





ARTIGO DE PESQUISA/RESEARCH PAPER

Explorando meta-variáveis para a classificação de persuasão em textos de memes políticos

Exploring meta-features for classifying persuasion in texts of political memes

Ana Beatriz Soares de Azevedo   [Escola Nacional de Ciências Estatísticas | anabeatriz98azevedo@gmail.com]

Eduardo Corrêa Gonçalves  [Escola Nacional de Ciências Estatísticas | eduardo.correa@ibge.gov.br]

 Coordenação de Graduação (CGRAD), Escola Nacional de Ciências Estatísticas (ENCE-IBGE), Rua André Cavalcanti, 106, Bairro de Fátima, Rio de Janeiro, RJ, 20231-050, Brasil.

Resumo. Este trabalho propõe uma abordagem baseada em engenharia de características para detectar estratégias persuasivas em textos de memes políticos. Foram definidos quatro grupos de meta-variáveis: (i) retóricas, (ii) de sentimento e discurso de ódio, (iii) estruturais e (iv) contextuais. Os experimentos utilizaram a base da Task 4 da competição SemEval 2024, com 7.000 instâncias de treino e 1.000 de teste. Os algoritmos Floresta Aleatória e Regressão Logística foram avaliados com e sem o tratamento do desbalanceamento entre as classes da base de dados de treinamento. O melhor resultado, com F1-macro de 0,701, foi obtido pela combinação dos grupos de meta-variáveis retóricas e estruturais. A abordagem proposta oferece uma alternativa computacionalmente eficiente e interpretável frente aos modelos baseados em redes neurais profundas.

Abstract. This work proposes a feature-engineering-based approach for detecting persuasive strategies in political memes. Four groups of meta-features were designed: (i) rhetorical, (ii) sentiment and hate speech, (iii) structural, and (iv) contextual features. Experiments were performed using the SemEval-2024 Task 4 challenge dataset with 7,000 training and 1,000 testing instances. Random Forest and Logistic Regression were evaluated with and without handling the imbalance between classes in the training dataset. The best result, with a macro-F1 value of 0.701, was achieved by combining rhetorical and structural meta-features. The proposed approach offers a computationally efficient and interpretable alternative to deep learning models.

Palavras-chave: meta-variáveis, persuasão, classificação de memes, classificação de textos curtos, aprendizado de máquina

Keywords: meta-features, persuasion, meme classification, short text classification, machine learning

Recebido/Received: 21 November 2025 • Aceito/Accepted: 09 January 2026 • Publicado/Published: 16 January 2026

1 Introdução

A constante e acelerada evolução da era digital causou uma profunda transformação em todos os aspectos da sociedade, principalmente na forma com que nos comunicamos [Weiss, 2019]. Neste cenário, os memes, com sua essência caracterizada pela combinação de informações visuais e textuais com humor e uma linguagem adaptável, se consolidaram como uma das mais populares formas de linguagem na internet [Dimitrov *et al.*, 2024]. Com o volume e velocidade proporcionados pelas redes, sua rápida disseminação e capacidade de sintetizar ideias complexas de maneira acessível os tornam poderosas ferramentas de comunicação em massa. Assim, muitos memes são instrumentalizados para além do humor ou crítica: tornam-se dispositivos de persuasão, capazes de influenciar opiniões e ideologias de forma muitas vezes sutil, porém eficaz [Halversen and Weeks, 2023].

Considerando a magnitude e amplitude do impacto potencial, a pesquisa científica das mais diversas áreas do conhecimento sobre esta questão intensificou-se. Um exemplo foi a competição SemEval 2024 Task 4 – Detecção Multilíngue de Técnicas de Persuasão em Memes [Dimitrov *et al.*, 2024]. Os organizadores da competição forneceram uma base de dados de treinamento composta por 7.000 memes políticos em inglês (além de 3 bases surpresa em árabe, búlgaro e macedônio,

disponibilizadas apenas na última fase da competição), que poderiam ou não conter técnicas de persuasão embutidas. O desafio era dividido em diferentes subtarefas, sendo as subtarefas 1 e 2a focadas em classificação multirrotulo [Gonçalves *et al.*, 2013] (criar um sistema classificador para identificar o conjunto de técnicas de persuasão presentes em memes persuasivos, dentre 21 técnicas possíveis) e a subtarefa 2b voltada para classificação binária (criar um sistema classificador para detectar se um meme é persuasivo ou não). O presente trabalho tem por foco a subtarefa 2b em inglês.

Dos sistemas vencedores da subtarefa 2b [Anghelina *et al.*, 2024; Li *et al.*, 2024; Yu *et al.*, 2024], todos adotaram uma solução multimodal (ou seja, o treinamento envolveu tanto o uso da imagem como o texto do meme) e basearam-se em modelos de redes neurais profundas – como CLIP [Radford *et al.*, 2021] e BERT [Devlin *et al.*, 2019] – para extrair embeddings das imagens e textos. Tais sistemas são poderosas ferramentas e produzem resultados impressionantes, entretanto, possuem conhecidas desvantagens [Caseli and Nunes, 2024; Navigli *et al.*, 2023], como a falta de interpretabilidade, o grande volume de dados necessários para treinamento e o alto custo computacional. Sendo a tarefa proposta na SemEval-2024 um novo e moderno desafio por si só, é impossível conter a curiosidade que surge neste preciso meio: como o modelo chegou a uma conclusão? Quais fórmulas

linguísticas proporcionam esse efeito? A literatura na área destaca a identificação de variáveis que influenciam a percepção do público alvo como sendo etapa fundamental do desenvolvimento de sistemas persuasivos e que responderiam estas questões [Braca and Dondio, 2023].

Com o objetivo de cobrir essa lacuna, o presente artigo investiga o poder preditivo de diferentes grupos de meta-variáveis para a classificação de memes persuasivos. Meta-variáveis são informações extraídas de dados brutos, como somatórios ou contagens de elementos textuais como palavras, frases, pontuações e outras características relevantes para a análise [Carvalho and Plastino, 2021; Santos *et al.*, 2019]. Entre as possíveis características, destaca-se a retórica aristotélica, sendo ela a arte de persuadir por meio da argumentação lógica (*logos*), da credibilidade e ética do orador (*ethos*), e do apelo à emoção do público (*pathos*), ainda relevantes e utilizadas pela linguística moderna para entender como a linguagem influencia o público [Braca and Dondio, 2023].

Mais precisamente, este trabalho investiga a efetividade de quatro diferentes grupos de meta-variáveis na tarefa de classificar um meme como persuasivo ou não em função de seu texto. O primeiro grupo é baseado na tríade retórica aristotélica que correspondem ao uso de apelos à lógica, ética e emoção nos textos. O segundo grupo é composto por meta-variáveis tradicionalmente utilizadas nos problemas de análise de sentimentos e detecção de discurso de ódio, com o intuito de identificar se as mesmas também são úteis para a classificação de persuasão. O terceiro abrange meta-variáveis que capturam a estrutura lexical e estatísticas básicas do texto. Por fim, o quarto grupo busca capturar o uso de recursos linguísticos que influenciam o tom da mensagem do texto, como negações e advérbios intensificadores. Diferente dos sistemas propostos para a competição SemEval-2024 – que foram projetados para maximizar o desempenho preditivo, mas são modelos caixa-preta – o presente trabalho tem por objetivo contribuir para o melhor entendimento do fenômeno da persuasão em memes, através da análise de modelos construídos com o uso de aprendizado de máquina tradicional e atributos interpretáveis.

Esse artigo é uma versão estendida de um trabalho publicado anteriormente [de Azevedo and Gonçalves, 2025], incluindo uma nova etapa para realizar a seleção de atributos relevantes durante a fase de pré-processamento, a apresentação dos resultados atualizados considerando essa nova etapa, a apresentação de resultados para modelos treinados utilizando todas as combinações de grupos de meta-variáveis e uma discussão detalhada sobre todas as meta-variáveis utilizadas, com ênfase nas baseadas na tríade aristotélica. O restante do artigo está dividido da seguinte maneira. Os trabalhos relacionados são discutidos na Seção 2. A metodologia é apresentada na Seção 3 e os resultados na Seção 4. Finalizando com as conclusões e ideias para trabalhos futuros na Seção 5.

2 Trabalhos Relacionados

Nesta seção serão discutidos trabalhos recentes que utilizam meta-variáveis para a análise e classificação de persuasão assim como artigos sobre duas tarefas relacionadas, a análise de sentimentos e a identificação de discurso de ódio (Subse-

ção 2.1). Também serão apresentados os sistemas propostos pelas três equipes com o melhor desempenho na competição SemEval-2024 (Subseção 2.2).

2.1 Análise de Persuasão, de Sentimentos e Discurso de Ódio com Meta-Variáveis

Mohamad [2022] propõe um modelo integrado de análise retórica dos apelos aristotélicos (*logos*, *ethos* e *pathos*) de acordo com três domínios psicolinguísticos (pensamento analítico, influência, tom emocional) e seus níveis de escrita: lexical, sentencial e textual. Compreende-se os níveis lexical-frasal pela escolha das palavras nas frases (semântica), o sentencial pela análise estrutural da composição da frase (sintaxe) e textual pelo tipo de escrita e estilo usados. O estudo aplica esse modelo, composto por nove dispositivos retóricos, à resumos acadêmicos, onde demonstraram existir correlações significativas entre os três dispositivos identificados para cada apelo.

Cruz *et al.* [2019] utilizaram meta-variáveis estruturais, como número médio de caracteres por sentença e frequência de letras maiúsculas combinadas com *embeddings* semânticos para identificar frases de artigos de notícias contendo conteúdo persuasivo. O modelo proposto foi treinado em uma base de dados contendo 350 artigos e 16.956 frases (4.720 rotuladas como persuasivas e 12.245 rotuladas como não persuasivas), alcançando F1 de 0,595.

Tanto as meta-variáveis de Mohamad [2022] como as de Cruz *et al.* [2019] foram originalmente aplicadas de forma isolada em textos maiores e formais, de natureza acadêmica e jornalística, respectivamente. Neste artigo, investiga-se o desempenho das mesmas quando combinadas e aplicadas aos textos mais curtos e informais dos memes.

Carvalho and Plastino [2021] abordaram o problema da classificação de sentimentos utilizando um conjunto de 130 meta-variáveis, sendo 70 delas obtidas a partir de léxicos de análise de sentimentos. Em experimentos realizados em 22 bases de dados, classificadores Floresta Aleatória (*Random Forest*) treinados com meta-variáveis obtiveram acurácia média de 0,859, resultado superior ao obtido pelos modelos treinados com *embeddings* estáticos e n-gramas.

Vargas *et al.* [2025] discutem os desafios para a detecção de discurso de ódio e as limitações dos recursos disponíveis para linguagens como português, e introduzem o léxico *Moral Offensive Lexicon* (MOL), composto por expressões e termos ofensivos, voltado para a detecção de linguagem ofensiva em contextos digitais. O MOL foi originalmente construído a partir de mensagens em português coletadas no Instagram, relacionadas ao contexto político brasileiro, e posteriormente traduzido para o inglês, francês, espanhol e turco, com o apoio de tradutores nativos a fim de manter os sentidos morais e ofensivos das expressões para o idioma final. O léxico foi avaliado em experimentos em corpora anotados em Português, Espanhol e Inglês obtendo F1 de 0,880, 0,820 e 0,750, respectivamente.

2.2 Trabalhos com Melhor Desempenho no SemEval-2024 Task 4

Esta subseção destaca os três sistemas com melhor desempenho na sub tarefa 2b do SemEval-2024 Task 4. Li *et al.* [2024] obtiveram o melhor resultado de F1-macro (0,810) –

a métrica oficial da competição – utilizando um sistema em duas etapas. Na primeira, um modelo de linguagem *decoder-only* (LLaMA 2 7B) é utilizado para gerar um texto maior do que o texto original do meme, contendo uma explicação dada pela LLM sobre o seu significado. Na segunda etapa, os *embeddings* desse texto expandido são combinados em *embeddings* CLIP [Radford *et al.*, 2021] extraídos da imagem e do texto originais do meme para o aprendizado do modelo de classificação final.

A proposta de Anghelina *et al.* [2024] resultou em um sistema que também alcançou valor de 0,810 para a F1-macro. Esse sistema utiliza um modelo em que *embeddings* BERT [Devlin *et al.*, 2019] produzidos com o texto do meme são concatenados com *embeddings* ViT [Dosovitskiy *et al.*, 2020] extraídos das imagens e então passados para uma camada neural responsável por gerar a classificação de persuasão.

O sistema com o terceiro melhor desempenho (F1-macro de 0,809) foi proposto por Yu *et al.* [2024] e utiliza o modelo CLIP para extrair os *embeddings* das imagens e um ensemble de modelos pré-treinados (BERT, RoBERTa e outros) para gerar os *embeddings* dos textos. Os *embeddings* do texto e da imagem são então concatenados e submetidos a uma rede de três camadas responsável por realizar a classificação da persuasão.

Os três sistemas descritos nos parágrafos anteriores são modelos caixa-preta voltados para maximizar o desempenho preditivo, sem qualquer preocupação com questões relacionadas ao entendimento do fenômeno ou à interpretabilidade do processo de classificação. Além disso, todos utilizaram uma abordagem multimodal, extraindo *embeddings* das imagens e dos textos dos memes. Este artigo visa contribuir para o melhor entendimento do fenômeno da persuasão em memes através do uso de modelos treinados com diferentes grupos de meta-variáveis interpretáveis extraídas apenas dos textos dos memes. A metodologia do trabalho é apresentada na seção a seguir.

3 Metodologia

Esta seção tem como principal objetivo descrever as meta-variáveis utilizadas para treinar os classificadores binários de persuasão propostos neste trabalho (Subseções 3.4 e 3.5). Além disso, realiza-se uma breve apresentação da base de dados (Subseção 3.1) e das tarefas de pré-processamento básico do texto (Subseção 3.2), seleção de atributos (Subseção 3.3) e tratamento do desbalanceamento de classes (Subseção 3.6).

3.1 Base de Dados

As bases de dados para o treinamento e teste utilizadas nos experimentos foram providas pela equipe organizadora da competição SemEval-2024 Task 4 [Dimitrov *et al.*, 2024], que extraíram os textos das imagens e os pré-processaram manualmente para a remoção de erros gramaticais e para que frases em diferentes áreas da imagem fossem separadas por quebras de linhas. A base de treino é composta por 7.000 instâncias e a de teste por 1.000 instâncias. Cada uma possui uma *id* única, o texto extraído da imagem, os rótulos de persuasão pré-definidos (caso se trate de um meme persuasivo) e a URL fonte do meme (de onde é possível obter a imagem). Um exemplo de meme anotado como persuasivo na base é



Figura 1. Um exemplo de meme persuasivo da base de dados de treinamento

mostrado na Figura 1. Neste caso, a persuasão é feita com o intuito de realizar a associação entre o terrorismo e o desgosto em relação ao presidente americano. Informações completas sobre processo de anotação da base de dados podem ser obtidas em Piskorski *et al.* 2023.

É importante salientar que quase 82% da base de treino é composta por memes persuasivos. O apelo ao *ethos* é o mais recorrente (72%), seguido pelo *logos* (60%) e, por fim, o *pathos* (44%). A soma é superior a 100% porque trata-se originalmente de um problema de classificação de natureza multirrótulo, em que cada objeto (nesse caso, o texto de um meme) pode estar associado zero, uma ou mais classes [Dimitrov *et al.*, 2024; Gonçalves *et al.*, 2013].

3.2 Pré-processamento dos Textos

O pré-processamento textual constituiu na limpeza das quebras de linha indesejadas a fim de padronizar os dados para a etapa subsequente de extração de meta-variáveis. Após a limpeza, os textos foram processados pela *pipeline* treinada *en_core_web_lg* da biblioteca de processamento de linguagem natural spaCy¹. Essa *pipeline* foi responsável por converter cada texto em um objeto do tipo *Doc*, que encapsula anotações linguísticas como tokens, lemas, entidades nomeadas e relações de dependência sintática [Caseli and Nunes, 2024]. Essa estrutura foi fundamental para a extração posterior de características linguísticas e retóricas. Além da spaCy, a biblioteca NLTK² foi utilizada para segmentação de frases, por meio da função *sent_tokenize*, que apresentou desempenho superior ao do spaCy nessa tarefa específica. Também foi empregado o analisador de sentimentos VADER³ para a obtenção de escores de polaridade afetiva dos textos [Hutto and Gilbert, 2015]. Importante destacar que o texto lematizado não foi utilizado na análise de sentimentos, pois o VADER

¹<https://spacy.io/>

²<https://www.nltk.org/>

³<https://pypi.org/project/vaderSentiment/>

considera a morfologia verbal como parte de seu critério de classificação.

3.3 Seleção de Atributos

Após a extração inicial de 46 meta-variáveis distintas, foi realizado um processo de seleção de atributos com o objetivo de identificar o subconjunto de meta-variáveis mais relevantes para a tarefa de classificação.

Tendo em vista que existem 2^n possíveis subconjuntos a serem levados em consideração, onde n é o número total de meta-variáveis, uma avaliação completa de todos esses subconjuntos é computacionalmente inviável [da Silva *et al.*, 2015]. Para lidar com essa questão, foi adotada uma solução simplificada, em que os p-valorés de cada meta-variável em relação às classes "Persuasivo" e "Não Persuasivo" foram calculados utilizando a biblioteca *gtsummary* [Sjoberg *et al.*, 2021] da Linguagem R. Como resultado do teste, 6 meta-variáveis com p-valorés superiores a 0.001 foram removidas do conjunto final utilizado na construção dos modelos apresentados na Seção 4. Ao remover essas variáveis do experimento, espera-se aumentar a efetividade do modelo de classificação através da mitigação do ruído e redução da complexidade. As meta-variáveis removidas estão representadas em itálico nas tabelas descritivas da subseção a seguir.

3.4 Descrição das Meta-Variáveis

Para treinar os classificadores de persuasão, as 40 meta-variáveis escolhidas através do processo de seleção de atributos foram conceitualmente organizadas em quatro grupos distintos para compor a representação vetorial numérica dos textos (cada instância é transformada em um vetor numérico de 40 posições em que cada posição corresponde a uma meta-variável). São eles (i) grupo das meta-variáveis baseadas na tríade aristotélica, (ii) grupo das meta-variáveis baseadas em léxicos de análise de sentimentos e discurso de ódio, (iii) grupo das meta-variáveis estruturais e estatísticas e (iv) grupo das meta-variáveis contextuais. É importante ressaltar que o Grupo 2 compartilha meta-variáveis com dois grupos: 6 meta-variáveis com o Grupo 1 e 2 meta-variáveis com o Grupo 4 (estas últimas acabaram sendo removidas após a análise de p-valor).

3.4.1 Grupo 1: Meta-Variáveis da Retórica Aristotélica

De acordo com Aristóteles and Kennedy [1991] apud Mohamad [2022], o apelo para a emoção (*pathos*) é feito por meios capazes de gerar reação emocional, como palavras e frases emotivas, figuras de linguagem e narrativas; o apelo para a ética (*ethos*) é refletido pelo uso de linguagem dinâmica e direcionada ao interlocutor para demonstrar a credibilidade e competência do locutor; e o apelo para a lógica (*logos*) é usado para induzir uma resposta racional com, por exemplo, o uso de citações de autoridades, estatísticas e linguagem formal.

Para detectar o uso da tríade aristotélica no texto, a extração das características foi realizada seguindo as referências do modelo proposto por Mohamad [2022] e adaptadas ao contexto do problema. As características foram agrupadas segundo seus respectivos níveis linguísticos (lexical-frasal, sentencial e textual). A seguir, descrevem-se os critérios ado-

tados:

Pathos

- Frases emotivas: frequência de lemas positivos e negativos (gerais, substantivos e adjetivos), com base no léxico de sentimentos de Hu and Liu [2004].
- Uso da voz passiva: avaliação binária da presença de sujeitos passivos (etiqueta *nsubjpass*) em verbos no particípio (VBN) com verbos auxiliares como dependentes.
- Sentimento: escores positivos, neutros, negativos e compostos obtidos via análise da abordagem VADER [Hutto and Gilbert, 2015].

Ethos

- Uso de pronomes: detecção de frases cujo sujeito seja um pronome, a partir das etiquetas *nsubj* e *PRP*.
- Tempo verbal futuro: presença de estruturas com sujeito (*nsubj*), verbo principal e verbo auxiliar futuro como filhos do mesmo núcleo verbal.
- Personalismo: avaliação binária da presença de advérbios e conjunções subordinativas, como marcadores de subjetividade.

Logos

- Dados numérico-nominais: detecção de tokens com etiqueta *CD* (número cardinal) seguidos diretamente por tokens nominais (*NN*, *NNS* etc.), assim como citações a pessoas e organizações.
- Frases complexas e longas: presença de orações compostas, com estruturas sintáticas envolvendo sujeito, verbo, adjuntos adverbiais e conectores subordinativos ou coordenativos, identificadas via *DependencyParser*.
- Formalidade: caracterizada pelo uso de voz passiva, construção em terceira pessoa e frases declarativas.

3.4.2 Grupo 2: Meta-Variáveis Geradas por Léxicos de Análise de Sentimentos e Detecção de Discurso de Ódio

O segundo conjunto foi composto por meta-variáveis de contabilização de frequências utilizando o léxico para análise de sentimentos de Hu and Liu [2004] e o léxico MOL [Vargas *et al.*, 2025] para a identificação de discurso de ódio. O principal objetivo em avaliar esse grupo foi identificar se meta-variáveis tradicionalmente utilizadas para a classificação de sentimentos e discurso de ódio também são efetivas para a classificação de persuasão.

Na análise de sentimentos foram computados separadamente o total de palavras e de classes gramaticais (substantivos, adjetivos, advérbios e verbos) presentes no texto de acordo com o léxico proposto por [Hu and Liu, 2004]. Cada palavra foi reduzida para o seu lema para uma contagem mais precisa.

Para gerar as meta-variáveis relacionadas ao discurso de ódio, dividiu-se o léxico MOL em termos e expressões conforme a anotação feita pelos próprios autores. Feito isso, identificou-se a presença dos elementos lexicais no texto por meio dos unigramas, bigramas e trigramas extraídos de cada texto. Dos unigramas identificados como ofensivos, suas classes gramaticais foram contadas.

3.4.3 Grupo 3: Meta-Variáveis Estruturais e Estatísticas

Este grupo é composto por 9 meta-variáveis contadoras e estatísticas, como total de frases e frequência de pontuação, utilizadas por Cruz *et al.* [2019] para a detecção de persuasão em frases de artigos de notícias. É importante ressaltar que algumas variáveis apresentam potencial correlação entre si como, por exemplo, o total de frases e a frequência de pontuação. Apesar disso, optou-se por mantê-las no experimento para manter a abordagem empregada por Cruz *et al.* [2019].

3.4.4 Grupo 4: Meta-Variáveis Contextuais

Para o último grupo, foram selecionadas as seguintes 4 meta-variáveis descritas em Carvalho and Plastino [2021]. Elas foram escolhidas e adaptadas para o contexto do trabalho atual pela potencial explicabilidade. São elas:

- Total de palavras de negação: contagem de palavras como “no”, “nothing” e “never”;
- Total de contextos negados: onde um contexto negado é definido pelo segmento de frase iniciado pela palavra de negação e finalizado na primeira pontuação após a palavra;
- Total de advérbios positivos e total de advérbios negativos: as duas meta-variáveis desta classe transmitem circunstâncias como tempo e intensidade, permitindo ao autor expressar ideias de forma mais precisa e persuasiva.

As Tabelas 1, 2, 3 e 4 apresentam, respectivamente, a relação completa de meta-variáveis dos Grupos 1, 2, 3 e 4. As presentes em mais de um grupo estão destacadas em negrito e as que foram removidas após o processo de seleção de variáveis descrito na Subseção 3.3 estão destacadas em itálico.

3.5 Exemplo de Extração de Meta-Variáveis

A seguir, apresenta-se um exemplo que ilustra como são obtidos os valores das metas-variáveis a partir do texto original de um meme. O exemplo em questão utiliza o texto de um meme persuasivo da base de dados de treino e mostra algumas das meta-variáveis que foram extraídas a partir dele.

*CROOKS & LIARS AS PRESIDENTS? WASN'T ONE CLINTON ENOUGH?*⁴

- **total_neg_nn (Grupo 2, Sentimento):** O valor é 2, contabilizado após a verificação dos lemmas dos substantivos “CROOKS” e “LIARS” estarem presentes no léxico de palavras negativas de Hu and Liu [2004].
- **subjectivity (Grupo 1, Ethos):** O valor é 1, indicando uso de linguagem subjetiva pelo advérbio “ENOUGH”.
- **numerical_nominal (Grupo 1, Logos):** O valor é 1, sinalizado pela combinação do número cardinal “ONE” seguido do substantivo pessoal “CLINTON”.
- **person_named (Grupo 1, Logos):** O valor é 1, extraído da entidade nomeada “CLINTON” (sobrenome reconhecido de um chefe de Estado).

⁴Exemplo (ID: 69660) retirado da base de treinamento fornecida por Dimitrov *et al.* [2024].

3.6 Balanceamento de Classes

Para lidar com o desbalanceamento de classes da base de dados fornecida pela competição, que possui mais de 80% dos memes da classe “Persuasivo”, foi aplicada a técnica *Synthetic Minority Over-sampling Technique* (SMOTE) [Chawla *et al.*, 2002] na base de treinamento gerada após a extração e seleção das meta-variáveis. O desbalanceamento pode levar a modelos que tendem a classificar a maioria dos exemplos na classe majoritária, prejudicando o desempenho na detecção da classe minoritária (neste caso, favorecendo a classe “Persuasivo” sobre “Não Persuasivo”).

A abordagem adotada pelo SMOTE consiste em gerar dados sintéticos da classe minoritária por interpolação de características entre exemplos vizinhos da mesma classe, em vez de simplesmente duplicar exemplos existentes, evitando assim o *overfitting*. No presente trabalho, o SMOTE foi aplicado até que a classe “não persuasivo” representasse 50% dos objetos da base.

É importante reportar que também foi avaliada uma outra estratégia de balanceamento de classes que consistiu na inclusão da coleção de slogans e expressões políticas não persuasivas presentes no corpus de Guerini *et al.* [2015]. Porém, os resultados foram muito inferiores aos obtidos com a técnica SMOTE.

4 Resultados

Esta seção reporta e discute os resultados de dois experimentos realizados para a avaliação dos grupos de meta-variáveis. No primeiro, apresenta-se o desempenho preditivo de modelos de classificação treinados com os algoritmos Floresta Aleatória (FA) e Regressão Logística (RL) [Witten *et al.*, 2011] para cada grupo isoladamente, com o objetivo de identificar qual deles é o mais eficaz para a classificação de persuasão. No segundo experimento, os modelos foram treinados combinando os grupos de meta-variáveis. As implementações de FA e RL avaliadas no experimento são as disponibilizadas pelo pacote *scikit-learn*⁵ da linguagem Python e os modelos foram treinados com os hiperparâmetros padrão desses dois algoritmos.

4.1 Experimento 1 - Grupos Isolados

A Tabela 5 apresenta os resultados de F1-macro obtidos pelos modelos treinados com cada grupo de meta-variáveis isoladamente (F1-macro foi o critério utilizado na competição SemEval 2024 Task 4). São apresentados resultados obtidos sem e com o balanceamento de classes através da técnica SMOTE.

Conforme mostra a tabela, em geral os modelos FA obtiveram desempenho superior ao dos modelos RL (exceto para o Grupo 4, que é composto por um número muito reduzido de variáveis). Os melhores valores de F1-macro foram obtidos pelos modelos FA treinados com o Grupo 1 (retórica aristotélica), que alcançou um F1-macro acima de 60%, seguido pelo Grupo 3 (meta-variáveis estruturais). O melhor desempenho geral foi obtido pelo modelo treinado com o Grupo 1 e FA + SMOTE (F1-macro de 0,645).

Por outro lado, é possível observar que o modelo treinado com as meta-variáveis do Grupo 2 – geradas por léxicos

⁵www.scikit-learn.org

Tabela 1. Descrição das 25 meta-variáveis do Grupo 1 (retórica aristotélica).

Meta-variável	Descrição
total_pos_w	Total de palavras positivas.
total_pos_jj	Total de adjetivos positivos.
total_pos_nn	Total de substantivos positivos.
total_neg_w	Total de palavras negativas.
total_neg_jj	Total de adjetivos negativos.
total_neg_nn	Total de substantivos negativos.
pos_sentiment_polarity	Escore de sentimento positivo (VADER).
neu_sentiment_polarity	Escore de sentimento neutro (VADER).
neg_sentiment_polarity	Escore de sentimento negativo (VADER).
comp_sentiment_polarity	Escore composto geral de sentimento (VADER).
has_passive	Indica presença de voz passiva (1 ou 0).
pronoun_sent	Indica se há frases cujo sujeito é um pronome (1 ou 0).
future_tense	Indica presença de tempo verbal no futuro (1 ou 0).
has_wh	Indica presença de pronomes interrogativos (1 ou 0).
subjecitivity	Presença de linguagem subjetiva (1 ou 0).
conditional	Indica presença de conjunções condicionais (1 ou 0).
numerical_nominal	Indica expressões numéricas com substantivos (1 ou 0).
organization_named	Total de entidades nomeadas do tipo organização.
person_named	Total de entidades nomeadas do tipo pessoa.
has_complex_sent	Indica presença de orações compostas (1 ou 0).
sum_log_sent	Logaritmo natural da soma dos caracteres do texto.
third_p_count	Total de pronomes na terceira pessoa.
has_period	Total de pontos finais.
<i>has_exclamation</i>	Total de pontos de exclamação.
<i>has_interrogation</i>	Indica se há interrogação (1 ou 0).

Tabela 2. Descrição das 16 meta-variáveis do Grupo 2 (léxicos de análise de sentimentos e discurso de ódio).

Meta-variável	Descrição
total_pos_w	Total de palavras positivas.
total_pos_jj	Total de adjetivos positivos.
total_pos_nn	Total de substantivos positivos.
total_neg_w	Total de palavras negativas.
total_neg_jj	Total de adjetivos negativos.
total_neg_nn	Total de substantivos negativos.
total_pos_vb	Total de verbos positivos.
total_neg_vb	Total de verbos negativos.
total_hate_jj	Total de adjetivos ofensivos (MOL).
total_hate_nn	Total de substantivos ofensivos (MOL).
total_hate_rb	Total de advérbios ofensivos (MOL).
total_hate_vb	Total de verbos ofensivos (MOL).
hate_speech_terms_total	Total de palavras ofensivas utilizadas (MOL).
<i>hate_speech_expressions_total</i>	Total de expressões ofensivas (MOL).
total_pos_rb	Total de advérbios positivos (MOL).
total_neg_rb	Total de advérbios negativos.

de análise de sentimentos e discurso de ódio – obtive desempenho inferior. Esse resultado sugere que meta-variáveis tradicionalmente utilizadas nestes dois problemas podem não ser boas preditoras para classificadores de persuasão, ao menos quando utilizadas isoladamente.

4.2 Experimento 2 - Grupos Combinados

A Tabela 6 apresenta os resultados obtidos por modelos treinados com combinações de grupos de meta-variáveis. A tabela indica que os modelos treinados com o algoritmo FA superaram os modelos RL em todas as situações, seja sem ou com o

emprego do balanceamento de classes via SMOTE.

O classificador que obteve o melhor resultado geral (F1-macro de 0,701) utilizou as meta-variáveis dos Grupos 1 e 3 combinadas, (retórica aristotélica + meta-variáveis estruturais) e foi treinado com o algoritmo FA utilizando o SMOTE. Observe ainda que outros modelos treinados com as meta-variáveis dos Grupos 1 e 3 com FA + SMOTE também obtiveram desempenho destacado (Grupos 1 + 2 + 3, Grupos 1 + 3 + 4 e Grupos 1 + 2 + 3 + 4). Conforme mostrado na subseção anterior, os Grupos de meta-variáveis 1 e 3 já haviam se destacado isoladamente no primeiro experimento.

Tabela 3. Descrição das 9 meta-variáveis do Grupo 3 (meta-variáveis estruturais).

Meta-variável	Descrição
total_of_sentences	Total de frases.
char_mean_sentence	Média de caracteres por frase.
char_variance_sentence	Variância de caracteres por frase.
total_char_text	Total de caracteres no texto.
char_variance_word	Variância de caracteres por palavra.
punctuation_frequency	Frequência de pontuação no texto.
upper_case_frequency	Frequência de letras maiúsculas no texto.
lemma_per_token	Proporção de tokens sobre palavras lematizadas.
char_mean_word	Média de caracteres por palavra.

Tabela 4. Descrição das 4 meta-variáveis do Grupo 4 (meta-variáveis contextuais).

Meta-variável	Descrição
total_negation_words	Total de palavras de negação.
total_negated_context	Total de contextos negados.
total_pos_rb	Total de advérbios positivos.
total_neg_rb	Total de advérbios negativos.

Tabela 5. Desempenho dos modelos treinados com os grupos de meta-variáveis isolamente (F1-macro)

Grupo de Meta-Variáveis	FA	RL	FA + SMOTE	RL + SMOTE
Grupo 1 (23 meta-variáveis)	0,602	0,565	0,645	0,608
Grupo 2 (13 meta-variáveis)	0,458	0,458	0,581	0,581
Grupo 3 (8 meta-variáveis)	0,587	0,515	0,613	0,599
Grupo 4 (2 meta-variáveis)	0,458	0,458	0,393	0,393

É interessante ainda observar que as meta-variáveis do Grupo 2 (léxicos de análise de sentimentos e discurso de ódio), quando combinadas com as dos Grupos 1 e 3 levaram a modelos com desempenho preditivo superior aos dos classificadores treinados com esses grupos isoladamente. Por exemplo, o valor de F1-macro do modelo treinado utilizando as meta-variáveis do Grupo 3 isoladamente com FA + SMOTE subiu de 0,613 para 0,640 quando as meta-variáveis do Grupo 2 foram acrescentadas.

Por fim, é importante observar que apesar dos bons desempenhos das combinações de trios e quadra, eles foram inferiores à combinação em par do Grupo 1 + 3, conforme mostrado nas 5 últimas linhas da tabela.

4.3 Discussão

O modelo proposto neste artigo teria obtido a 15ª colocação na subtarefa 2b do SemEval-2024 Task 4 com um F1-macro de 0,701, diferença de cerca de 11% em relação ao resultado dos modelos vencedores (os resultados completos podem ser consultados em Dimitrov *et al.* [2024]).

Entretanto, é importante observar que não apenas as equipes campeãs, mas todos os competidores propuseram modelos que visavam maximizar o desempenho preditivo, sem se importar em entender o fenômeno da persuasão em memes (ou seja, não houve tentativa de identificar as variáveis que melhor caracterizam um texto de meme persuasivo). Além disso, o classificador proposto neste artigo possui custo computacional baixo e nem mesmo utiliza a imagem do meme no processo de treinamento. Dentre todos os competidores, Takahashi [2024] foi o único que também não utilizou a imagem, porém em um modelo baseado em redes neurais em grafos (obtendo F1-macro de 0,714), o que compromete o custo e

interpretabilidade.

5 Conclusão

Este trabalho explorou abordagens interpretáveis para a detecção de persuasão em textos de memes, utilizando meta-variáveis como alternativa aos modelos neurais amplamente utilizados pelas equipes que disputaram a tarefa 4, subtarefa 2b da competição SemEval-2024. Baseando-se em estudos existentes, foram definidos 4 grupos de meta-variáveis fundamentadas na retórica clássica aristotélica, em léxicos afetivos e ofensivos, nas características linguísticas estruturais e nas características linguísticas contextuais.

A combinação das variáveis retóricas (Grupo 1) com as variáveis estruturais (Grupo 3) produziu o melhor desempenho entre os experimentos, em especial após o balanceamento da base por meio da técnica SMOTE. Destaca-se também o desempenho superior dos modelos de Floresta Aleatória em relação à Regressão Logística, indicando que, de modo geral, as variáveis não seguem uma relação linear com o rótulo de persuasão. Apesar de não alcançar os níveis de desempenho preditivo dos modelos baseados em redes neurais profundas, a perspectiva proposta oferece maior interpretabilidade e menor custo computacional, tornando-a uma alternativa viável para contextos com recursos limitados ou que demandam interpretabilidade nos critérios de decisão.

Como trabalhos futuros, pretende-se explorar novas técnicas de seleção de atributos, buscando identificar o subconjunto de meta-variáveis capaz de maximizar o desempenho preditivo do modelo de classificação. Também se considera o emprego de novas meta-variáveis baseadas em estudos de outras áreas de pesquisa, como a identificação de notícias fal-

Tabela 6. Desempenho dos dois melhores modelos treinados combinando grupos de meta-variáveis (F1-macro)

Grupo de Meta-Variáveis	FA	RL	FA + SMOTE	RL + SMOTE
Grupos 1+2 (30 meta-variáveis)	0,603	0,569	0,646	0,601
Grupos 1+3 (31 meta-variáveis)	0,659	0,610	0,701	0,630
Grupos 1+4 (25 meta-variáveis)	0,599	0,572	0,630	0,617
Grupos 2+3 (21 meta-variáveis)	0,635	0,552	0,640	0,631
Grupos 2+4 (15 meta-variáveis)	0,463	0,458	0,615	0,615
Grupos 3+4 (10 meta-variáveis)	0,588	0,538	0,610	0,597
Grupos 1+2+3 (38 meta-variáveis)	0,657	0,609	0,696	0,632
Grupos 1+2+4 (32 meta-variáveis)	0,614	0,572	0,647	0,614
Grupos 1+3+4 (33 meta-variáveis)	0,661	0,607	0,696	0,633
Grupos 2+3+4 (23 meta-variáveis)	0,617	0,560	0,637	0,637
Grupos 1+2+3+4 (40 meta-variáveis)	0,659	0,611	0,688	0,644

sas [de Moraes *et al.*, 2020; Vargas *et al.*, 2023] e a utilização de léxicos adicionais de análise de sentimentos, como os avaliados por Carvalho and Plastino [2021], visando enriquecer a representação linguística dos textos.

Declarações complementares

Agradecimentos

Este trabalho foi apoiado pelo Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da Escola Nacional de Ciências Estatísticas (ENCE-IBGE).

Financiamento

Ana Beatriz Soares de Azevedo foi parcialmente financiada por bolsa Iniciação Científica concedida pelo CNPq (Processo: 800183/2024-0)

Contribuições dos autores

Eduardo Corrêa Gonçalves contribuiu para a revisão bibliográfica e concepção deste estudo. Ana Beatriz Soares de Azevedo é a principal contribuidora e escritora, tendo desenvolvido a metodologia e realizado os experimentos. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses

Disponibilidade de dados e materiais

Os conjuntos de dados utilizados e scripts produzidos durante o estudo atual estão disponíveis em https://github.com/fullon-Bia/pibic_persuasao.

Referências

Anghelina, I., Buță, G., and Enache, A. (2024). SuteAlbastre at SemEval-2024 task 4: Predicting propaganda techniques in multilingual memes using joint text and vision transformers. In *Proc. of the 18th Int'l. Wksh on Semantic Evaluation (SemEval-2024)*, pages 443–449. ACL. DOI: 10.18653/v1/2024.semeval-1.68.

Aristóteles and Kennedy, G. A. (1991). *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press.

Braca, A. and Dondio, P. (2023). Developing persuasive systems for marketing: the interplay of persuasion techniques, customer traits and persuasive message design. *Ital. J. Mark*, 2023:369—412. DOI: 10.1007/s43039-023-00077-0.

Carvalho, J. and Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artif. Intell. Rev.*, 54(3):1887—1936. DOI: 10.1007/s10462-020-09895-6.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3 edition.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357. DOI: 10.1613/jair.953.

Cruz, A. F., Rocha, G., and Cardoso, H. L. (2019). On sentence representations for propaganda detection: From hand-crafted features to word embeddings. In *Proc. of the 2nd Wkshp on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 107–112. ACL. DOI: 10.18653/v1/D19-5015.

da Silva, P. N., Gonçalves, E. C., Rios, E. H., Muhammad, A., Moss, A., Pritchard, T., Glassborow, B., Plastino, A., and de Vasconcellos Azeredo, R. B. (2015). Automatic classification of carbonate rocks permeability from 1h nmr relaxation data. *Expert Systems with Applications*, 42(9):4299–4309. DOI: 10.1016/j.eswa.2015.01.034.

de Azevedo, A. B. S. and Gonçalves, E. C. (2025). An evaluation of meta-features for automated detection of persuasion in texts of political memes. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 145–152. SBC. DOI: 10.5753/kdmile.2025.247776.

de Moraes, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon Jr, S. (2020). A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in brazilian portuguese. *iSys - Brazilian Journal of Information Systems*, 13(4):126–149. DOI: 10.5753/isys.2020.833.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL. DOI: 10.18653/v1/N19-1423.

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., and Da San Martino, G. (2024). SemEval-2024 task 4: Multilingual detection of persuasion techniques

- in memes. In *Proc. of the 18th Int'l. Wkshp on Semantic Evaluation (SemEval-2024)*, pages 2009–2026. ACL. DOI: 10.18653/v1/2024.semeval-1.275.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houselby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929. DOI: 10.48550/arXiv.2010.11929.
- Gonçalves, E. C., Plastino, A., and Freitas, A. A. (2013). A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, pages 469–476. IEEE. DOI: 10.1109/ICTAI.2013.76.
- Guerini, M., Özbal, G., and Strapparava, C. (2015). Echoes of persuasion: The effect of euphony in persuasive communication. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1483–1493. ACL. DOI: 10.3115/v1/N15-1172.
- Halversen, A. and Weeks, B. E. (2023). Memeing politics: Understanding political meme creators, audiences, and consequences on social media. *Social Media + Society*, 9(4):20563051231205588. DOI: 10.1177/20563051231205588.
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proc. of the 10th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, page 168–177. ACM. DOI: 10.1145/1014052.1014073.
- Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. of the 8th Int'l. Conf. on Weblogs and Social Media*, pages 216–225. PKP/PS. DOI: 10.1609/icwsm.v8i1.14550.
- Li, S., Wang, Y., Yang, L., Zhang, S., and Lin, H. (2024). LMEME at SemEval-2024 task 4: Teacher student fusion - integrating CLIP with LLMs for enhanced persuasion detection. In *Proc. of the 18th Int'l. Wkshp on Semantic Evaluation (SemEval-2024)*, pages 628–633. ACL. DOI: 10.18653/v1/2024.semeval-1.92.
- Mohamad, H. A. (2022). Analysis of rhetorical appeals to logos, ethos and pathos in ENL and ESL research abstracts. *Malaysian Journal of Social Sciences and Humanities*, 7(3). DOI: 10.47405/mjssh.v7i3.1314.
- Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2). DOI: 10.1145/3597307.
- Piskorski, J., Stefanovitch, N., Bausier, V.-A., Faggiani, N., Linge, J., Kharazi, S., Nikolaidis, N., Teodori, G., De Longueville, B., Doherty, B., *et al.* (2023). News categorization, framing and persuasion techniques: Annotation guidelines. Technical Report JRC132862, European Commission, Ispra.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proc. of the 38th Int'l. Conf. on Machine Learning*, volume 139, pages 8748–8763. PMLR.
- Santos, J., Alves, A., and Gonçalo Oliveira, H. (2019). Asapppy: a python framework for portuguese STS. In *Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symposium in Information and Human Language Technology (STIL 2019)*, volume 2583, pages 14–26. CEUR-WS.org.
- Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., and Larmarange, J. (2021). Reproducible summary tables with the gtsummary package. *The R Journal*, 13:570–580. DOI: 10.32614/RJ-2021-053.
- Takahashi, H. (2024). Hidetsune at SemEval-2024 task 4: An application of machine learning to multilingual propagandistic memes identification using machine translation. In *Proc. of the 18th Int'l. Wkshp on Semantic Evaluation (SemEval-2024)*, pages 370–373. ACL. DOI: 10.18653/v1/2024.semeval-1.57.
- Vargas, F., Carvalho, I., Pardo, T. A. S., and Benevenuto, F. (2025). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456. DOI: 10.1017/nlp.2024.18.
- Vargas, F., Jaidka, K., Pardo, T., and Benevenuto, F. (2023). Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206. INCOMA Ltd., Shoumen, Bulgaria".
- Weiss, M. C. (2019). Sociedade sensorizada: a sociedade da transformação digital. *Estudos Avançados*, 33(95):203–214. DOI: 10.1590/s0103-4014.2019.3395.0013.
- Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers Inc., 3rd edition.
- Yu, E., Wang, J., Qiao, X., Qi, J., Li, Z., Lin, H., Zong, L., and Xu, B. (2024). DUTIR938 at SemEval-2024 task 4: semi-supervised learning and model ensemble for persuasion techniques detection in memes. In *Proc. of the 18th Int'l. Wkshp on Semantic Evaluation (SemEval-2024)*, pages 642–648. ACL. DOI: 10.18653/v1/2024.semeval-1.94.