Estudo e comparação de Modelos de Língua para detecção de Fake News em português

Otávio Augusto Teixeira

UNESP - Instituto de Biociências, Letras e Ciências Exatas
São José do Rio Preto

Orientador: Prof. Dr. Lucas Correia Ribas
Curso: Bacharelado em Ciência da Computação

2025

Sumário

Introdução

Contextualização: O Fenômeno das Fake News

- **Ameaça à Democracia:** Disseminação massiva de informações fabricadas que emulam notícias reais.
- **Impacto:** Polarização política, erosão da confiança pública e manipulação de processos eleitorais (ex: 2018 no Brasil).
- **Desafio:** "Enganosas por design".
  - Difícil distinção para humanos e máquinas.
  - Volume massivo impede checagem manual (Fact-checking).

Motivação e Problema

O Problema

Métodos tradicionais de detecção (baseados em contagem de palavras) muitas vezes falham em capturar nuances semânticas e o contexto profundo do texto.

A Oportunidade (LLMs)

O avanço dos Grandes Modelos de Linguagem (Transformers) permitiu capturar relações contextuais complexas.

Lacuna de Pesquisa

- A maioria dos estudos foca na língua inglesa.
- Trabalhos em português ainda dependem muito de métodos clássicos (TF-IDF, Bag-of-Words) ou modelos pré-Transformers (LSTMs).
- **Lacuna:** Falta de um estudo comparativo sistemático que avalie a eficácia de LLMs modernos (GPT, Mistral, BERTimbau) especificamente para *fake news* em português.

Objetivos

Objetivo Geral

Investigar a eficácia de LLMs modernos na detecção de fake news em português, comparando-os com técnicas tradicionais de representação vetorial.

- **Específicos:**
  - Avaliar embeddings gerados por LLMs (Open Source vs Proprietários).
  - Comparar com Baselines (TF-IDF, Word2Vec).
  - Analisar o impacto do pré-processamento (Stopwords).
  - E investigar se as LLMs demonstram desempenho superior aos modelos baselines na língua portuguesa

Fundamentação Teórica

Representação Vetorial de Texto (Embeddings)

O Conceito

Computadores não processam "palavras", apenas números. **Embeddings** mapeiam palavras para vetores em um espaço multidimensional.

- **Propriedade Chave:** Palavras com significados similares ficam geometricamente próximas.
- Permite operações matemáticas com significados (Álgebra Linear aplicada à Semântica).

Abordagens Tradicionais: TF-IDF

TF-IDF (Estatístico)

- *Term Frequency - Inverse Document Frequency.*
- **Como funciona:** Calcula um peso para cada palavra.
- **Lógica:** Se uma palavra aparece muito no documento atual, mas é rara no resto do corpus (ex: "vacina"), ela é **importante**.
- **Limitação:** Gera vetores esparsos (muitos zeros) e ignora totalmente a ordem e o contexto das palavras.

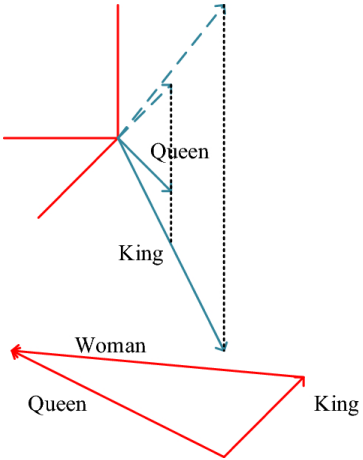
$$\text{idf}(t_i) = \log\left(\frac{N}{n_i}\right)$$

Ponderação estatística baseada em frequência.

Abordagens Tradicionais: Word2Vec

Word2Vec (Neural Raso)

- **Inovação:** Aprende vetores densos (números reais) observando a vizinhança das palavras.
- **Hipótese Distribucional:** Palavras que aparecem nos mesmos contextos tendem a ter significados parecidos.
- **Limitação:** Cria um Embedding **Estático**. A palavra "Manga"terá o mesmo vetor fixo, seja fruta ou parte da camisa.

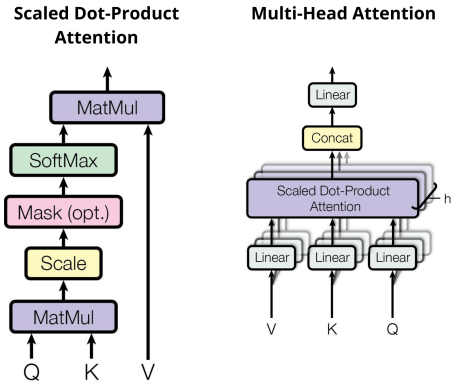


Fonte: TCC - Figura 2.4.2 (Álgebra Linear com Semântica).

A Revolução dos Transformers e o Mecanismo de Atenção

O Salto Tecnológico [Vaswani et al. 2017]:

- **Atenção (Self-Attention):** O modelo pondera a influência de *todas* as palavras da frase simultaneamente.
- **Embedding Contextual:** A representação de uma palavra muda dependendo do contexto.
- Resolve o problema da polissemia e captura dependências de longo prazo.



Fonte: TCC - Figura 2.4.4 (Mecanismo de Atenção Multi-Cabeça)

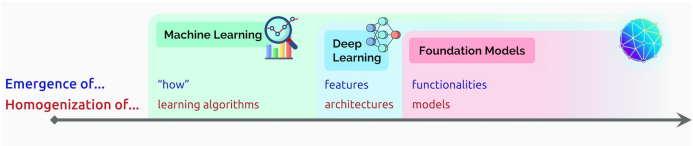
Grandes Modelos de Linguagem (LLMs)

Foundation Models

Modelos treinados em volumes massivos de dados (escala da internet) que aprendem padrões estatísticos, sintáticos e semânticos profundos.

Características Principais:

- **Emergência:** Habilidades complexas surgem implicitamente do treinamento em escala.
- **Homogeneização:** Um único modelo base (ex: GPT, BERT) serve para várias tarefas (classificação, tradução, resumo).



Emergência e Homogeneização [Bommasani et al. 2021]

Métricas de Avaliação

Para medir o desempenho, utilizamos a Matriz de Confusão (Verdadeiros/Falsos Positivos e Negativos):

1. Acurácia (Visão Geral)

$$\frac{\text{Acertos Totais}}{\text{Total de Amostras}}$$

Indica a taxa global de acerto. Como o dataset é balanceado (50/50), é uma métrica válida.

2. Precisão (Confiabilidade)

$$\frac{TP}{TP + FP}$$

Evita Falsos Positivos. De tudo que o modelo disse que era Fake, quanto era realmente?

Métricas de Avaliação

3. Revocação (Abrangência)

$$\frac{TP}{TP + FN}$$

Evita Falsos Negativos. O modelo deixou passar alguma Fake News?

4. F1-Score (Equilíbrio)

$$2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Média harmônica. É a **principal métrica** deste estudo para evitar modelos enviesados.

Trabalhos correlatos

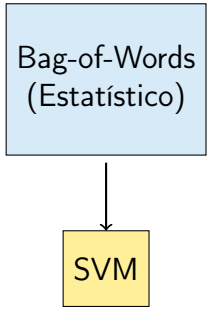
1. O Pioneirismo no Brasil: [Monteiro et al. 2018]

Contribuição: Corpus Fake.Br

- Primeiro grande corpus em PT-BR.
- **7.200 notícias** (Balanceado).
- Metodologia de pareamento (uma falsa para uma verdadeira sobre o mesmo tema).

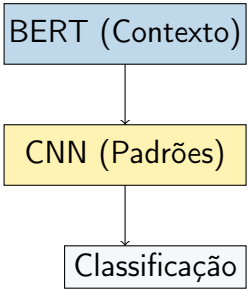
Resultados:

- Usaram SVM + Bag-of-Words.
- Surpreendente F-Measure de **89%** com métodos simples.



Representação simplificada da abordagem clássica.

2. Deep Learning Avançado: [Kaliyar, Goswami e Narang 2021]



Modelo FakeBERT

- **Híbrido:** Combina a força do BERT (texto) com Redes Convolucionais (CNN).
- Foco em inglês.
- **Estado da Arte:** 98.9% de acurácia.

Gap: Mostra o poder dos Transformers, mas não foi testado extensivamente em PT-BR.

3. O estudo em português brasileiro: [Giordani et al. 2023]

Comparação Sistemática

Este trabalho comparou diversos algoritmos clássicos no corpus Fake.Br.

Algoritmos Testados:

- SVM, Random Forest, Regressão Logística, LightGBM.
- Representations: TF-IDF vs Word2Vec.

Melhor Resultado

LightGBM + TF-IDF alcançou 96% de acurácia.

Conclusão: Métodos baseados em frequência ainda são muito fortes em português.

Metodologia

Pipeline Metodológico

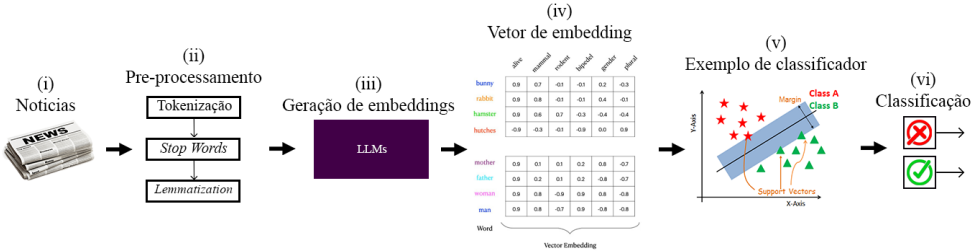


Figura: Fluxo: Coleta → Pré-processamento → Embeddings → Classificação → Avaliação.

Dataset: FakeRecogna

Características

- **11.902 notícias** coletadas entre 2019 e 2021.
- **Balanceado:** 50% verdadeiras / 50% falsas.
- Fontes: G1, UOL (Verdadeiras) e Boatos.org (Falsas).
- Não pareado (Cenário realista).

Estrutura dos Dados

Coluna	Descrição
Title	Título do artigo
Sub-title	Breve descrição/subtítulo
News	Texto completo da notícia
Category	Tópico (Saúde, Política, etc.)
Date	Data de publicação
Author	Autor da publicação
Class	**0 (Fake News) / 1 (Real News)**

Baseado na estrutura do FakeRecogna (Garcia et al., 2022).

Estratégias de Pré-processamento

Foram adotadas duas abordagens distintas para testar o impacto da limpeza no desempenho dos modelos:

1. Limpeza Extensiva (Limpeza Total)

- **O que foi feito:** Remoção de pontuação, caracteres especiais e, principalmente, remoção de **Stopwords** (artigos, preposições).
- **Justificativa:** Reduzir a dimensionalidade e ruído.
- **Foco:** Modelos tradicionais (TF-IDF, Word2Vec) que dependem da frequência de palavras-chave.

2. Limpeza Moderada (Manutenção)

- **O que foi feito:** Remoção apenas de pontuações ruidosas/links. **Mantiveram-se as Stopwords**.
- **Justificativa:** As palavras de ligação definem a estrutura sintática e o contexto da frase.
- **Foco:** Modelos baseados em Transformers (LLMs), que precisam do contexto completo.

Modelos de Representação Vetorial (1/2)

Baselines e Modelos Leves/Médios

Modelo	Tipo	Dimensão	Parâmetros	Arquitetura Base
TF-IDF	Estatístico	5.000	N/A	Baseado em Frequência
Word2Vec	Estático	300	N/A	Skip-gram
BERTimbau (base)	Contextual	768	110M	BERT Base (pt-br)
all-MiniLM-L6-v2	Contextual	384	22M	MiniLM / BERT
nomic-embed-text-v1.5	Contextual	768	137M	NomicBERT (RoPE + Flash Attention)
gte-modernbert-base	Contextual	768	149M	ModernBERT (2024)
granite-embedding-278m	Contextual	768	278M	XLM-RoBERTa (IBM Granite)

Modelos de Representação Vetorial (2/2)

Modelos de Larga Escala e Proprietários

Modelo	Tipo	Dimensão	Parâmetros	Arquitetura Base
KaLM-embedding-v2.5	Contextual	896	500M	Qwen2-0.5B (instruções)
multilingual-E5-large	Contextual	1.024	560M	XLM-RoBERTa / InfoXLM
jina-embeddings-v3	Contextual	1.024	570M	Modelo com LoRA adapters
persian-embeddings	Contextual	1.024	560M	XLM-RoBERTa bilíngue (Persa/Ing.)
SERAFIM-900M-PT	Contextual	1.536	900M	Família Albertina (pt-br)
SFR-Embedding-Mistral	Contextual	4.096	7B	Mistral-7B (MTEB SOTA)
OpenAI-3-small (API)	Contextual	1.536	N/A (Proprietário)	GPT (OpenAI)
Google embedding-001 (API)	Contextual	768	N/A (Proprietário)	Vertex AI (Google)

Classificadores Supervisionados

Para avaliar a qualidade dos embeddings gerados, foram utilizados três algoritmos de paradigmas diferentes:

SVM (Support Vector Machine)

Busca o hiperplano ótimo que separa as classes. Eficaz em espaços de alta dimensão.

Random Forest

Ensemble de árvores de decisão. Captura relações não-lineares e interações entre features.

Regressão Logística

Modelo linear probabilístico. Serve como baseline de complexidade mínima.

Validação: Divisão 80/20 estratificada e Otimização de Hiperparâmetros (Grid/Random Search).

Ferramentas e Configuração

- **Software:** Python, Scikit-learn, PyTorch, Hugging Face Transformers, RAPIDS cuML (aceleração GPU).
- **Hardware:**
  - GPU NVIDIA RTX 3060 (12GB VRAM).
  - CPU Ryzen 7 5700x.
- **Métricas:**
  - F1-Score (Principal - média harmônica).
  - Acurácia.
  - Precisão e Revocação (Sensibilidade).

Resultados

Resultados: Baselines

Embedding	Classificador	Stopwords	F1-Score
TF-IDF	SVM	Com	**0.9773**
TF-IDF	Reg. Logística	Sem	0.9706
Word2Vec	Random Forest	Com	0.8563

Tabela: Destaque para o desempenho surpreendente do TF-IDF.

- TF-IDF superou amplamente o Word2Vec.
- Baselines tradicionais ainda são muito competitivos neste corpus.

Resultados: LLMs e Transformers (1/2)

- **Dominância do SVM:** Foi o melhor classificador para 12 dos 13 modelos testados.
- **Impacto das Stopwords:** Diferente do TF-IDF, os LLMs performaram melhor **mantendo** as stopwords (preservação do contexto sintático).

Modelo	Classificador	F1-Score	Acurácia
BERTimbau	LogisticRegression	0.9672	0.9672
multilingual-E5-large	SVM	0.9693	0.9693
KaLM-embedding-v2.5	SVM	0.9622	0.9622
jina-embeddings-v3	SVM	0.9475	0.9475
granite-embedding-278m	SVM	0.9475	0.9475
nomic-embed-text-v1.5	SVM	0.9374	0.9374

Resultados: LLMs e Transformers (2/2)

Continuação dos Resultados (Modelos Modernos e APIs):

Modelo	Classificador	F1-Score	Acurácia
gte-modernbert-base	SVM	0.9185	0.9185
persian-embeddings	SVM	0.9320	0.9320
all-MiniLM-L6-v2*	SVM	0.9139	0.9139
SFR-Embedding-Mistral	**SVM**	**0.9727**	**0.9727**
SERAFIM-900M-PT	**SVM**	**0.9693**	**0.9693**
OpenAI-3-small (API)	**SVM**	**0.9698**	**0.9698**
Google-embedding-001 (API)	SVM	0.9614	0.9614

**Único modelo onde "Sem Stopwords"foi ligeiramente superior. Demais usaram "Com Stopwords".*

Resultados: Destaques

Melhores Modelos (Top 3 - F1-Score)

Os modelos que se destacaram antes da otimização de hiperparâmetros:

- **SFR-Embedding-Mistral (0.9727):** O melhor desempenho geral entre os LLMs. Baseado na arquitetura Mistral-7B.
- **OpenAI-3-small (0.9698):** Melhor modelo proprietário (API), leve e rápido.
- **SERAFIM-900M-PT (0.9693):** Modelo treinado nativamente em português (família Albertina), mostrando a força de modelos específicos para o idioma.

Otimização de Hiperparâmetros e Melhor Resultado

Modelo	Configuração	F1-Score Final	Ganho
TF-IDF	SVM Otimizado	0.9798	+0.25%
SFR-Mistral	SVM Otimizado	0.9761	+0.34%
OpenAI-3-small	**Reg. Logística Otimizada**	**0.9832**	**+1.34%**

Melhor resultado

A combinação **OpenAI + Regressão Logística** estabeleceu o melhor resultado do estudo

Análise de Erros (Qualitativa)

Falsos Positivos:

- Notícias de *fact-checking* (contêm a mentira, mas para desmenti-la).
- O modelo confunde a negação da fake news com a própria fake news.

Falsos Negativos:

- Fake news bem estruturadas (sem erros gramaticais).
- Simulação de autoridade científica.
- Mistura de fatos verdadeiros com distorções sutis.

Conclusão

Conclusão e Contribuições

- **Hipótese Confirmada:** LLMs modernos geram embeddings superiores para a detecção de fake news, atingindo 98.32% de F1-Score.
- **Benchmark:** Estabelecimento de um comparativo abrangente para o português (13+ modelos).
- **Pré-processamento:** Contrariando a intuição clássica, para LLMs, **não se deve remover stopwords**.
- **Acessibilidade:** Modelos Open Source (ex: Mistral) são altamente competitivos com modelos proprietários (OpenAI).

Limitações e Trabalhos Futuros

Limitações

- Uso de apenas um corpus (FakeRecogna).
- Custo computacional/financeiro das APIs para grandes volumes.

Trabalhos Futuros

- Aplicar *Fine-tuning* nos modelos Open Source (BERTimbau, Serafim).
- Testar generalização em outros datasets (Fake.Br).
- Investigar detecção explicável (XAI) para entender o "porquê"da classificação.

Referências I

BOMMASANI, R. et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. Disponível em: <https://arxiv.org/abs/2108.07258>.

GIORDANI, L. et al. *fakenewsbr: A Fake News Detection Platform for Brazilian Portuguese*. 2023. Disponível em: <https://arxiv.org/abs/2309.11052>.

KALIYAR, R. K.; GOSWAMI, A.; NARANG, P. Fakebert: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, v. 80, n. 8, p. 11765–11788, mar. 2021. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-020-10183-2>.

Referências II

- MONTEIRO, R. et al. Contributions to the study of fake news in portuguese: New corpus and automatic detection results: 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings. In: _____. [S.l.]: PROPOR, 2018. p. 324–334. ISBN 978-3-319-99721-6.
- VASWANI, A. et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>.

Obrigado!

Perguntas?

Otávio Augusto Teixeira

UNESP - IBILCE