# ESTUDO PRELIMINAR DA CORRELAÇÃO ENTRE NOTÍCIAS FALSAS E DISCURSOS DE ÓDIO EM TEXTOS DA LÍNGUA PORTUGUESA

CIPRIANO LANGTON CIPRIANO PARAFINO[1]; GABRIEL KUSTER DE AZEVEDO[2]; LARISSA ASTROGILDO DE FREITAS[3]; BRENDA SALENAVE SANTANA[4]

*[1]Universidade Federal de Pelotas – clcparafino@inf.ufpel.edu.br*
*[2]Universidade Federal de Pelotas – gkuster@inf.ufpel.edu.br*
*[3]Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*
*[4]Universidade Federal de Pelotas – bssalenave@inf.ufpel.edu.br*

## 1. INTRODUÇÃO

As redes sociais, que já contavam em 2024 com aproximadamente 5.04 bilhões de identidades de usuários ativos no mundo, ou seja, mais de 62% da população mundial (REPORTAL, 2024), tornaram-se o principal espaço para comunicação e circulação de informações, mas também para a disseminação de conteúdos nocivos, dentre os quais se destacam as notícias falsas, tratadas como notícias inventadas para manipular o leitor (ALLCOTT; GENTZKOW, 2017) e o discurso do ódio, definidas como sendo as linguagens discriminatórias dirigidas aos grupos identitários (RICHARDSON-SELF, 2018).

Apesar da diferença, esses fenômenos compartilham estratégias retóricas semelhantes, como exploração emocional, narrativas polarizadas e difusão massiva algorítmica (VOSOUGHI et al., 2018). Estudos mostram que campanhas de desinformação utilizam narrativas de ódios (KIM et al., 2024) e as notícias falsas favorecem a hostilidade (FAUSTINO, 2020), sugerindo que é necessária uma abordagem integrada, que possa se mostrar mais eficaz comparativamente as abordagens isoladas (LIMA, 2024). Diante desse cenário, este estudo apresenta os primeiros resultados sobre a detecção automática desses fenômenos em português, utilizando os corpora Fakebr (MONTEIRO *et al*., 2018) e HateBR (FORTUNA; NUNES, 2018). Para além das análises isoladas, realizamos experimentos de classificação cruzada, a fim de simular a sobreposição entre desinformação e hostilidade que ocorre nas redes sociais.

## 2. METODOLOGIA

O estudo foi realizado em três etapas principais: (i) identificação e caracterização dos *datasets*, (ii) aplicação das técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) e (iii) análise e interpretação dos resultados. Foram utilizados dois corpora balanceados: Fakebr, composto 7.200 notícias, sendo 3.600 falsas e 3.600 verdadeiras e HateBR, constituído por 7.000 comentários de redes sociais, 3.500 de ódio e 3.500 neutros.

Na etapa experimental, os textos foram examinados com representações vetoriais com *Bag of Words (BoW)*, *TF-IDF*, *Word2Vec (W2V)* e *BERTimbau*, juntamente com os classificadores *Naive Bayes (NB)* e *Support Vector Machine (SVM)*. Para medir o desempenho dos modelos, utilizamos as métricas do

aprendizado de máquina: acurácia (acc), que mostra a taxa do erro total, precisão (Prec), a proporção de amostras corretamente classificadas como positivas em relação ao total de amostras preditas como positivas, *recall* (Rec), mede a taxa de positivos verdadeiros, isto é, a fração de amostras positivas corretamente classificadas, e o F1-score (F1), a média harmônica de precisão e recall, especialmente útil em bases usadas neste trabalho. Além das análises isoladas, também foi realizada análise cruzada, em que o melhor modelo do Fakebr foi aplicado ao HateBR, assim como vice-versa. O objetivo foi examinar a capacidade de generalização dos classificadores e investigar a sobreposição semântica entre desinformação e hostilidade em redes sociais.

## 3. RESULTADOS E DISCUSSÃO

Os primeiros resultados foram analisados de forma isolada para ressaltar as diferenças de desempenho de cada método, dependendo da natureza de cada corpus de treinamento. A Tabela 1(a) apresenta os resultados dos modelos no Fakebr e a Tabela 1(b) apresenta os resultados obtidos no HateBR.

Tabela 1: Desempenho dos modelos nos corpora Fakebr (1a) e HateBR (1b).

**(a) Corpus Fakebr**

| Modelo | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| BoW+NB | 0.86 | 0.87 | 0.87 | 0.87 |
| BoW+SVM | 0.95 | 0.95 | 0.96 | 0.95 |
| TF-IDF+NB | 0.86 | 0.87 | 0.87 | 0.86 |
| TF-IDF+SVM | **0.96** | 0.97 | 0.97 | **0.96** |
| W2V+NB | 0.75 | 0.76 | 0.76 | 0.75 |
| W2V+SVM | 0.85 | 0.86 | 0.85 | 0.85 |
| BERTimbau+NB | 0.74 | 0.75 | 0.74 | 0.75 |
| BERTimbau+SVM | 0.84 | 0.85 | 0.85 | 0.84 |

**(b) Corpus HateBR**

| Modelo | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| BoW+NB | 0.85 | 0.86 | 0.85 | 0.85 |
| BoW+SVM | 0.84 | 0.85 | 0.85 | 0.84 |
| TF-IDF+NB | 0.85 | 0.85 | 0.85 | 0.85 |
| TF-IDF+SVM | 0.84 | 0.84 | 0.84 | 0.84 |
| W2V+NB | 0.72 | 0.73 | 0.73 | 0.72 |
| W2V+SVM | 0.79 | 0.80 | 0.80 | 0.79 |
| BERTimbau+NB | 0.81 | 0.82 | 0.82 | 0.81 |
| BERTimbau+SVM | **0.86** | 0.86 | 0.86 | **0.85** |

Para o Corpus Fakebr na Tabela 1(a), o modelo TF-IDF+SVM alcançou o F1 de 0.96, enquanto o BoW+SVM muito próximo, alcançou o F1 de 0.95. Os resultados confirmam que as notícias falsas apresentam estruturas lexicais consistentes, o que possibilita que modelos baseados em frequência identifiquem padrões discriminativos com elevada precisão. Para o Corpus HateBR na Tabela 1(b), o melhor desempenho foi alcançado pelo BERTimbau+SVM que alcançou de F1 de 0.86. Esse resultado indica que a detecção de discurso de ódio está muito mais ligada à captura semântica e contextual, uma vez que o fenômeno é muito mais variável em termos de linguística, ironia e nuance do que as notícias falsas. Os resultados isolados sinalizam contrastes relevantes: enquanto o Fakebr está na dependência de abordagens estatísticas simples, com resultados geralistas muito elevados, o HateBR precisa de representações profundas de linguagem embora com valores globais modestos.

### 3.1. Resultados da análise cruzada

Além da análise isolada, foi realizada uma análise cruzada entre os corpora, com o intuito de avaliar a variabilidade de generalização dos modelos ao serem testados fora do seu domínio de treino. Os resultados estão apresentados na Tabela 2:

Tabela 2: Resultados da avaliação cruzada entre os corpora Fakebr e HateBR.

| Modelo (treinado em) | Testado em | Rótulo atribuído | Proporção |
|---|---|---|---|
| TF-IDF+SVM (Fakebr) | HateBR | “ódio” | 42,2% |
| BERTimbau+SVM(HateBR) | Fakebr | “notícia falsa” | 99,9% |

A análise cruzada mostrou que houve baixa generalização entre domínios para os modelos. As notícias falsas foram rotuladas como ódio e vice-versa no caso de mensagens de ódio serem marcadas como notícias falsas, refletindo a sobreposição lexical encontrada entre os fenômenos e o viés no rótulo dominante do corpus de treinamento. Isso sugere que, fora do seu domínio original, os modelos se fixam em categorias únicas e não captam as nuances contextuais necessárias. Esses resultados podem ser melhor compreendidos na Tabela 3 com exemplos de casos em que uma notícia falsa foi classificada como discurso de ódio e, inversamente, uma mensagem de ódio foi interpretada como notícia falsa.

Tabela 3: Exemplos de erros de classificação observados na avaliação cruzada entre os corpora Fakebr e HateBR

| Texto original | Rótulo correto | Predição cruzada |
|---|---|---|
| “katia abreu diz vai colocar expulsao [...] ficar chorando pitangas todos cantos ” | Notícia falsa | Ódio |
| “Essa nao tem vergonha na cara!!” | Ódio | Notícia falsa |

A Tabela 3 demonstra que, na análise cruzada, os modelos frequentemente confundem notícias falsas com discurso de ódio, o que se deve à existência de pistas lexicais parecidas, ou seja, palavras ou expressões que ocorrem em ambos os fenômenos (como “mentira”, “fraude”, “vergonha” ou termos pejorativos), funcionando como evidências estatísticas que o modelo aprende a vincular a certos rótulos, e ao viés originado pelo rótulo predominante no corpus de treinamento. Desse modo, uma notícia falsa foi considerada como ódio por incluir termos pejorativos, ao passo que uma mensagem ofensiva foi classificada como notícia falsa. Esses exemplos demonstram que a sobreposição lexical entre os fenômenos desafia os modelos a identificar com precisão os contextos quando utilizados além do domínio original.

## 4. CONCLUSÕES E TRABALHOS FUTUROS

Este estudo, ainda em progresso, forneceu resultados tanto para as tarefas individuais quanto para a análise comparativa entre os corpora Fakebr e HateBR. Os resultados são encorajadores e confirmam a possibilidade de uma estratégia

unificada para identificar notícias falsas e discursos de ódio, considerando a coocorrência dos fenômenos, utilizando seus padrões lexicais comuns e abordando-os como questões interdependentes. Os experimentos mostraram indícios de sobreposição entre padrões linguísticos dos dois fenômenos, ao passo que evidenciaram limitações na capacidade de generalização dos modelos em diferentes domínios. A principal contribuição deste estudo é a análise dos resultados obtidos e a compreensão de como esses padrões se tornaram possíveis, oferecendo interpretações que podem orientar pesquisas futuras. Seria interessante que estudos futuros investigassem esses fenômenos de forma interligada.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

ALLCOTT, H.; GENTZKOW, M. **Social media and fake news in the 2016 election. Journal of economic perspectives** 31.2, vol. 31, n. 2, p. pp. 211–236, 2017.

FAUSTINO, A. **Fake news: a liberdade de expressão nas redes sociais na sociedade da informação**: Lura Editorial, 2020.

FORTUNA, P.; NUNES, S. **A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR)** 51.4, vol. 51, n. 4, p. pp. 1–30, 2018.

KIM, M.; ELMAS, T.; MENCZER, F. **Toxic Synergy Between Hate Speech and Fake News Exposure. Workshop Proceedings of the 18th Intl. AAAI Conf. on Web and Social Media (ICWSM CySoc: 5th International Workshop on Cyber Social Threats)**, vol., n., p., 2024.

LIMA, F. R. **DISCURSO DE ÓDIO E FAKE NEWS NAS REDES SOCIAIS: SUTURAS E SILENCIAMENTOS ÀS CAMPANHAS DE VACINAÇÃO CONTRA A COVID-19**. **Revista Docência e Cibercultura** 8.2, vol. 8, n. 2, p. pp. 01–20, 2024.

MONTEIRO, R. A.; SANTOS, R. L.; PARDO, T. A.; DE ALMEIDA, T. A.; RUIZ, E. E.; VALE, O. A. **Contributions to the study of fake news in portuguese: New corpus and automatic detection results. Em: International Conference on Computational Processing of the Portuguese Language**, 2018. Pp. 324–334.

REPORTAL, D. (2024). **Digital 2024: Indonesia. Data Reportal**.

RICHARDSON-SELF, L. **Woman-Hating: On Misogyny, Sexism, and Hate Speech. Hypatia** 33.2, vol. 33, n. 2, p. pp. 256–272, 2018.

VOSOUGHI, S.; ROY, D.; ARAL, S. **The spread of true and false news online. science** 359.6380, vol. 359, n. 6380, p. pp. 1146–1151, 2018.