



Ciclo de
Encontros:
4 x PLN

Detecção automática de notícias falsas

Renato Moraes Silva

Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística
Computacional (NILC), ICMC, USP

Roney Lira de Sales e Santos

Instituto de Ciência, Tecnologia e Inovação
(ICTI), UFBA

06/02/2025

Sumário

- **Introdução**
- **Corpora**
- **Métodos de detecção automática**
- **Impactos para a sociedade**
- **Considerações finais**

Introdução

O Ecossistema da Desinformação

Notícias falsas (fake news)

Informações falsas que imitam conteúdos de agências de notícias.

Trote/Embuste (*hoax*)

Boato ou farsa sensacionalista que tem o objetivo de enganar, assustar e viralizar.

Conteúdo humorístico

Ironia, sarcasmo e sátira que podem ser mal interpretados como fatos.

Revisões falsas

Avaliações manipuladas de produtos ou serviços.

Clickbait

Conteúdo sensacionalista que visa atrair cliques.

O Ecossistema da Desinformação

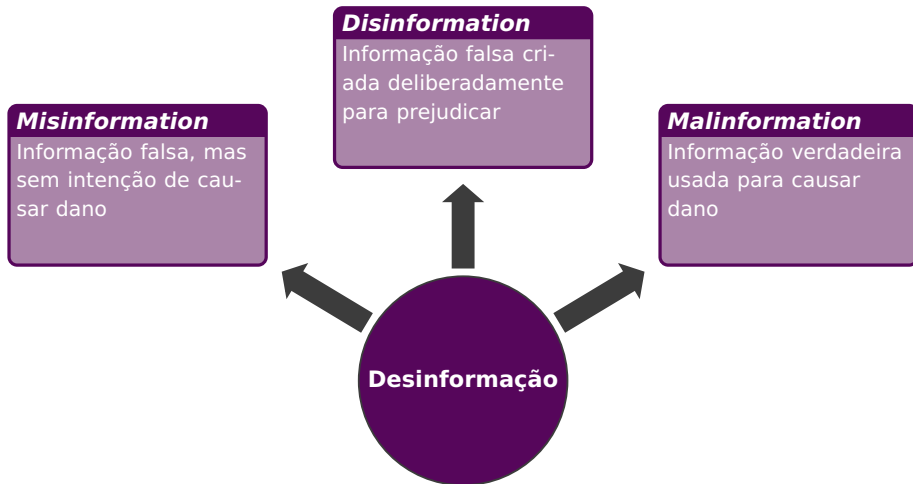
- ❑ **Notícias falsas:** problema antigo, mas a tecnologia ampliou seu alcance e impacto.
- ❑ **Exemplos de impacto:** eleições, pandemias e crises de saúde pública.
- ❑ **Consequências:** danos à saúde, prejuízos econômicos e até mesmo mortes.



The cow-pock - or - the wonderful effects of the new inoculation (Gillray, James, 1802)

- ❑ Charge usada pela campanha antivacina contra a Varíola.

Classificação da Desinformação

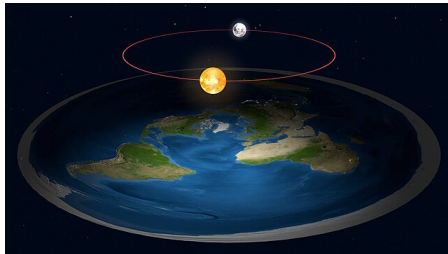


Misinformation

- ❑ “Leite com Manga faz mal.” ❌



- ❑ “A Terra é plana.” ❌



- ❑ Teorias e observações desde o século V a.C.
 - » Reforçada por Galileu Galilei (1564-1642), Isaac Newton (1643-1727) e outros

Disinformation

❑ “Leite com Manga faz mal.” ❌

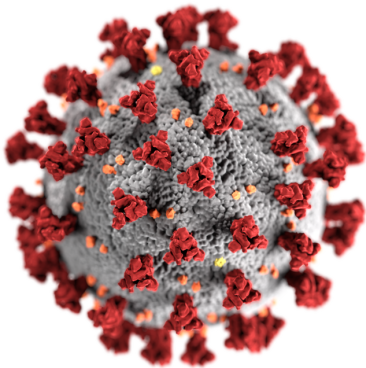


❑ “Diabetes é causada por vermes” ❌

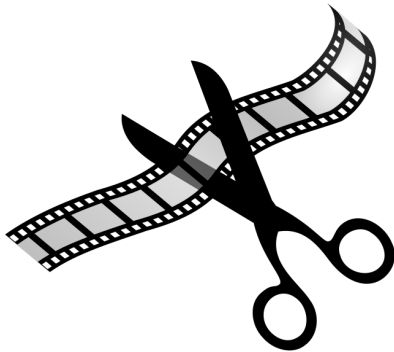


Malinformation

- ❑ Extrapolar tratamentos médicos específicos para contextos gerais.



- ❑ Informações seletivas.



Deepfakes

❑ Redes Generativas Adversárias (GANs – *Generative Adversarial Networks*)



Fonte: Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410)

❑ Imagens, áudios, vídeos falsos

g1

SOROCABA E JUNDIAÍ



Aluno de escola particular denunciado por criar nudes de adolescentes com IA presta depoimento na delegacia

Caso é investigado pela Delegacia da Infância e da Juventude de Sorocaba (SP) como ato infracional, análogo 'a utilizar-se de criança ou adolescente em cena de sexo explícito'.

Por Maira Fernandes, Moniele Nogueira, TV TEM
23/08/2024 12h16 - Atualizado há 5 meses

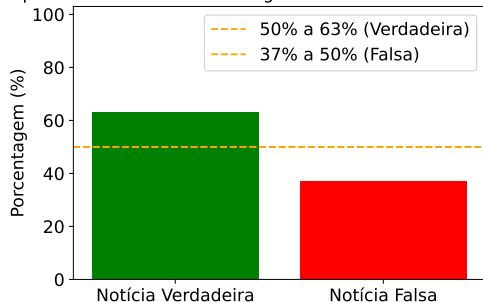


Aluno de escola particular que criou falsos nudes de adolescentes com IA presta depoimento



Desafios na Detecção de Notícias Falsas: A Dificuldade Humana

Capacidade humana em distinguir notícias falsas de verdadeiras



- ❑ Falta de conhecimento prévio sobre o assunto;
- ❑ Viés de confirmação;
 - » Reforçar crenças existentes.
- ❑ Confiança em fontes próximas, sem verificação da informação.
- ❑ Dificuldade em seguir dicas para identificar notícias falsas.

Diretrizes do Facebook para identificação de notícias falsas

- ❑ Seja cético com as **manchetes**.
- ❑ Olhe atentamente para a **URL**.
- ❑ Investigue a **fonte**.
- ❑ Fique atento a **formatações incomuns**.
- ❑ Considere as **fotos**.
- ❑ Confira as **datas**.
- ❑ Verifique as **evidências**.
- ❑ Busque **outras reportagens**.
- ❑ A história é uma **farsa** ou uma **brincadeira**?
- ❑ Algumas histórias são **intencionalmente falsas**.

Combate à Desinformação: Iniciativas

❑ Checagem Manual de Fatos



Lupa

❑ Redes Sociais: implementam medidas para sinalizar e remover conteúdo falso. Será?



Combate à Desinformação: Iniciativas

» **Redes Sociais Descentralizadas:** problema ou solução?



□ Pesquisa em PLN:

- » Métodos computacionais podem ser mais objetivos do que os humanos
 - Humanos são propensos a erros e vieses;
 - As máquinas não são?

Corpora

Corpora

Corpus	Qtd.	Categorias	Período	Tipo
Fake.br [Monteiro et al. 2018]	7.200	falsa, verdadeira	2016 a 2018	Notícias
FACTCK.BR [Moreno e Bressan 2019]	1.309	falsa, verdadeira, meia-ver-	2016 a 2019	Notícias
Bracis2019FakeNews – Twitter [Faustini e Ferreira Covões 2019]	8.981	falsa, verdadeira	—	X (Twitter)
Bracis2019FakeNews – WhatsApp [Faustini e Ferreira Covões 2019]	177	falsa, verdadeira	—	WhatsApp
FakeTweet.Br [Cordeiro e Pinheiro 2019]	299	falsa, verdadeira	—	X (Twitter)

Corpora

Corpus	Qtd.	Categorias	Período	Tipo
FakeWhastApp.BR [Cabral et al. 2021]	5.284	falsa, verdadeira	—	WhatsApp
FakeRecogna [Garcia et al. 2022]	11.902	falsa, verdadeira	2018 a 2021	notícias
FakePedia [Charles et al. 2022]	12.398	falsa, verdadeira	2013 a 2021	notícias
FakeTrueBR [Chavarro et al. 2023]	3.582	falsa, verdadeira	—	notícias
Fake tweets Covid-19 [Geurgas e Tessler 2024]	14.000	falsa, verdadeira	—	X (Twitter)
FakeRecogna 2.0 [Garcia et al. 2024]	52.800	falsa, verdadeira	2002 a 2023	notícias

Notícias alinhadas

Notícia falsa

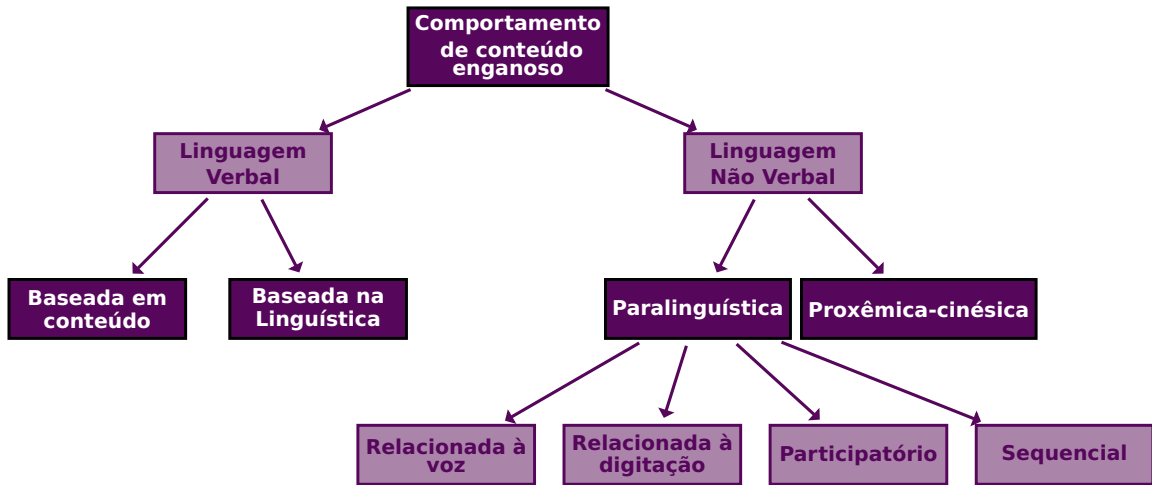
Polos magnéticos da Terra podem se inverter e causar colapso mundial: A Terra ficaria inabitável. Ao menos 100 mil pessoas morreriam por ano pela alta nos níveis de radiação espacial. Se o campo magnético continuar a diminuir e os polos magnéticos se inverterem, a Terra pode acabar como Marte – um local seco, árido e incapaz de preservar a vida.

Notícia verdadeira

Inversão dos polos magnéticos da Terra pode ocorrer mais rápido do que o previsto. Segundo afirmações, essas ocorrências são, a princípio, indistinguíveis das verdadeiras mudanças nos polos. Apesar dessas reversões não representarem qualquer ameaça à humanidade, os especialistas alertam que poderão gerar falhas nos satélites que orbitam a Terra.

Métodos de detecção automática

Abordagens de PLN para Detecção de Notícias Falsas



Aprendizado de máquina com atributos linguísticos

Corpus Fake.Br (adaptado de [Silva et al. 2020]).

Atributos	True news	Fake news
Número médio de tokens	1,268.5	216.1
Tamanho médio das palavras (em caracteres)	4.8	4.8
Razão <i>type</i> -token	0.47	0.68
Número médio de sentenças	54.8	12.7
Tamanho médio das sentenças (em palavras)	21.1	15.3
Número médio de verbos	13.4	14.3
Número médio de substantivos	24.6	24.5
Número médio de adjetivos	4.4	4.1
Número médio de advérbios	4.0	3.7
Número médio de pronomes	5.2	5.0
Número médio de stopwords	32.8	31.0
Proporção de textos com erros ortográficos	3.0	36.0
Pausalidade (pontuação / número de sentenças)	3.04	2.46
Emotividade ([adj. + adv.] / [subs. + verb.])	0.21	0.20
Incerteza (verbos modais e voz passiva)	2.11	2.39
Não imediatismo (pronomes de 1ª e 2ª pessoa)	0.235	0.249

Aprendizado de máquina com atributos linguísticos

□ Técnicas de representação [Silva et al. 2020]:

» *Bag-of-words*

- » Frequência dos termos

- » Binário

- » *Term Frequency-Inverse Document Frequency* (TF-IDF)

» Word2Vec

» FastText

» Combinação com atributos linguísticos (*ensemble* e *stacking*)

Aprendizado de máquina com atributos linguísticos

□ Métodos [Silva et al. 2020]:

- » Regressão logística (LR)
- » Máquinas de vetores de suporte (SVM)
- » Árvores de decisão (DT)
- » Floresta aleatória (RF)
- » Agregação bootstrap (bagging)
- » Boosting adaptativo (AdaBoost)

Aprendizado de máquina com atributos linguísticos

- ❑ Modelos de linguagem [Garcia et al. 2024]:
 - » LSTM
 - » GRU
 - » CNN
 - » BERT: BERTimbau e PTT-5 (*Portuguese, Tagalog, Turkish, Tamil e Telugu*)
- ❑ Outros modelos
 - » DistilBERT

Aprendizado de máquina com atributos linguísticos

□ Grandes Modelos de Língua

- » Usar o próprio modelo de língua para a classificação [Gôlo et al. 2024]
- Não há etapa de treinamento



Aprendizado de máquina com atributos linguísticos

❑ Usar *Retrieval-Augmented Generation* (RAG)



Fonte: https://python.langchain.com/docs/use_cases/question_answering/

Abordagens de PLN para Detecção de Notícias Falsas

- ❑ **Recuperação de Informação:** busca por evidências da veracidade da notícia em fontes confiáveis.
 - » Análise de resultados de busca no Google
 - » Construção de grafos de conhecimento
- ❑ **Análise de Credibilidade:** avaliação da confiabilidade da fonte da notícia e do autor.
 - » Exemplo: Investigação da reputação do site, histórico de publicações, análise de rede de fontes.

Impactos para a sociedade

Impactos Sociais das Notícias Falsas

- ❑ **Saúde:** Disseminação de informações falsas sobre tratamentos médicos e medidas de prevenção.
 - » Exemplo: Infodemia durante a pandemia de COVID-19.
- ❑ **Economia:** Manipulação do mercado financeiro com base em notícias falsas.
- ❑ **Violência:** Incitação à violência e linchamentos virtuais.
- ❑ **Polarização Política:** Erosão da confiança nas instituições e aumento da divisão social.

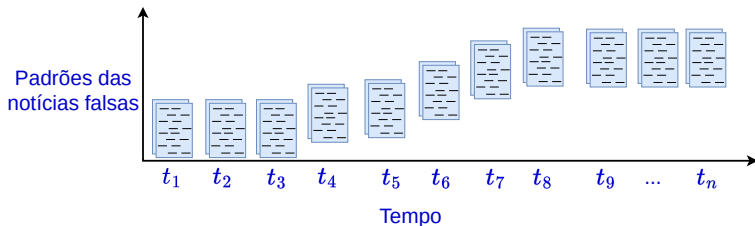


Considerações finais

Desafios e Considerações Finais

❑ Desafios:

- » **Natureza multimodal da desinformação:** Necessidade de integrar texto, imagens, vídeos e áudio.
- » **Disponibilidade de dados:** Coleta e rotulagem de dados de qualidade para treinar sistemas de detecção.
- » **Aspectos temporais:** Modelagem da dinâmica das notícias e atualização dos sistemas em relação a novos eventos.



Desafios e Considerações Finais

❑ Desafios:

- » **Fenômenos complexos:** meias verdades, pós-verdades e outros tipos de desinformação.
- » **Impacto dos erros:**
 - O que é pior: falso positivo ou falso negativo?

		Classe predita	
		Notícia Falsa	Notícia Verdadeira
Classe real	Notícia Falsa	Verdadeiro Positivo	Falso Negativo
	Notícia Verdadeira	Falso Positivo	Verdadeiro Negativo



Desafios e Considerações Finais

❑ Pilares do Combate à Desinformação:

- » **Tecnologia:** desenvolvimento de ferramentas de detecção e plataformas de checagem de fatos.
- » **Educação:** promover a alfabetização digital e o pensamento crítico.
- » **Legislação:** criar leis que responsabilizem os disseminadores de notícias falsas, respeitando a liberdade de expressão.



Referências



Cabral, L., Monteiro, J. M., da Silva, F., Wellington, J., Mattos, C. L. C., e Mourão, P. J. C. (2021).

FakeWhastApp.BR: NLP and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages.

[In Proceedings of the 23rd International Conference on Enterprise Information Systems \(ICEIS 2021\) - Volume 1, pages 63–74.](#)



Charles, A. C., Ruback, L., e Oliveira, J. (2022).

Fakepedia corpus: A flexible fake news corpus in portuguese.

[In Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings, page 37–45, Berlin, Heidelberg. Springer-Verlag.](#)

Referências



Chavarro, J., Carvalho, J., Portela, T., e Silva, J. (2023).

FakeTrueBR: Um corpus brasileiro de notícias falsas.

In [Anais da XVIII Escola Regional de Banco de Dados](#), pages 108–117, Porto Alegre, RS, Brasil. SBC.



Cordeiro, P. R. e Pinheiro, V. (2019).

Um corpus de noticias falsas do twitter e verificacao automática de rumores em lingua portuguesa.

In [Proceedings of the Symposium in Information and Human Language Technology](#), pages 219–228.

Referências



Faustini, P. e Ferreira Covões, T. (2019).

Fake news detection using one-class classification.

In Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19), pages 592–597, Salvador, BA, Brazil. IEEE.



Garcia, G. L., Afonso, L. C., e Papa, J. P. (2022).

Fakerecogna: a new brazilian corpus for fake news detection.

In International Conference on Computational Processing of the Portuguese Language, pages 57–67. Springer.

Referências



Garcia, G. L., Paiola, P. H., Jodas, D. S., Sugi, L. A., e Papa, J. P. (2024).

Text summarization and temporal learning models applied to Portuguese fake news detection in a novel Brazilian corpus dataset.

In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., e Amaro, R., editors, Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 86–96, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.



Geurgas, R. e Tessler, L. R. (2024).

Automatic detection of fake tweets about the covid-19 vaccine in portuguese.

Social Network Analysis and Mining, 14(1):55.

Referências



Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano-Neto, V. V., de Lima, E. A., e Marcacini, R. M. (2024).

On the use of large language models to detect brazilian politics fake news.

In Proceedings of the 21st National Meeting on Artificial and Computational Intelligence (ENIAC'2024), pages 1–12, Belém, PA, Brazil. Brazilian Computer Society.



Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., e Vale, O. A. (2018).

Contributions to the study of fake news in portuguese: New corpus and automatic detection results.

In 13th International Conference on Computational Processing of the Portuguese Language (PROPOR'2018), pages 324–334, Canela, Rio Grande do Sul, Brazil. Springer International Publishing.

Referências



Moreno, J. e Bressan, G. (2019).

FACTCK.BR: a new dataset to study fake news.

In [Proceedings of the 30th Brazilian Symposium on Multimedia and the Web \(WebMedia 2024\)](#), WebMedia '19, page 525–527, New York, NY, USA. Association for Computing Machinery.



Silva, R. M., de Sales Santos, R. L., Pardo, T. A. S., e Almeida, T. A. (2020).

Towards automatically filtering fake news in portuguese.

[Expert Systems with Applications](#), 146:1–48.