An Evaluation of Meta-Features for Automated Detection of Persuasion in Texts of Political Memes

Ana B. S. de Azevedo[1], Eduardo C. Gonçalves[1]

Escola Nacional de Ciências Estatísticas (ENCE-IBGE), Brazil
anabeatriz98azevedo@gmail.com, eduardo.correa@ibge.gov.br

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 used the SemEval-2024 Task 4 dataset with 7,000 training and 1,000 testing instances. Random Forest and Logistic Regression were evaluated with and without SMOTE to handle class imbalance present on the training dataset. The best result, with a macro-F1 value of 0.698, was achieved by combining rhetorical and structural features. The proposed approach offers a computationally efficient and interpretable alternative to neural-based models.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

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

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 multirrótulo hierárquica (identificar o conjunto de técnicas de persuasão presentes em memes persuasivos) e a 2b voltada para classificação binária (identificar se o 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 [Ferreira and Paraboni 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 no 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 [Carvalho and Plastino 2021] que podem ser extraídas do conteúdo textual dos memes para a classificação de memes persuasivos. O trabalho investiga quatro diferentes grupos de meta-variáveis. O primeiro grupo é baseado na tríade retórica aristotélica que correspondem ao uso de apelos à lógica (logos), ética (ethos) e emoção (pathos) 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.

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

Esta seção apresenta os trabalhos sobre classificação com meta-variáveis mais diretamente relacionados ao método proposto no presente artigo (Subseção 2.1) e os sistemas das três equipes com o melhor desempenho na competição SemEval-2024 (Subseção 2.2).

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

[Carvalho and Plastino 2021] abordaram o problema da classificação de sentimentos utilizando 130 meta-variáveis, sendo 70 delas obtidas a partir de léxicos de sentimentos. Em experimentos realizados em 22 bases de dados, classificadores Floresta Aleatória (*Random Forest*) treinados com meta-variáveis obtiveram desempenho superior aos 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 MOL (Moral Offensive Lexicon), 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.

2.2 Trabalhos com Melhor Desempenho no SemEval-2024 Task 4

Esta subseção destaca os três sistemas com melhor desempenho na subtarefa 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 (combinados em um processo de pooling). Os embeddings do texto e da imagem são então concatenados e submetidos a uma rede de 3 camadas responsável por realizar a classificação da persuasão.

Os três sistemas descritos acima 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 de persuasão propostos neste trabalho (Subseção 3.3). Além disso, realiza-se uma breve apresentação da base de dados (Subseção 3.1) e das operações de pré-processamento (Subseção 3.2).

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 e a URL fonte do meme (de onde é possível obter

a imagem). Um exemplo de meme anotado como persuasivo na base é mostrado na Figura 1. Neste caso, a persuasão é feita com o intuito de de questionar a reputação de uma instituição (o governo).

Fig. 1. Um exemplo de meme persuasivo

É importante salientar que quase 82% da base de treino é composta por memes persuasivos. O apelo é o 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 multirrótulo).

3.2 Pré-processamento

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 utilizando a pipeline treinada *en_core_web_lg* da biblioteca spaCy[1]. 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. Essa estrutura foi fundamental para a extração posterior de características linguísticas e retóricas. Além da spaCy, a biblioteca NLTK[2] 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 VADER[3] para a obtenção de escores de polaridade afetiva dos textos. Importante destacar que o texto lematizado não foi utilizado na análise de sentimentos, pois o VADER considera a morfologia verbal como parte de seu critério de classificação.

Tendo em vista que a base de dados fornecida pela competição é desbalanceada, com mais de 80% dos memes da classe "persuasivo", após a geração das meta-variáveis, a técnica SMOTE [Chawla et al. 2002] foi aplicada na base de treinamento para criar dados sintéticos da classe minoritária (textos não persuasivos) até que esta representasse 50% dos objetos da base.

3.3 Descrição das Meta-Variáveis

Para treinar os classificadores de persuasão, foram elaborados quatro conjuntos distintos de meta-variáveis para compor a representação vetorial dos textos. São elas (i) baseadas na tríade aristotélica, (ii) baseadas em léxicos de análise de sentimentos e discurso de ódio, (iii) meta-variáveis estruturais e estatísticas e (iv) meta-variáveis contextuais.

[1] `https://spacy.io/`
[2] `https://www.nltk.org/`
[3] `https://pypi.org/project/vaderSentiment/`

3.3.1 *Grupo 1: Meta-Variáveis da Retórica Aristotélica.* Para detectar o uso da tríade aristotélica no texto, a extração 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, sentencial e textual). A seguir, descrevem-se os critérios adotados:

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: scores positivos, neutros, negativos e compostos obtidos via análise do VADER.

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 interrogações, 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 ou imperativas.

3.3.2 *Grupo 2: Meta-Variáveis Geradas por Léxicos de Análise de Sentimentos e Detecção de Discurso de Ódio.* O segundo conjunto foi composto por 16 meta-variáveis frequentistas utilizando o léxico de sentimentos de [Hu and Liu 2004] e o léxico MOL [Vargas et al. 2025] para a identificação de linguagem de ódio. O principal objetivo em avaliar esse grupo foi identificar se meta-variáveis tradicionalmente utilizadas para a classificação de de sentimentos e discurso de ódio também são efetivas para a classificação de persuasão.

Na análise de sentimentos foi computado separadamente o total de palavras e 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 categorização dos 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.

3.3.3 *Grupo 3: Meta-Variáveis Estruturais e Estatísticas.* Este grupo é composto por 9 variáveis frequentistas e estatísticas, como total de frases e frequência de pontuação, utilizadas por [Cruz et al. 2019] para a detecção de persuasão textos 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].

3.3.4 *Grupo 4: Meta-Variáveis Contextuais.* Para o último grupo, foram adotadas variáveis descritas em [Carvalho and Plastino 2021], 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 negativos: transmite circunstâncias como tempo e intensidade, permitindo ao autor expressar ideias de forma mais precisa e persuasiva.

A relação completa de meta-variáveis encontra-se disponível em `https://github.com/edubd/bd/blob/main/meta_variaveis_persuasao.pdf`, sendo os Grupo 1, 2, 3 e 4 compostos por 22, 16, 9 e 4 meta-variáveis, respectivamente. Algumas pertencem a mais de um grupo, como é o caso do "total de palavras positivas", que pertence aos Grupos 1 e 2.

4. RESULTADOS

Esta seção reporta 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) 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[4] da linguagem Python e os modelos foram treinados com os hiperparâmetros padrão.

A Tabela I 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 do 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) e com o Grupo 3 (Meta-Variáveis Estruturais) com F1 acima de 60%. O melhor desempenho geral foi obtido pelo modelo treinado com o Grupo 1 e FA + SMOTE (F1-macro de 0,647). Por outro lado, é possível observar que o modelo treinado com as meta-variáveis do Grupo 2 – geradas por léxicos de análise de sentimentos e discurso de ódio – obteve 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.

Grupo de Meta-Variáveis	FA	RL	FA + SMOTE	RL + SMOTE
Grupo 1 (22 meta-variáveis)	0,624	0,558	**0,647**	0,604
Grupo 2 (16 meta-variáveis)	0,458	0,458	0,561	0,559
Grupo 3 (9 meta-variáveis)	0,612	0,531	0,605	0,602
Grupo 4 (4 meta-variáveis)	0,458	0,458	0,403	0,403

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

A Tabela II apresenta os resultados obtidos por modelos treinados com pares de grupos de meta-variáveis combinados. A tabela mostra que os modelos treinados com o algoritmo FA superaram os

[4] `www.scikit-learn.org`

modelos RL em todas as situações. O classificador que obteve o melhor resultado geral (F1-macro de 0,698) 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. Esses mesmos grupos já haviam se destacado individualmente no primeiro experimento. 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 F1-macro do modelo treinado utilizando as meta-variáveis do Grupo 3 isoladamente com RA + SMOTE subiu de 0,605 para 0,672 quando as meta-variáveis do Grupo 2 foram acrescentadas. Por fim, é importante reportar que também foram realizados experimentos combinando três grupos e os quatro grupos, porém os resultados não foram superiores aos obtidos pelos modelos relacionados na Tabela II.

Grupo de Meta-Variáveis	FA	RL	FA + SMOTE	RL + SMOTE
Grupos 1+2 (32 meta-variáveis)	0,625	0,558	0,651	0,610
Grupos 1+3 (31 meta-variáveis)	0,644	0,590	**0,698**	0,629
Grupos 1+4 (26 meta-variáveis)	0,622	0,558	0,653	0,599
Grupos 2+3 (25 meta-variáveis)	0,620	0,557	0,672	0,630
Grupos 2+4 (20 meta-variáveis)	0,463	0,458	0,582	0,550
Grupos 3+4 (13 meta-variáveis)	0,590	0,514	0,641	0,605

Table II. Desempenho dos modelos treinados combinando pares de grupos de meta-variáveis (F1-macro)

O modelo proposto neste artigo teria obtido a 15ª colocação na subtarefa 2b do SemEval-2024 Task 4, com resultado um pouco acima da média de 0,691 para o F1-macro, computada considerando o resultado de todas as equipes que participaram do desafio. Já a diferença de F1-macro em relação ao resultado dos modelos vencedores é de cerca de 11% (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 (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. CONCLUSÕES E TRABALHOS FUTUROS

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 de treino 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 avaliar uma estratégia de balanceamento de classes baseada na inclusão da coleção de textos políticos não persuasivos do léxico de [Guerini et al. 2015]. Também

pretende-se fazer a análise da importância das features no modelo FA e dos pesos do modelo RL com o objetivo de identificar o subconjunto de meta-variáveis mais relevantes para a classificação, idependente do grupo a qual pertençam. Por fim, intenciona-se realizar uma análise qualitativa dos erros de classificação com o intuito de identificar padrões ou lacunas relevantes, contribuindo para o avanço do entendimento sobre o problema.

REFERENCES

Anghelina, I., Buță, G., and Enache, A. 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)*. ACL, Mexico City, Mexico, pp. 443–449, 2024.

Braca, A. and Dondio, P. Developing persuasive systems for marketing: the interplay of persuasion techniques, customer traits and persuasive message design. *Ital. J. Mark* vol. 2023, pp. 369–412, 2023.

Carvalho, J. and Plastino, A. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artif. Intell. Rev.* 54 (3): 1887–1936, 2021.

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

Cruz, A. F., Rocha, G., and Cardoso, H. L. On sentence representations for propaganda detection: From handcrafted features to word embeddings. In *Proc. of the 2nd Wkshp on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. ACL, Hong Kong, China, pp. 107–112, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*. ACL, Minneapolis, Minnesota, pp. 4171–4186, 2019.

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., and Da San Martino, G. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proc. of the 18th Int'l. Wkshp on Semantic Evaluation (SemEval-2024)*. ACL, Mexico City, Mexico, pp. 2009–2026, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv* vol. abs/2010.11929, 2020.

Ferreira, T. C. and Paraboni, I. Geração de linguagem natural. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 ed., H. M. Caseli and M. G. V. Nunes (Eds.). BPLN, Book chapter 16, 2024.

Guerini, M., Özbal, G., and Strapparava, C. Echoes of persuasion: The effect of euphony in persuasive communication. In *Proc of the 2015 HLT-NAACL Conf.* ACL, Denver, Colorado, pp. 1483–1493, 2015.

Halversen, A. and Weeks, B. E. Memeing politics: Understanding political meme creators, audiences, and consequences on social media. *Social Media + Society* 9 (4): 20563051231205588, 2023.

Hu, M. and Liu, B. Mining and summarizing customer reviews. In *Proc. of the 10th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*. KDD '04. ACM, New York, NY, USA, pp. 168–177, 2004.

Li, S., Wang, Y., Yang, L., Zhang, S., and Lin, H. 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)*. ACL, Mexico City, Mexico, pp. 628–633, 2024.

Mohamad, H. A. Analysis of rhetorical appeals to logos, ethos and pathos in enl and esl research abstracts. *Malaysian Journal of Social Sciences and Humanities* 7 (3), 2022.

Navigli, R., Conia, S., and Ross, B. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality* 15 (2), 2023.

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. Learning transferable visual models from natural language supervision. In *Proc. of the 38th Int'l. Conf. on Machine Learning*. Vol. 139. PMLR, pp. 8748–8763, 2021.

Takahashi, H. 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)*. ACL, Mexico City, Mexico, pp. 370–373, 2024.

Vargas, F., Carvalho, I., Pardo, T. A. S., and Benevenuto, F. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing* 31 (2): 435–456, 2025.

Weiss, M. C. Sociedade sensoriada: a sociedade da transformação digital. *Estudos Avançados* 33 (95): 203–214, 2019.

Yu, E., Wang, J., Qiao, X., Qi, J., Li, Z., Lin, H., Zong, L., and Xu, B. 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)*. ACL, Mexico City, Mexico, pp. 642–648, 2024.