclang	Identifica a presença de linguagem tóxica no texto
2	NILCMetrix-Medidas Psicolinguísticas	idade_de_aquisição	Valores de idade média de aquisição de palavras de conteúdo de texto
3	NILCMetrix-Frequência de palavras	freq_bra	Média dos valores de frequência das palavras no texto na escala logarítmica Zipf via Corpus Brasileiro
4	NILCMetrix-Medidas Psicolinguísticas	imageabilidade_25_4_ratio	Proporção de palavras com valor de imageabilidade entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto
5	NILCMetrix-Medidas psicolinguísticas	imageabilidade_mean	Imageabilidade média das palavras de conteúdo no texto
6	NILCMetrix - Medidas Descritivas	syllabes_per_content_word	Número médio de sílabas por palavra de conteúdo no texto
7	TweetSentBR- Sentimento neutro	sent_neu	Proporção de palavras com emoção neutra em relação a todas as palavras do texto
8	NILCMetrix-Frequência de palavras	freq_brwac	Média dos valores das frequências das palavras do texto na escala logarítmica Zipf via BrWac
9	NILCMetrix-Medidas Psicolinguísticas	imageabilidade_4_55_ratio	Proporção de palavras com valor de imageabilidade entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
10	NILCMetrix-Medidas descritivas	sentence_length_standard_deviation	Desvio padrão do número de palavras por frase

Tabela 3: Top-10 *features* com melhor correlação com a qualidade geral da argumentação.

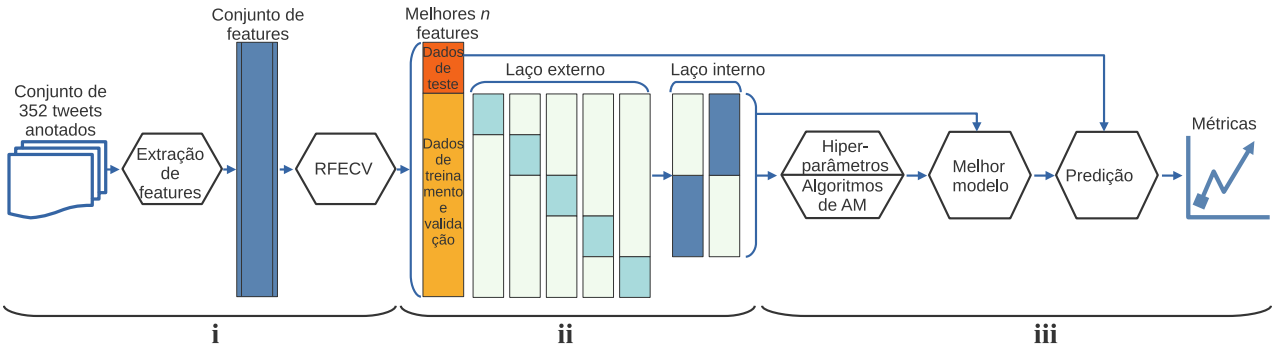

Figura 8: Configuração experimental adotada para a geração de modelos de AM baseados em *features*, dividida em: i. extração e seleção de *features*, ii. otimização com validação cruzada aninhada; e iii. treinamento e teste.

neurais treinados em português. O BERTimbau foi treinado no BrWaC (*Brazilian Web as Corpus*), um grande corpus em português, por 1 milhão de passos, usando *full word mask*. O RobertaTwitterBR foi treinado com aproximadamente 7 milhões de *tweets* em português.

Nestes experimentos, somente o texto das mensagens foi utilizado, ao contrário dos experimentos com os algoritmos de AM tradicional que utilizaram as *features* linguísticas. O único processamento realizado com as mensagens para estes experimentos foi a remoção dos nomes dos usuários. Um script Python foi usado para realizar essa tarefa.

A Figura 9 ilustra esse processo que foi realizado em duas etapas: (i) ajuste fino do modelo neural; e (ii) avaliação do modelo no conjunto de teste e geração das métricas. Vale ressaltar que as mesmas partições de treinamento e teste foram usadas tanto nos experimentos com AM baseado em *features* quanto nos experimentos com os modelos neurais, de forma estratificada.

5. Resultados

Neste artigo, investigou-se como o AM pode ser aplicado para classificar a qualidade da argumentação em *tweets* do domínio da política brasileira. Para isso, foram testados diferentes classi-

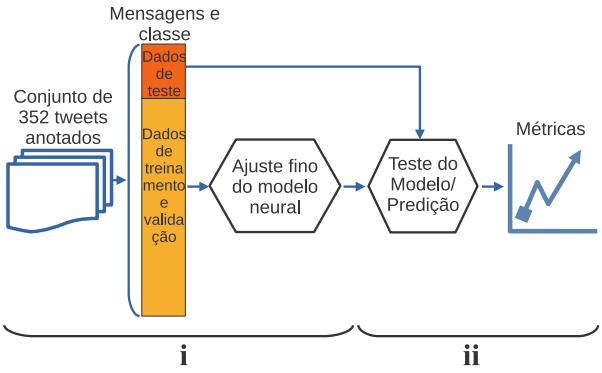

Figura 9: A configuração experimental proposta foi dividida em duas etapas: i. ajuste fino dos modelos neurais e ii. avaliação do modelo.

ficadores com diferentes hiperparâmetros em um corpus de 352 *tweets* do domínio da política brasileira. Desses, 281 *tweets* foram usados para treinamento ou ajuste fino dos modelos e os restantes 71 *tweets* foram usados para teste.

As seções a seguir apresentam as análises quantitativas e qualitativas dos resultados obtidos nos experimentos.

5.1. Análise quantitativa

Em termos das medidas quantitativas usualmente aplicadas na avaliação de modelos computacionais, apresentadas na Tabela 4, concluímos que o modelo neural obtido com o ajuste fino do BERTimbau obteve os melhores valores: 63,38% de acurácia, 69,65% de precisão, 63,38% de cobertura e uma medida F de 63,61%.[31]

Entre os algoritmos de AM baseado em *features*, o desempenho variou de 35% a 54% de medida F, ficando a pelo menos cerca de 9 pontos percentuais abaixo do melhor modelo neural. Quanto ao baixo desempenho desses modelos, vale mencionar que a alta quantidade de *features* usadas no treinamento pode ter influenciado esses resultados. Mesmo com a etapa de seleção de *features*, realizada como descrito na Figura 8, os modelos tradicionais foram treinados com 290 *features* para 281 instâncias e essa alta dimensionalidade pode ter confundido os algoritmos no momento de gerar os modelos. Contudo, a configuração experimental do AM baseado em *features* foi proposta para dar autonomia ao processo de AM na definição das *features* mais relevantes (sem a interferência do especialista humano), de

[31] Os melhores resultados foram alcançados com os seguintes hiperparâmetros otimizados: número de épocas= 20; *batch size*= 8; *early stop*= 2; *learning rate*=1e-5. Os mesmos hiperparâmetros foram utilizados para o BERTimbau e RoBERTaTwitterBR.

Classificador	Acurácia	Precisão	Cobertura	Medida F
LR	45,07%	45,34%	45,07%	44,56%
LR + RFECV	46,48%	46,53%	46,48%	45,75%
K-NN	47,89%	49,20%	47,89%	46,86%
K-NN + RFECV	39,44%	41,68%	39,44%	37,80%
DT	46,48%	32,57%	46,48%	37,71%
DT + RFECV	46,48%	32,57%	46,48%	37,71%
SVM	43,66%	42,99%	43,66%	41,10%
SVM + RFECV	35,21%	35,95%	35,21%	35,39%
RF	54,93%	54,90%	54,93%	54,44%
RF + RFECV	45,07%	46,53%	45,07%	43,24%
NB	45,07%	45,06%	45,07%	44,29%
NB + RFECV	50,70%	50,89%	50,70%	50,49%
RobertaTwitterBR	49,30%	48,41%	49,30%	48,11%
BERTimbau	63,38%	69,65%	63,38%	63,61%

Tabela 4: Valores das medidas de avaliação obtidas para a qualidade geral da argumentação nos modelos baseados em *features* (LR, K-NN, DT, RF e NB) com (+ RFECV) e sem a etapa de seleção de *features* e nos modelos neurais (RobertaTwitterBR e BERTimbau).

modo semelhante ao que fazem os modelos neurais durante o ajuste fino.

É importante destacar que os experimentos conduzidos com DT apresentaram resultados idênticos com e sem a etapa de seleção de *features*. Isso se deve às características do próprio algoritmo, que já faz a seleção de *features* durante o treinamento para definir qual *feature* vai em cada nó da árvore. O DT trabalha dividindo recursivamente o conjunto de *features* em subconjuntos homogêneos, até que um critério de parada seja atingido. Durante o processo de construção da árvore, o algoritmo avalia diferentes *features* e escolhe aquela que melhor separa as classes ou minimiza o erro.

Nota-se, também, que os experimentos conduzidos com SVM sem a etapa de seleção de *features* apresentaram melhores resultados se comparados com o experimento com a seleção de *features*. Isso também se deve às características do algoritmo, que trabalha transformando as *features* em um espaço de alta dimensão, no qual é mais provável que as classes sejam linearmente separáveis: quanto mais pontos, melhor o detalhamento na definição das bordas. Em seguida, o SVM encontra o hiperplano que separa as classes, minimizando o erro ou a perda (Vapnik, 1999).

A partir dos valores das medidas de avaliação apresentados nesta seção é possível concluir que os modelos de AM basedos em *features* não são os mais indicados para identificar automaticamente os *tweets* do campo da política brasileira que tenham bons argumentos. Embora os experimentos tenham sido realizados de maneira bastante abrangente em termos de recursos linguístico-computacionais usados para o português, *features*

e algoritmos, não é possível apontar claramente quais desses fatores levaram ao baixo desempenho.

Em relação aos resultados dos modelos neurais, a Tabela 5 e a Figura 10 apresentam os resultados detalhados do modelo de melhor desempenho obtido com o ajuste fino do BERTimbau. Esse modelo teve uma excelente precisão para a classe Alta. Das 19 instâncias da classe Alta presentes no corpus de teste, o modelo foi capaz de predizer 9 delas corretamente (resultando em uma cobertura de 47,37%) mas com uma precisão de 100%. Assim, embora a amostra do corpus de teste seja pequena, esses resultados trazem fortes indícios de uma excelente assertividade do modelo para prever *tweets* com alta qualidade da argumentação. Esses *tweets* são apresentados na Tabela 6.

	Precisão	**Cobertura**	**Medida F**
Baixa	64,52%	71,43%	67,80%
Média	51,61%	66,67%	58,18%
Alta	100,00%	47,37%	64,29%

Tabela 5: Resultados detalhados para a predição da qualidade geral da argumentação retornados pelo modelo obtido com o ajuste fino do BERTimbau.

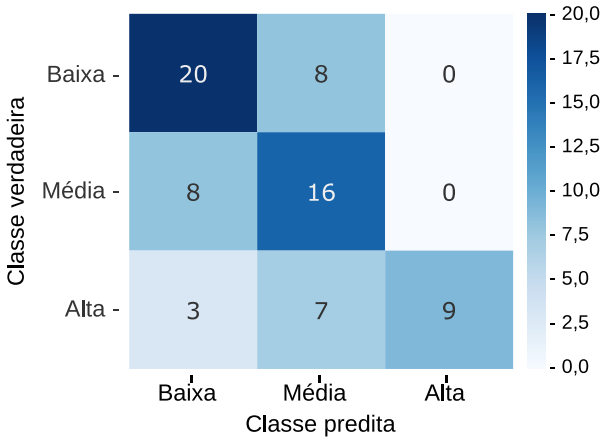

Figura 10: Matriz de confusão do modelo obtido com o ajuste fino do BERTimbau.

5.2. Análise qualitativa

Para complementar a análise quantitativa, esta seção traz uma análise qualitativa das classificações do modelo de melhor desempenho para a predição da qualidade da argumentação nos *tweets* do domínio da política brasileira.

5.2.1. Análise da classe de Alta qualidade de argumentação

Como mencionado anteriormente, o objetivo dos experimentos apresentados neste artigo foi gerar um modelo capaz de predizer, com um bom desempenho, aqueles *tweets* que tenham alta qualidade de argumentação. Assim, a Tabela 6 apresenta os 9 *tweets* do corpus de teste que o modelo ajustado a partir do BERTimbau apontou corretamente como sendo de alta qualidade de argumentação.

Vale mencionar que todos foram considerados pelos anotadores humanos como sendo claros, organizados e com nenhum ou poucos erros de português. Além dessas características gerais que levaram os anotadores a considerar tais *tweets* como de alta qualidade da argumentação, outras pistas linguísticas (destacadas em negrito) incluem: uso de tratamento cordial (ex: Desculpe, Caro, senhora, Sr.), fato histórico (ex: como foi feito na Alemanha), numérico (ex: 56 milhões de eleitores) ou citação de figura/entidade pública que dá respaldo a alguma informação (ex: Arthur Lira, conselho de ética). A presença de marcadores discursivos (ex: como, porque, ainda mais) também foi apontada como positiva para melhorar a qualidade argumentativa porque eles unem, contrapõem ou interligam ideias.

Como exemplo, na mensagem 6.1, da Tabela 6, dentre as principais pistas linguísticas que fizeram com que os humanos considerassem esse *tweet* como de alta qualidade de argumentação, destacam-se a presença de: (i) termo especializado "Impeachment" (apontado por todos os anotadores); (ii) encadeamento cronológico expresso pela presença do "quando", usado com sentido temporal (apontado por todos os anotadores), e (iii) fato midiático (pista anotada pela maioria dos anotadores).

Por outro lado, outros 10 *tweets*, que correspondem às mensagens 7.1 até 7.10, da Tabela 7, considerados de alta qualidade argumentativa pelos anotadores, foram classificados pelo modelo BERTimbau de modo diferente: 70% deles como Média e 30% deles como Baixa.

Comparando os *tweets* da Tabela 7 com os da Tabela 6 podemos notar uma maior presença de abreviações (ex: n, q, ñ, tds) e grafias diferentes para palavras (ex: "K0V1D" para representar Covid, ou "vac1na" para vacina), além de uma ausência notável de pontuação (como o uso da vírgula). Essas características retratam bem as peculiaridades do gênero textual e aparentemente tiveram bastante peso na classificação do modelo BERTimbau.

Ex.	Texto
6.1	Rodrigo Maia, você hoje já falou que se arrepende do apoio a Bolsonaro no segundo turno. Parabéns por admitir isto. Agora... **quando** virá o arrependimento de não ter ao menos colocado para a frente algum dos pedidos de **Impeachment**?
6.2	Vc propôs essa emenda, esperando que passe ou apenas para constar? Com a **postagem do** seu presidente da câmara, que até já considerou que o Dep. Daniel Silveira contrapôs à democracia, **mesmo não tendo** sido julgado e condenado pelo STF, espera que essa sua proposta tenha sucesso? https://t.co/uJjvgcwqEt
6.3	**Desculpe senhora** deputada, **cansei** de vcs ! Ninguém faz nada, ninguém! Vcs brincam com o povo! **Se** hoje um governador maluco fizer um forno, **como foi feito na Alemanha** e começar a matar as pessoas,tudo bem , os caras que jamais devem ser citados, deram o direito !
6.4	**Caro** Deputado, não sei se irá ler **meu** posicionamento. **Mas**, **calaram a voz** de uma Deputado q foi eleito para **PODER FALAR POR NÓS!** Um PODER, calou a não a voz do Daniel, calou foi a NOSSA! Ontem foi deputado pondo mordaça da boca de outro deputado e traçando o fim do CONGRESSO.
6.5	Está na hora de exigir o respeito com seriedade, **impeachment** se faz mais que necessário, ele está tentando rebaixar a **Câmara dos Deputados** a seu serviço, uma ação judicial enérgica imediata. Ação do **Arthur Lira** agora, **se** deixar passar perderá a força 💪 👊 🇧🇷 🇧🇷 🇧🇷 🇧🇷
6.6	**Ao Sr**. Apresentar esse material ao **conselho de ética**, vamos ver **se** esse conselho de ética e justo, **ou** hipócrita **ou** incoerente. **Se** esse conselho disser que esses deputados e senadores que cometeram crime de ofensa, no cometeram crime **porque** tem **imunidade parlamentar**. Aí tem!!!
6.7	A prisão é ilegal já no momento que ele usa um **inquérito que** serve pra **tudo**, sem requerimento da polícia ou do **MP** o resto só invalida **ainda mais** a prisão, **o que esse ministro fez é caso de impeachment e prisão**
6.8	Mas também deputada, com essa oposição que **tudo** que o governo federal faz vocês acham que está errado. Imagine **se** o povo estivesse **todos** seguindo o **FIQUE EM CASA, A ECONOMIA A GENTE VER DEPOIS**. **Sou** a favor que sejam seguidos os protocolos: máscara, lavar as mãos e não aglomerar.
6.9	Deveria abrir uma **CPI para investigar** as **sabotagens** que o ex-presidente da câmara comandou durante seu mandato, **causando prejuízos incalculáveis** à **toda nação** e desrespeitando a vontade de mais de **56 milhões de eleitores**!

Tabela 6: 9 *tweets* de alta qualidade presentes no corpus de teste corretamente classificados pelo modelo ajustado a partir do BERTimbau.

5.2.2. Análise dos erros

A seguir, apresenta-se uma breve discussão das principais disparidades observadas entre a anotação humana e a classificação com o modelo ajustado a partir do BERTimbau. Na Tabela 7, são apresentadas algumas mensagens do conjunto de testes que demonstraram divergência entre as avaliações humanas e automáticas.

Os exemplos 7.8, 7.9 e 7.10, da Tabela 7, são alguns dos casos nos quais foram observadas as principais divergências, uma vez que o modelo os classificou como Baixa argumentatividade e os anotadores os avaliaram como Alta argumentatividade. Características como ausência de vírgulas no exemplo 7.8, presença de maiúsculas no exemplo 7.9 e de *hashtags* no exemplo 7.10 podem ter influenciado o modelo a classificá-los como de Baixa qualidade. Por outro lado, os anotadores humanos provavelmente os classificaram como de Alta qualidade com base no conteúdo e apelo emocional e não na forma superficial do texto.[32] No restante dos exemplos (7.11, 7.12, 7.13 e 7.14) houve diferenças entre Baixa/Média e Média/Alta argumentatividade. Essas diferenças foram vistas como plausíveis, considerando o contexto semântico do argumento.

[32]Os exemplos 7.8 e 7.9 foram classificados como de Alta qualidade pela maioria dos anotadores, e o exemplo 7.10 foi classificado como de Alta qualidade por todos os anotadores.

Além dos exemplos de erros apresentados na Tabela 7, outras 12 mensagens foram classificadas de forma equivocada pelo modelo. Destas, 6 de Baixa argumentatividade foram classificadas como Média e 6 de Média argumentatividade foram classificadas como Baixa. Importante ressaltar que todas as mensagens classificadas como de Alta qualidade da argumentação pelo modelo ajustado a partir do BERTimbau foram avaliadas da mesma forma pelos anotadores, isso indica uma precisão absoluta em classificar mensagens de alta argumentatividade.

6. Conclusões, limitações e trabalhos futuros

Este artigo apresentou experimentos para a predição da qualidade da argumentação em *tweets* do domínio da política brasileira. Os experimentos foram realizados com algoritmos de AM baseado em *features* e em redes neurais. O modelo que apresentou o melhor desempenho foi o obtido a partir do ajuste fino do BERTimbau.

Em relação às questões de pesquisa inicialmente propostas para este trabalho, a partir dos experimentos aqui apresentados pode-se concluir que: (Q1) é possível predizer automaticamente a qualidade da argumentação em *tweets* do domínio da política brasileira e (Q2) a maneira que se mostrou mais adequada para fazê-lo foi usando o modelo neural gerado a partir do

Ex.	Texto	Real	BERTimbau
7.1	**Acho que** o foco deveria ser nas desastrosas intervenções nas estatais (**que já** n deveriam ser mais estatais) e nos preços de energia, que logo se avizinham. N adianta ser **um leão contra** a prisão (q deveria ter discurso técnico), e **um gato contra** o desastre das políticas do gov.!	Alta	Média
7.2	**Precisamos tds** deixar bem claro que essa ñ representa o povo e que esses parlamentares são responsáveis pelas mortes diárias de brasileiros. É imoral que estejam votando um projeto de lei para se protegerem da lei **em vez de** salvar a vida de brasileiros!	Alta	Média
7.3	Quem sabe né o seu Jair da uma dentro. O dublê de ministro da saúde tá sendo um fiasco, um fracasso, uma vergonha nacional. Uma coisa é usina hidroelétrica outra coisa é a Petrobrás. Veremos.	Alta	Média
7.4	Hoje fui comprar 1kg de carne moída para o almoço e deu R$ 43,00 achei um absurdo! Em que mundo estamos com um custo tão alto de carne assim?! Mas de boa estamos comprando carne de primeira para nossos representantes políticos, então pq reclamar?! 😳🧐	Alta	Média
7.5	Cadê a dengue, cadê a gripe comum, a H1N1, cadê as demais doenças respiratórias, tão comuns em nosso país nessa época do ano? Sumiram ou estão se somando aos números do K0V1D? Tudo pela ciência e pela saúde! Principalmente agora que governadores e prefeitos podem comprar vac1na.	Alta	Média
7.6	O que o Brasil precisa é um sistema bancário que ofereca juros selic 2% ao ano para qualquer um empreender. Reduzir esse spread avassalador. Infelizmente politico so pensa em imposto e nao em melhorar a eficiencia.	Alta	Média
7.7	Esse é a teoria, . Na vida real, na prática brasileira, vendem água mineral a 30 reais, quando tem desastre natural... Privatizem mesmo! Abram o mercado sim! Quebrem os monopólios! Mas, acima de tudo, FISCALIZEM os espertinhos, senão eles deitam e rolam!	Alta	Média
7.8	Ah pelo amor Coco Bambú vive lotado se não tem grana para manter os funcionários durante o Lockdown tem péssima administração! Empatia zero e o governo federal precisa ajudar as pessoas essa que é a realidade!	Alta	Baixa
7.9	NAO É CRIME DE OPINIAO. VOU REPETIR PRO SARGENTO GARCIA ENTENDER. NAO É CRIME DE OPINIAO, e nao está coberto peka imunidade material so parlamentar, basta saber ler o art. 53. Mas é esperar muito de bolsonaristas. Que vergonha pro meu Parana ter deputado assim.	Alta	Baixa
7.10	Pode jogar a nossa Constituição no lixo, pois a acabam de transformar os 11 ministros do STF nos novos IMPERADORES do Brasil. #CamaraDosDeputadosVergonhaNacional #CâmaraDosDeputadosRasgandoAConstituição #DanielSilveiraFalouPeloPovo #STFVergonhaMundial	Alta	Baixa
7.11	Você pode ter 6 seguranças armados né deputado, e defender bandido, porque não pede ao STF para mandar investigar todo o governo do Rio de janeiro incluindo vocês de Dourados, vereadores e senadores, talvez assim consigam desarmar os traficantes dos morros	Média	Baixa
7.12	Os e o Sr perderam o meu respeito não acredito mais nessa turma de cristão não tem nada lembrei de Bararabas e dos lideres do sinedrio foi igual a passagem veio a minha mente no voto de muitos da AD e da IURD	Média	Baixa
7.13	Do jeito que esses deputados são, depois que vimos semana passada, acho difícil conseguir as assinaturas necessárias. São um bando de covarde, só pensam neles, no próprio ego. Nem a voz do povo eles escutam. 2022 é logo ali. #TodoPoderEmanaDoPovo #Democracia	Baixa	Média
7.14	Nada melhora pro povo já que não há instituições neste país apodreceu tdo e fede faz tempo, mas pelo menos é uma dívida a menos nas costas dos pagadores de impostos, mais conhecidos pela elite como trouxas	Baixa	Média

Tabela 7: Exemplos de erros de predição do modelo neural.

ajuste fino do BERTimbau. Tal modelo foi capaz de predizer com 100% de precisão instâncias da classe de Alta qualidade da argumentação, satisfazendo o objetivo deste trabalho que é o de encontrar *tweets* com boa capacidade argumentativa.

Embora o objetivo do trabalho tenha sido alcançado, algumas limitações são apresentadas na próxima subseção, seguida de algumas propostas de trabalhos futuros.

6.1. Limitações

Uma das principais limitações deste trabalho é o tamanho do corpus. Embora seja compatível com o corpus desenvolvido por Wachsmuth et al. (2017b) quanto ao número de instâncias, os resultados aqui apresentados têm seus impactos limitados pela pouca quantidade de instâncias usadas na geração e na avaliação dos modelos.

Outra limitação está relacionada à decisão de projeto adotada no momento da anotação do corpus que foi a de realizar a anotação da qualidade da argumentação de forma isolada, ou seja, desconsiderando o *tweet* semente. Diante disso, assumiu-se que o argumento pode ser classificado independentemente do tópico. Trabalhos da literatura (Fromm et al., 2019; Hidayaturrahman et al., 2021) apontam que, muitas vezes, frases contendo argumentos são estruturalmente semelhantes a frases puramente informativas sem qualquer posicionamento sobre o tópico e que considerar a informação do tópico é crucial para a tarefa, pois ele define o contexto semântico de um argumento. Para lidar com essa limitação, uma possível estratégia é concatenar as mensagens avaliadas com o *tweet* semente de modo a incorporar contexto e, consequentemente, melhorar o desempenho do modelo.

Outra decisão de projeto que pode ter impactado o desempenho do modelo está relacionada ao modo como os aspectos foram combinados para definir a Qualidade Geral da argumentação. Conforme visto na Tabela 7, algumas instâncias de teste anotadas como de Alta qualidade foram classificadas pelo modelo BERTimbau como de Média ou Baixa qualidade o que, analisando o conteúdo de tais *tweets* pode ter fundamento. Assim, uma proposta de trabalho futuro é fazer uma revisão dos critérios e do modo como os aspectos são combinados para definir a Qualidade Geral.

É importante considerar as características específicas do Twitter. Diferentemente de outras mídias sociais, o Twitter não possui uma política das mais rígidas para restringir ou filtrar o conteúdo das postagens ou o comportamento abusivo dos usuários. Por causa disso, postagens que contêm palavrões e discurso de ódio são comuns. No campo da política brasileira atual, essas características são ainda mais acentuadas, com postagens contendo *fake news*, ataques pessoais a políticos ou às famílias deles, ideologia política, entre outros. Assim, esses textos (*tweets*) costumam ter várias marcas que impactam negativamente a qualidade deles e, consequentemente, reduzem a qualidade geral da argumentação.

Por fim, outra limitação relacionada ao Twitter é o número muito limitado de caracteres (280) permitido para cada mensagem, o que dificulta o uso de estratégias mais elaboradas de argumentação linguística pelos autores das postagens.

6.2. Trabalhos futuros

Como trabalhos futuros, destacam-se três caminhos que trariam maior benefício às propostas aqui apresentadas: (i) como apresentado na Seção 6.1, concatenar as mensagens avaliadas com o *tweet* semente de modo a incorporar contexto e, consequentemente, melhorar o desempenho do modelo; (ii) testar uma combinação (*ensemble*) de classificadores, que constitui-se em múltiplos classificadores, treinados de forma individual, cujos resultados são combinados, também com o objetivo de buscar uma melhora no desempenho do modelo; (iii) aumentar o corpus com a finalidade de capturar novos contextos, além de um acréscimo da sua própria dimensão.

Além dessas, outra possibilidade seria gerar modelos específicos para cada um dos aspectos da qualidade da argumentação (Clareza, Organização, Credibilidade e Polaridade e Intensidade do Apelo Emocional) com o intuito de combiná-los para definir automaticamente a qualidade geral da argumentação, como foi feito de modo manual, pelos anotadores, no momento da geração do corpus (Silva et al., 2021).

Agradecimentos

Os autores agradecem aos linguistas e coautores em (Silva et al., 2021), que anotaram o corpus e contribuíram para a elaboração das diretrizes de anotação, utilizados neste trabalho: Amanda Pontes Rassi, Jackson Wilke da Cruz Souza, Renata Ramisch e Roger Alfredo de Marci Rodrigues Antunes. Agradecem, também, ao Sidney Evaldo Leal e ao Núcleo Interinstitucional de Linguística Computacional (NILC) pelos recursos linguístico-computacionais disponibilizados para o desenvolvimento desta pesquisa. Por fim, os autores agradecem ao Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Universidade Federal de São Carlos (UFSCar) e à Rede Gonzaga de Ensino Superior (REGES), pelo apoio a este trabalho.

Referências

Adi, Sumarni, Yoga Pristyanto & Andi Sunyoto. 2019. The best features selection method and relevance variable for web phishing classification. Em *International Conference on Information and Communications Technology (ICOIACT)*, 578–583. 10.1109/ICOIACT46704.2019.8938566.

Al-Khatib, Khalid, Henning Wachsmuth, Matthias Hagen, Jonas Köhler & Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 1395–1404. 10.18653/v1/N16-1165.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel. L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan. Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez–Martínez & Miriam. A. Scalco. 2006. Open-source Portuguese–Spanish machine translation. Em *VII Encontro para o Processamento Computacional da Língua Portuguesa (PROPOR)*, 50–59.

Balage Filho, Pedro P., Thiago Alexandre Salgueiro Pardo & Sandra M. Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 215–219.

Bench-Capon, Trevor JM & Paul E Dunne. 2007. Argumentation in artificial intelligence. *Artificial intelligence* 171(10-15). 619–641. 10.1016/j.artint.2007.05.001.

Bilu, Yonatan & Noam Slonim. 2016. Claim synthesis via predicate recycling. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 525–530. 10.18653/v1/P16-2085.

Blair, J. Anthony. 2012. Rhetoric, dialectic, and logic as related to argument. *Philosophy & Rhetoric* 45(2). 148–164. 10.5325/philrhet.45.2.0148.

Boudry, Maarten, Fabio Paglieri & Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4). 431–456. 10.1007/s10503-015-9359-1.

Brum, Henrico & Maria Graças Volpe Nunes. 2018. Building a sentiment corpus of tweets in Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4167–4172.

Capellaro, Leonardo & Helena Caseli. 2021. Análise de polaridade e de tópicos em tweets no domínio da política no Brasil. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 47–55. 10.5753/stil.2021.17783.

Carlile, Winston, Nishant Gurrapadi, Zixuan Ke & Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. Em *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 621–631. 10.18653/v1/P18-1058.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 4171–4186. 10.18653/v1/N19-1423.

Eemeren, Frans H & Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation* 1(3). 283–301. 10.1007/BF00136779.

Eemeren, Frans H. van & Rob Grootendorst. 2003. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press. 10.1017/CBO9780511616389.

Feng, Vanessa Wei, Ziheng Lin & Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. Em *25th International Conference on Computational Linguistics (COLing)*, 940–949.

Fonseca, Erick. Rocha & João Luís G. Rosa. 2013. Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 98–107.

Fromm, Michael, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman & Thomas Seidl. 2022. Towards a holistic view on argument quality prediction. 10.48550/ARXIV.2205.09803. ArXiv cs.CL.

Fromm, Michael, Evgeniy Faerman & Thomas Seidl. 2019. TACAM: Topic and context aware argument mining. Em *IEEE/WIC/ACM International Conference on Web Intelligence*, 99–106. 10.1145/3350546.3352506.

García-Gorrostieta, Jesús Miguel & Aurelio López-López. 2018. Identifying argumentative paragraphs: Towards automatic assessment of argumentation in theses. Em *International Conference on Applications of Natural Language to Information Systems*, 83–90. 10.1007/978-3-319-91947-8_9.

García-Gorrostieta, Jesús M., Aurelio López-López & Samuel González-López. 2018. Automatic argument assessment of final project reports of computer engineering students. *Computer Applications in Engineering Education* 26(5). 1217–1226. 10.1002/cae.21996.

Gleize, Martin, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov & Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. Em *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 967–976. 10.18653/v1/P19-1093.

Gretz, Shai, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov & Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. 10.48550/ARXIV.1911.11408. ArXiv cs.CL.

Habernal, Ivan & Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2127–2137. 10.18653/v1/D15-1255.

Habernal, Ivan & Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1589–1599. 10.18653/v1/P16-1150.

Habernal, Ivan & Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics* 43(1). 125–179. 10.1162/COLI_a_00276.

Hidayaturrahman, Emmanuel Dave, Derwin Suhartono & Aniati Murni Arymurthy. 2021. Enhancing argumentation component classification using contextual language model. 10.1186/s40537-021-00490-2.

Krippendorff, Klaus. 2011. Computing krippendorff 's alpha-reliability. Relatório técnico. University of Pennsylvania. http://repository.upenn.edu/asc_papers/43/.

Lauscher, Anne, Lily Ng, Courtney Napoles & Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. Em *27th International Conference on Computational Linguistics (COLing)*, 4563–4574. 10.18653/v1/2020.coling-main.402.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2022. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. 10.48550/ARXIV.2201.03445. ArXiv cs.CL.

Leite, João Augusto, Diego Silva, Kalina Bontcheva & Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 914–924.

Lytos, Anastasios, Thomas Lagkas, Panagiotis Sarigiannidis & Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56(6). 102055. 10.1016/j.ipm.2019.102055.

Marcuschi, Luiz Antônio et al. 2002. Gêneros textuais: definição e funcionalidade. Em *Gêneros textuais e ensino*, Lucerna.

Maćkiewicz, Andrzej & Waldemar Ratajczak. 1993. Principal components analysis (PCA). *Computers & Geosciences* 19(3). 303–342. 10.1016/0098-3004(93)90090-R.

Misra, Puneet & Arun Singh Yadav. 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal of Emerging Technologies* 11(3). 659–665.

Peldszus, Andreas & Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 938–948. 10.18653/v1/D15-1110.

Persing, Isaac & Vincent Ng. 2015. Modeling argument strength in student essays. Em *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 543–552. 10.3115/v1/P15-1053.

Potthast, Martin, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein & Matthias Hagen. 2019. Argument search: Assessing argument relevance. Em *42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1117–1120. 10.1145/3331184.3331327.

Putra, Jan Wira Gotama, Simone Teufel & Takenobu Tokunaga. 2021. Annotating argumentative structure in English-as-a-foreign-language learner essays. *Natural Language Engineering* 28(6). 797–823. 10.1017/S1351324921000218.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410.

Reimers, Nils, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab & Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. Em *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 567–578. 10.18653/v1/P19-1054.

Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni &

Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 440–450. 10.18653/v1/D15-1050.

Rosenfeld, Ariel & Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems* 6(4). 1–33. 10.1145/2983925.

Rossini, Patrícia. 2019. Disentangling uncivil and intolerant discourse in online political talk. Em *A Crisis of Civility?*, 142–157. Routledge. 10.4324/9781351051989-9.

Rossini, Patrícia. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research* 49(3). 399–425. 10.1177/0093650220921314.

Schaefer, Robin & Manfred Stede. 2021. Argument mining on twitter: A survey. *it – Information Technology* 63(1). 45–58. 10.1515/itit-2020-0053.

Silva, Cássio, Amanda Rassi, Jackson Souza, Renata Ramisch, Roger Antunes & Helena Caseli. 2021. Quality of argumentation in political tweets: what is and how to measure it / qualidade da argumentação em tweets de política: o que e como avaliar. *Estudos da Linguagem* 29(4). 2537–2586. 10.17851/2237-2083.29.4.2537-2586.

Skitalinskaya, Gabriella, Jonas Klaff & Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. Em *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1718–1729. 10.18653/v1/2021.eacl-main.147.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Brazilian Conference on Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28.

Stab, Christian & Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. Em *25th International Conference on Computational Linguistics (COLing)*, 1501–1510.

Stab, Christian & Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43(3). 619–659. 10.1162/COLI_a_00295.

Stab, Christian & Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 980–990.

Swanson, Reid, Brian Ecker & Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. Em *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226. 10.18653/v1/W15-4631.

Toledo, Assaf, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov & Noam Slonim. 2019. Automatic argument quality assessment: New datasets and methods. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)*, 5625–5635. 10.18653/v1/D19-1564.

Toulmin, Stephen E. 2003. *The uses of argument*. Cambridge university press.

Vapnik, Vladimir. 1999. *The nature of statistical learning theory*. Springer.

Wachsmuth, Henning, Khalid Al-Khatib & Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. Em *26th International Conference on Computational Linguistics (COLing)*, 1680–1691.

Wachsmuth, Henning, Johannes Kiesel & Benno Stein. 2015. Sentiment flow - a general model of web review argumentation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 601–611. 10.18653/v1/D15-1072.

Wachsmuth, Henning, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych & Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. Em *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 250–255. 10.18653/v1/P17-2039.

Wachsmuth, Henning, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst & Benno Stein. 2017b. Computational argumentation quality assessment in natural language. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 176–187.

Wachsmuth, Henning, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann,

Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff & Benno Stein. 2017c. Building an argument search engine for the web. Em *4th Workshop on Argument Mining (ArgMining 2017)*, 49–59.

Wachsmuth, Henning, Benno Stein & Yamen Ajjour. 2017d. "PageRank" for argument relevance. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1117–1127.

Wachsmuth, Henning & Till Werner. 2020. Intrinsic quality assessment of arguments. Em *28th International Conference on Computational Linguistics*, 6739–6745. 10.18653/v1/2020.coling-main.592.

Walton, Douglas N & David N Walton. 1989. *Informal logic: A handbook for critical argument.* Cambridge University Press.

Wei, Zhongyu, Yang Liu & Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 195–200. 10.18653/v1/P16-2032.

Weltzer-Ward, Lisa, Beate Baltes & Laura Knight Lynn. 2009. Assessing quality of critical thought in online discussion. *Campus-Wide Information Systems* 26(3). 168–177. 10.1108/10650740910967357.

Zhang, Justine, Ravi Kumar, Sujith Ravi & Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 136–141. 10.18653/v1/N16-1017.