Discursos de Ódio e Notícias Falsas: Estudo preliminar da coocorrência entre esses fenômenos em textos em português

Cipriano Parafino [1], João Tomaszewski [1], Larissa A. de Freitas [1], Brenda S. Santana [1]

[1]Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas
Rua Gomes Carneiro 1 – Centro – CEP 96010610 – Pelotas, RS – Brasil

{clcparafino,jptomaszewski,larissa,bssalenave}@inf.ufpel.edu.br

Abstract. *Social media has evolved significantly, becoming fertile ground for the spread of fake news and hate speech. These phenomena are dynamic in nature and have a complex relationship, in which one can feed the other, hindering the circulation of reliable information. This paper presents a preliminary study that aims to analyze the co-occurrence of these two phenomena in Portuguese texts, using Natural Language Processing and Machine Learning techniques applied to specific corpora.*
Keywords: *Fake news; Hate speech; Social networks; Machine Learning; Natural Language Processing.*

Resumo. *As redes sociais evoluíram de forma acentuada, tornando-se assim um terreno fértil para a propagação de notícias falsas e discursos de ódio. Estes fenômenos são dinâmicos por natureza e têm uma relação complexa, na qual um pode alimentar o outro, dificultando a circulação de informações confiáveis. Este trabalho apresenta um estudo preliminar que tem como objetivo analisar a coocorrência entre esses dois fenômenos em textos em português, por meio de técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina aplicadas a corpora específicos.*
Palavras chave: *Notícias falsas; Discursos de ódio; Redes sociais; Aprendizado de Máquina; Processamento de Linguagem Natural.*

1. Introdução

O crescimento exponencial das redes sociais criaram ambientes que passaram a ser os espaços centrais da comunicação contemporânea. Contudo, esses ambientes também se configuram como terreno fértil para a propagação de notícias falsas e discursos de ódio, ao mesmo tempo estes fenômenos estão inter-relacionados, especialmente em momentos de crise e desconfiança política. Embora sejam fundamentais como canais de comunicação, as redes sociais amplificam a polarização social e são aliadas da disseminação de conteúdos discriminatórios [Vosoughi et al. 2018].

Neste contexto, este trabalho apresenta resultados preliminares para a identificação de notícias falsas e discursos de ódio em corpora diferentes de português brasileiro, bem como o próximo passo na análise de sua coocorrência. Neste estudo, analisamos separadamente os corpora de notícias falsas e de discurso de ódio, utilizando diferentes representações textuais e diferentes modelos de classificação. Este passo é crucial para investigar posteriormente de que maneira esses fenômenos se relacionam em análises conjuntas.

2. Fundamentação teórica

As chamadas notícias falsas são informações enganosas criadas de forma intencional com o propósito de manipular o leitor, segundo definição clássica de [Allcott and Gentzkow 2017]. Já os discursos de ódio se refere a utilização de linguagem hostil, insultuosa ou discriminatória contra qualquer pessoa ou grupos com base em características identitárias [Richardson-Self 2018].

A relação de notícias falsas e discursos de ódio, tem ganhado destaque, pois ambos fenômenos compartilham dinâmicas semelhantes de disseminação. Segundo [Faustino 2020] as notícias falsas produzem um ambiente adequado para estimular discursos hostis, enquanto estes reforçam narrativas tendenciosas. Estudos conduzidos por [Kim et al. 2024] mostraram que campanhas coordenadas de desinformação empregam narrativas de ódio para manipular a opinião pública, por sua vez [Lima 2024] indicou que a fim de romper esse ciclo exige estratégias integradas de detecção, em vez de ações isoladas. Com base nesses estudos, estruturamos nossa proposta metodológica em três etapas complementares, descritas a seguir.

3. Proposta

A pesquisa proposta foi organizada em três etapas principais: (i) seleção e caracterização dos *datasets*; (ii) pré-processamento dos textos e aplicação das técnicas de Processamento de Linguagem Natural e Aprendizado de Máquinas; e (iii) avaliação dos modelos e análise dos resultados obtidos.

3.1. Seleção e caracterização dos *datasets*

Foram utilizados dois conjuntos de dados: ***Corpus FakeBr*** [Monteiro et al. 2018], com 7200 notícias sendo 3600 verdadeiras e 3600 falsas, e ***Corpus HateBr***, [Fortuna and Nunes 2018], com 7000 instâncias composto por 3500 de ódio e 3500 neutras. A escolha por esses corpora foi por conta de sua representatividade e do balanceamento entre classes.

3.2. Procedimentos metodológicos

O processo metodológico consistiu na análise dos modelos clássicos e neurais que foram utilizados separadamente para as tarefas de detecção de notícias falsas (*Corpus FakeBr*) e discurso de ódio (*Corpus HateBr*). Para representar os textos, utilizamos quatro abordagens diferentes: *Bag of Words (BoW)*, *TF-IDF*, *Word2Vec* (W2V) e *BERTimbau*. Os classificadores que utilizamos foram o *Naive Bayes (NB)* e o *Support Vector Machine (SVM)*.Optamos por esses modelos porque são bastante populares na classificação de textos e se complementam de forma equilibrada, combinando abordagens probabilísticas com a ideia de margem máxima. Assim, conseguimos comparar desde representações baseadas em frequência até *embeddings* contextuais, analisando os ganhos de desempenho em diferentes níveis de complexidade. Embora os autores de referência do Corpus FakeBr já tenham testado esse conjunto com SVM, este trabalho vai além, explorando diferentes representações textuais e utilizando também o classificador *NB*, permitindo observar as diferenças entre as abordagens clássicas e as neurais.

As métricas utilizadas para avaliar os modelos incluíram acurácia, precisão, *recall* e *F1-score*. Além de analisar cada corpus individualmente, também realizamos uma

análise cruzada, onde o modelo que teve o melhor desempenho no FakeBr foi testado no HateBr e vice-versa. O objetivo dessa abordagem foi identificar indícios de coocorrência linguística e explorar a possibilidade de generalização dos classificadores em relação a diferentes domínios textuais.

4. Resultados preliminares

Os primeiros experimentos demonstraram resultados expressivos para a detecção de ambos os fenômenos. A Tabela 1(a) apresenta o desempenho dos modelos testados no Corpus *Fakebr*, enquanto a Tabela 1(b) mostra os resultados para o Corpus *HateBR*.

Tabela 1. Desempenho dos modelos nos corpora *Fakebr* e *HateBR*.

(a) Corpus Fakebr

Modelo	Acc	Prec	Rec	F1
BoW + NB	0.86	0.87	0.87	0.87
BoW + SVM	0.95	0.96	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**

No Corpus *Fakebr*, os modelos clássicos se destacaram, com o *TF-IDF+SVM* alcançando 96% em *F1*, e o *BoW + SVM* logo atrás alcançou 95% em *F1*. Eles superaram o *BoW+NB* que alcançou 87% em *F1* e o *TF-IDF+NB*, que ficou com 86% em *F1*. No caso do *Corpus HateBR*, o modelo *BERTimbau+SVM* alcançou 85% em *F1*, seguido pelo *TF-IDF+NB* e *BoW + NB* ambos também alcançaram 85% *F1*. Já o *TF-IDF+SVM* e o *BoW+SVM* alcançaram 84% em *F1*.

4.1. Resultados da análise cruzada

Foi feita uma avaliação cruzada entre os corpora, onde frequentemente as notícias falsas eram classificadas como discurso de ódio e as mensagens de ódio como notícias falsas. Esses resultados indicam sobreposição lexical entre os fenômenos e problemas de generalização dos modelos. A Tabela 2 resume os resultados e suas implicações.

Tabela 2. Percentagem de classificação cruzada entre os Corpus *FakeBr* e *HateBr*.

Modelo (Treinado em)	Testado em	Classificação
TF-IDF+SVM (*FakeBr*)	HateBr	42,2%
BERTimbau+SVM (*HateBr*)	FakeBr	99,9%

Como demonstrado na Tabela 2, 42,2% das notícias falsas do *Corpus FakeBr* foram classificados como discurso de ódio, ao passo em que 99,9% dos discursos de ódio presentes no *Corpus HateBr* foram classificados como noticias falsas. Esses dados mostram uma sobreposição lexical entre os dois fenômenos, mas também apresentam restrições de generalização dos modelos, na sua classificação cruzada.

5. Discussão dos resultados

Os resultados preliminares evidenciam diferenciais relevantes entre os corpora. Para o *FakeBr*, a alta performance do modelo foi o *TF-IDF+SVM* com 96% em *F1*, sugere que as notícias falsas tendem a seguir padrões linguísticos um tanto homogêneos, suscetíveis de serem capturados por representações vetoriais de frequência. Para o *HateBr*, o melhor desempenho foi obtido pelo *BERTimbau+SVM* com 86% em *F1*, confirmando que o discurso de ódio supõe a captura do contexto semântico e de relações sutis entre palavras, característica das representações neurais. Na análise cruzada, observou-se que 42,2% das notícias falsas do *FakeBr* foram rotuladas como ódio no modelo treinado no *HateBr*, ao passo que 99,9% dos discursos de ódio do *HateBr* foram rotulados como notícias falsas no modelo treinado no *FakeBr*. Esses números ilustram que, de forma geral, apesar de seus diferentes sentidos, esses dois fenômenos possuem similaridades linguísticas que provocam erros nos classificadores.

6. Conclusão

Este estudo apresentou os resultados iniciais tanto nas tarefas isoladas quanto na análise cruzada dos corpora FakeBr e HateBr. Isso estabelece uma base para a proposta de uma estratégia unificada para identificar notícias falsas e discursos de ódio. Os experimentos indicaram uma possível sobreposição entre esses dois fenômenos, mas também revelaram as limitações dos modelos ao tentar generalizar para diferentes domínios. Desse modo, este estudo constitui um ponto de partida para a criação de metodologias de detecção integradas a serem investigadas nas fases seguintes.

Referências

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

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

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Kim, M., Elmas, T., and Menczer, F. (2024). 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)*.

Lima, F. R. (2024). 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):01–20.

Monteiro, R. A., Santos, R. L., Pardo, T. A., De Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

Richardson-Self, L. (2018). Woman-hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2):256–272.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151.