Revista Eletrônica de Iniciação Científica em Computação, 202X, XX:1
ISSN: 1519-8219 • doi: 10.5753/reic.202X.XXXXXX

Artigo de Pesquisa/Research Paper

Aprendizado Supervisionado na Era da Desinformação: Estratégias para Detecção de Fake News

Supervised Learning in the Age of Disinformation: Strategies for Detecting Fake News

Davi dos Reis de Jesus [Universidade Federal de São João del-Rei | *davireisjesus@aluno.ufsj.edu.br*]
Guilherme Francis Carvalho [Universidade Federal de São João del-Re | *Guilherme2036@aluno.ufsj.edu.br*]

Ciência da Computação, Universidade Federal de São João del-Rei, Av. Leite de Castro, 847 - Fábricas, São João del Rei - MG, 36301-182, Brazil.

Resumo. Este trabalho explora a detecção de notícias falsas como uma tarefa de aprendizado supervisionado. Utilizando o conjunto de dados WELFake, que contém amostras de notícias reais e falsas, o estudo propõe uma abordagem híbrida que combina características linguísticas extraídas com o kit de ferramentas LFTK e representações vetoriais (embeddings) de modelos de linguagem. O objetivo é aprimorar o desempenho e a eficiência da classificação comparando modelos treinados com embeddings, características linguísticas e suas combinações em diferentes algoritmos, como SVM e LLMs.

Abstract. This work explores fake news detection as a supervised learning task. Using the WELFake dataset, which contains real and fake news samples, the study proposes a hybrid approach that combines linguistic features extracted with the LFTK toolkit and vector representations (embeddings) from language models. The objective is to enhance classification performance and efficiency by comparing models trained with embeddings, linguistic features, and their combination across different algorithms, such as SVM and LLMs.

Palavras-chave: Anais de evento, Modelo, SBC OpenLib, Indexação

Keywords: Fake News Detection; Supervised Learning; NLP; Linguistic Features; Embeddings; Machine Learning.

Recebido/Received: DD Month YYYY • **Aceito/Accepted:** DD Month YYYY • **Publicado/Published:** DD Month YYYY

1 Introdução

A disseminação de notícias falsas (fake news) tornou-se um problema central na era digital, com impactos políticos, econômicos e sociais. Episódios como as eleições dos EUA em 2016 e a pandemia de COVID-19 ilustram seu poder de influenciar decisões coletivas e gerar instabilidade. A velocidade e o alcance das redes sociais tornaram o trabalho manual de checagem insuficiente, evidenciando a necessidade de ferramentas automáticas e escaláveis para identificação de desinformação assim como é indentificado no artigo de Jawaher Alghamdi [2023].

Avanços em Inteligência Artificial e Processamento de Linguagem Natural popularizaram a formulação da detecção de fake news como um problema de classificação supervisionada, em que modelos aprendem a distinguir textos verdadeiros de falsos a partir de exemplos rotulados. O desempenho desses modelos depende fortemente de dois fatores: os dados de entrada e o algoritmo de aprendizado empregado.

Em paralelo, cresceu o interesse por características linguísticas explícitas extraídas por ferramentas como o Linguistic Feature Toolkit (LFTK), que oferecem métricas sobre complexidade sintática, variedade lexical e legibilidade [Lee and Lee, 2023]. Diversos estudos mostram que essas medidas capturam nuances estruturais e estilísticas relevantes, capazes de diferenciar notícias falsas de verdadeiras, complementando representações puramente semânticas [Ferreira, 2020; Silva *et al*., 2020; Garg and Suthar, 2022].

O presente trabalho propõe uma abordagem híbrida que integra as métricas extraídas pelo LFTK com embeddings gerados por modelos de linguagem), a ideia é combinar informações analíticas (sintáticas/estilísticas) e contextuais (semânticas) em uma representação unificada, aproveitando sinais complementares para formar vetores textuais mais ricos e informativos.

A hipótese central deste estudo é que a fusão entre características linguísticas explícitas e vetoriais semânticas enriquece a base de aprendizado dos modelos, ampliando sua precisão e capacidade de generalização na classificação de desinformação. Para verificar essa hipótese, foram aplicadas diferentes técnicas de aprendizado supervisionado, contemplando tanto modelos clássicos de AM, como o Support Vector Machine (SVM), quanto modelos baseados em linguagem natural de última geração, como o RoBERTa-base e o LLaMA 3.1, ambos fundamentados na arquitetura Transformer. O objetivo é gerar uma solução híbrida que combine o melhor potencial de ambas as representações, enriquecendo os dados para treinamento desses modelos e traga ganhos na eficácia para a tarefa em questão.

2 Fundamentação Teórica

2.1 Processo de detecção de Fake News

A detecção automática de notícias falsas (*fake news*) é, essencialmente, um problema de classificação binária no campo do Aprendizado de Máquina (AM). Dada a vasta quantidade de desinformação circulando digitalmente, cuja verificação manual é inviável, a criação de modelos preditivos se torna um imperativo técnico e social. O crescimento das plataformas

digitais, aliado à velocidade com que informações são compartilhadas, faz com que a desinformação se espalhe em escala e ritmo sem precedentes, exigindo soluções automatizadas, precisas e escaláveis.

O Aprendizado Supervisionado, vertente do AM adotada neste estudo, opera sob a premissa de que o modelo pode aprender a distinguir entre classes ("Fake News" e "Real") a partir de um conjunto de dados previamente rotulado. Esse processo envolve três fases fundamentais: pré-processamento dos dados, extração e representação das características relevantes do texto, e treinamento do classificador. O objetivo é treinar um modelo que não apenas memorize as características do conjunto de treinamento, mas que seja capaz de generalizar esse aprendizado para prever a classe de notícias nunca antes vistas.

Para ter sucesso no processo de classificação, o processo depende criticamente de duas etapas: a representação eficiente do texto e a escolha e otimização do algoritmo de classificação. A representação textual é responsável por transformar os dados brutos em uma forma compreensível pelos algoritmos, enquanto o modelo de classificação define como o sistema irá aprender a partir desses dados. Uma representação rica e diversificada permite que o classificador capture nuances sutis da linguagem, enquanto a escolha adequada do modelo garante a capacidade de aprendizado e generalização.

2.2 Representação do texto

O problema de classificação textual requer a transformação do texto bruto em um formato numérico processável, uma vez que os algoritmos de aprendizado não operam diretamente sobre palavras, mas sobre vetores numéricos. O conjunto de dados utilizado neste trabalho, o **WELFake**, é um dataset público composto pela fusão de quatro bases consolidadas ja existentes: os datasets do Kaggle, McIntire, Reuters e BuzzFeed, a ideia era mesclar esses diferentes conjuntos para aumentar a variedade e reduzir o viés, gerando assim um dataset mais genérico e robusto para treinar classificadores de fake news[1]. Essa base é composta por 72.134 instâncias (35,028 reais e 37,106 fake news), cada uma definida por quatro atributos principais: **ID** (identificador numérico), **Título** (título da notícia), **Texto** (conteúdo da notícia) e a **Classe** da instância (0 para *fake news* e 1 para notícia real). A presença de atributos textuais extensos, como título e corpo da notícia, permite que o estudo explore representações ricas e variadas, analisando não apenas o conteúdo, mas também o estilo de escrita.

Tradicionalmente, representações como *Bag-of-Words* (BoW) e *TF-IDF* (*Term Frequency–Inverse Document Frequency*) eram amplamente utilizadas para converter texto em vetores. No entanto, essas técnicas desconsideram a ordem e o contexto das palavras, limitando sua capacidade de capturar o significado real das frases. Com o avanço do *Processamento de Linguagem Natural* (PLN), novas abordagens surgiram, buscando representar textos de forma mais contextualizada e semântica, como as representações vetoriais densas, conhecidas como *embeddings*.

[1] https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification

2.2.1 Características linguísticas

Esta abordagem foca na engenharia de características (*feature engineering*) e na captura de características relacionadas à construção linguística dos textos. Textos de *fake news* frequentemente exibem padrões de escrita distintos das notícias reais, como maior apelo emocional [Ferreira, 2020], uso excessivo de pontuação, erros gramaticais ou sintáticos, e uma alta dose de subjetividade. Essas características são indícios importantes, pois refletem estratégias retóricas utilizadas para convencer o leitor, mesmo sem base factual.

A utilização de ferramentas como o *Linguistic Feature Toolkit (LFTK)* permite extrair características que quantificam esses atributos de forma sistemática. O LFTK é uma solução robusta e de código aberto que compila e categoriza mais de 220 características linguísticas amplamente reconhecidas pela literatura, como métricas de complexidade sintática, densidade lexical, uso de advérbios e adjetivos, e frequência de pontuação [Lee and Lee, 2023]. O benefício primário dessa representação é sua interpretabilidade: ao final do processo, é possível identificar exatamente quais traços linguísticos contribuíram para a classificação.

Além disso, a análise linguística explícita possui relevância teórica, pois aproxima o aprendizado de máquina da linguística computacional, permitindo que o comportamento do modelo seja compreendido em termos humanos. Essa interpretabilidade é fundamental em aplicações sensíveis, como a detecção de desinformação, em que a explicação do resultado é quase tão importante quanto a própria predição. No entanto, essa abordagem apresenta limitações, uma vez que as características linguísticas explícitas não capturam a semântica profunda do texto, ou seja, o significado contextual e as relações entre palavras e frases.

2.2.2 Representação vetorial

Com a evolução do PLN, as representações vetoriais densas (*embeddings*), especialmente as geradas por *Modelos de Linguagem Pré-treinados* (PLMs), tornaram-se o padrão-ouro na representação textual. Modelos baseados na arquitetura Transformer, sejam eles *Large Language Models* (LLMs) ou suas versões mais compactas, *Small Language Models* (SLMs), geram *embeddings* contextuais que capturam relações semânticas e sintáticas de forma eficaz.

A arquitetura Transformer, introduzida por Vaswani *et al.* [2017], revolucionou o Processamento de Linguagem Natural ao substituir estruturas sequenciais tradicionais pelo mecanismo de atenção automática (self-attention). Diferentemente de modelos recorrentes, que processam o texto palavra por palavra, o Transformer analisa toda a sequência simultaneamente, permitindo que cada termo seja interpretado a partir de sua relação com todos os demais elementos do enunciado. Esse mecanismo possibilita a geração de embeddings contextuais, representações vetoriais capazes de adaptar o significado de uma palavra ao contexto específico em que aparece, por exemplo, distinguindo se "banco" refere-se a uma instituição financeira ou a um assento. Entre suas principais características destacam-se: processamento paralelo, captura eficiente de dependências de longo alcance e produção de representações semânticas profundas.

A partir dessa arquitetura, surgem os Large Language Models (LLMs), modelos de grande porte que podem con-

ter centenas de milhões ou até bilhões de parâmetros. Seu tamanho expressivo permite aprender padrões linguísticos, sintáticos e semânticos complexos, além de incorporar conhecimento factual distribuído em grandes volumes de texto. Como consequência, os LLMs apresentam desempenho superior em tarefas sofisticadas, oferecendo maior capacidade de generalização e compreensão contextual. No entanto, essa alta complexidade acarreta desvantagens relevantes: elevado custo computacional, maior tempo de treinamento e inferência, além de uma redução significativa na interpretabilidade, tornando seu uso mais apropriado para aplicações específicas que exigem análise linguística profunda ou tomada de decisão baseada em contexto complexo.

Por outro lado, os Small Language Models (SLMs) constituem versões compactas baseadas na mesma arquitetura Transformer, porém com muito menos parâmetros, geralmente variando entre algumas dezenas de milhões e poucos bilhões. Essa redução torna os SLMs mais leves, rápidos e energeticamente eficientes, permitindo treinamento e execução em hardware menos robusto, muitas vezes em uma única GPU. Apesar de seu tamanho reduzido, esses modelos apresentam desempenho competitivo em diversas tarefas, especialmente quando aplicados a domínios específicos ou com dados bem estruturados. Suas limitações incluem menor capacidade de capturar nuances semânticas profundas e maior sensibilidade à falta de dados, o que pode impactar sua generalização em cenários mais complexos. Ainda assim, representam uma solução prática e eficiente quando o objetivo é equilibrar desempenho e custo computacional.

Esses vetores representam as palavras em um espaço contínuo de alta dimensionalidade, onde palavras com significados semelhantes ficam próximas entre si. Dessa forma, o modelo é capaz de compreender o contexto e a semântica do texto de uma melhor forma. Quando ajustados por meio de *fine-tuning*, esses modelos transferem o conhecimento adquirido em grandes volumes de texto para tarefas específicas, como a detecção de *fake news*.

Entretanto, essa abordagem apresenta desafios. O custo computacional e o tempo de treinamento necessário para os LLMs são elevados, exigindo recursos de hardware especializados, como GPUs de alta performance. Além disso, a interpretabilidade dos modelos baseados em *embeddings* é reduzida, pois as decisões de classificação são derivadas de representações matemáticas complexas e difíceis de rastrear. Assim, embora os LLMs ofereçam desempenho superior, sua aplicação em larga escala ainda enfrenta barreiras de acessibilidade e transparência.

2.2.3 Modelos de classificação em análise

A Tabela 1 apresenta os principais modelos de classificação utilizados nesta pesquisa, destacando suas naturezas, vantagens estratégicas e limitações operacionais. A comparação entre algoritmos de diferentes paradigmas — clássicos e baseados em redes neurais — permite compreender como a complexidade do modelo influencia o desempenho na detecção de fake news, especialmente quando variamos o tipo de representação textual.

Assim, a análise comparativa entre esses três grupos de modelos possibilita avaliar não apenas a eficácia técnica (em termos de desempenho e acurácia), mas também a eficiência prática, considerando tempo de treinamento, consumo de recursos e viabilidade de implementação. Essa abordagem híbrida e comparativa contribui para identificar o ponto de equilíbrio entre custo computacional e qualidade preditiva no contexto da detecção automática de fake news.

2.3 Abordagem híbrida

A proposta central do trabalho é investigar a abordagem híbrida, na qual as características linguísticas explícitas são combinadas com os vetores de *embeddings* (sejam de LLMs ou SLMs). O racional estratégico para essa combinação baseia-se em três pilares principais:

- **Enriquecimento da Representação:** Os *embeddings* fornecem a semântica (contexto), enquanto as características linguísticas descrevem mais a parte sintática (características com valores absolutos). A desinformação frequentemente se apoia em elementos linguísticos manipulativos e estilísticos, como ênfases emocionais, hipérboles e apelos sentimentais (Kai Shu [2017]). A combinação dessas representações enriquece a entrada do modelo, fornecendo informações complementares que podem ser ignoradas por modelos puramente semânticos.
- **Melhoria da Eficácia:** A hipótese é que um classificador treinado com representação híbrida terá mais detalhes e informações sobre a estrutura e semântica do texto, de forma conjunta, assim, aumentando suas capacidades de predição.
- **Aumento da Interpretabilidade:** Ao incorporar métricas linguísticas explícitas, o modelo torna-se mais compreensível e transparente. Essa característica é crucial em aplicações sensíveis, nas quais é necessário justificar por que determinado conteúdo foi classificado como falso.

A abordagem híbrida, portanto, busca unir o melhor dos dois paradigmas: a *interpretação linguística* das técnicas clássicas e a *profundidade semântica* dos modelos modernos de linguagem. Ao comparar o desempenho e o custo entre as abordagens puras e a híbrida em diferentes algoritmos, este estudo visa determinar a combinação ideal que maximize a performance da detecção de *fake news*.

3 Trabalhos Relacionados

A detecção automática de notícias falsas é uma área de pesquisa em rápido crescimento dentro do campo do *Processamento de Linguagem Natural* (PLN). O tema ganhou destaque especialmente após a intensificação da disseminação de desinformação em contextos eleitorais e de saúde pública, motivando a criação de modelos cada vez mais robustos para reconhecimento de padrões linguísticos e semânticos associados à falsidade textual.

Os trabalhos existentes nessa área podem ser amplamente classificados em três categorias: **(i)** abordagens baseadas em conteúdo, que analisam o texto propriamente dito; **(ii)** abordagens baseadas em metadados, que utilizam informações sobre a origem e disseminação das notícias; e **(iii)** abordagens híbridas, que integram múltiplas fontes de informação para obter maior precisão. O presente trabalho

Tabela 1. Modelos de Classificação em Análise e Suas Características Estratégicas

Algoritmo	Natureza e Vantagens	Desvantagens e Inadequações
Support Vector Machine (SVM)	Modelo clássico de AM (não neural). Excelente em encontrar o hiperplano que melhor separa as classes, sendo particularmente eficaz em espaços de alta dimensionalidade. É robusto a *overfitting*, requer pouco ajuste e é rápido na inferência.	É menos eficaz na captura de relações não lineares complexas em grandes volumes de dados. Sua performance depende fortemente da qualidade da representação de entrada e da escolha adequada do *kernel*.
SLMs (Small Language Models)	Modelos baseados em *Transformers* com quantidade significativamente menor de parâmetros quando comparados aos LLMs. Oferecem um equilíbrio entre desempenho e custo computacional, utilizando *embeddings* contextuais mais compactos e rápidos de processar. São adequados para aplicações em ambientes com recursos limitados.	Apesar da eficiência, ainda exigem mais recursos que os modelos clássicos e podem apresentar limitações em tarefas que exigem compreensão semântica profunda.
LLMs (Large Language Models)	Modelos de ponta baseados em *Transformers*. Capturam nuances linguísticas complexas, contexto de longo alcance e apresentam alta capacidade de generalização. Representam o estado da arte em tarefas de PLN.	Altíssimo custo computacional e de memória, além da menor interpretabilidade. Exigem grandes volumes de dados e hardware especializado para treinamento e inferência.

concentra-se na análise baseada em conteúdo, que busca compreender as diferenças linguísticas, discursivas e semânticas entre textos reais e falsos, explorando tanto técnicas tradicionais de *Machine Learning* quanto modelos modernos de *Deep Learning*.

3.1 Abordagens Baseadas em Características Linguísticas

As abordagens baseadas em características linguísticas constituem as raízes da detecção automática de *fake news*. Elas se fundamentam na hipótese de que a linguagem utilizada em textos falsos difere significativamente da linguagem de textos autênticos, especialmente em termos de estrutura, estilo e intenção comunicativa. Tais métodos utilizam a chamada *engenharia de características* (*feature engineering*), em que o pesquisador seleciona manualmente quais traços linguísticos devem ser considerados pelo modelo de aprendizado.

Diversos estudos corroboram essa perspectiva. Silva *et al*. [2020] e Almeida [2023], por exemplo, realizaram análises aprofundadas das características linguísticas em textos noticiosos em português, utilizando o *Fake.Br Corpus*, uma das principais bases de dados disponíveis para o idioma. Os resultados mostraram que notícias falsas tendem a apresentar maior frequência de advérbios de intensidade, uso exagerado de pontuação, construções gramaticais menos complexas e expressões de incerteza e subjetividade. Esses elementos funcionam como marcadores linguísticos que, quando quantificados, auxiliam o modelo a identificar padrões de manipulação discursiva.

Nesse contexto, algoritmos clássicos como *Support Vector Machines* (SVM), *Random Forest*, *Naïve Bayes* e *XGBoosting* demonstraram desempenho notável, especialmente quando combinados a representações baseadas em *TF-IDF* e características linguísticas explícitas. A principal vantagem dessas abordagens reside na alta interpretabilidade, permitindo compreender exatamente quais padrões de linguagem levam à classificação de um texto como falso ou verdadeiro.

Para este trabalho, três modelos clássicos serão empregados: SVM, XGBoosting e Naïve Bayes. O *Support Vector Machines* (SVM) é um classificador baseado na maximização de margens, buscando encontrar o hiperplano ótimo capaz de separar as classes no espaço de características, sendo amplamente utilizado em tarefas textuais devido à sua eficiência em dados de alta dimensionalidade. O *Naïve Bayes*, por sua vez, é um classificador probabilístico que assume independência condicional entre as características; apesar dessa simplificação, apresenta desempenho robusto em cenários com textos curtos e vetores esparsos, além de oferecer alta eficiência computacional. Já o *XGBoosting* é um método baseado em *gradient boosting* de árvores de decisão, conhecido por sua elevada capacidade preditiva, controle de *overfitting* e eficiência em termos de tempo de execução. Sua estrutura aditiva permite capturar interações mais complexas entre características linguísticas, tornando-o particularmente competitivo em tarefas de classificação textual.

No entanto, tais métodos possuem limitações claras. Embora consigam capturar elementos superficiais do texto, eles não modelam adequadamente as relações semânticas de longo alcance nem o contexto pragmático. Essa deficiência impulsionou o surgimento das abordagens basadas em *Deep Learning*, que buscam representar o significado e o contexto de forma mais precisa e autônoma.

3.2 Abordagens Baseadas em Representação Profunda (*Embeddings*)

Com o avanço das redes neurais e da arquitetura *Transformer*, o campo de detecção de *fake news* passou por uma revolução metodológica. Modelos de linguagem pré-treinados, como BERT [Devlin *et al*., 2019], RoBERTa [Liu *et al*., 2019], Dis-

tilBERT [Sanh *et al*., 2020] e GPT [Yenduri *et al*., 2023], tornaram-se o novo paradigma em tarefas de classificação textual, pois são capazes de aprender representações contextuais ricas — os chamados *embeddings*. Esses vetores densos permitem capturar nuances semânticas, relações sintáticas e dependências contextuais entre palavras que escapam à capacidade dos métodos baseados em características explícitas.

Conneau *et al*. [2020] demonstrou que modelos multilíngues, como o XLM-RoBERTa, conseguem generalizar o aprendizado de padrões linguísticos entre diferentes idiomas, ampliando a aplicabilidade em contextos multiculturais. Já Corrêa *et al*. [2024] investigou a adaptação de *embeddings* específicos para o português, utilizando modelos como BERTimbau e RoBERTa-base-portuguese, alcançando resultados superiores aos métodos tradicionais em tarefas de classificação binária de *fake news*. Esses estudos reforçam que a incorporação de *embeddings* contextuais eleva significativamente o desempenho em termos de precisão e recall, especialmente quando há grande volume de dados disponíveis para ajuste fino (*fine-tuning*).

Outros autores, como Guarise [2019], já exploravam o potencial do *Deep Learning* antes da popularização dos Transformers, utilizando redes neurais convolucionais (CNNs) e recorrentes (RNNs) para aprender representações hierárquicas do texto. Esses modelos conseguiram capturar padrões sintáticos e semânticos complexos, embora exigissem grande poder de processamento. Ainda assim, sua principal limitação residia na dificuldade de interpretar o processo de decisão — problema que persiste, em certa medida, nos LLMs modernos.

Em contrapartida, pesquisas recentes vêm discutindo as desvantagens de se utilizar exclusivamente grandes modelos de linguagem (*Large Language Models – LLMs*). Embora apresentem alta acurácia em cenários específicos, pela sua capacidade de captar a semântica dos textos coma utilização das suas *head self attention layers*, esses modelos demandam recursos computacionais significativos e, em muitos casos, não superam métodos mais simples quando o conjunto de dados é limitado ou bem estruturado [Wang *et al*., 2025]. Além disso, sua natureza de "caixa-preta"dificulta a explicação das decisões tomadas, o que é um desafio crítico em aplicações de natureza ética e social, como a detecção de desinformação.

3.3 Abordagens Híbridas

Diante das limitações das abordagens puramente linguísticas e das restrições de custo associadas aos modelos de linguagem de grande porte, surge um novo campo de estudo voltado às abordagens híbridas. Essas metodologias combinam o poder preditivo das representações profundas com a expressividade das características explícitas, buscando unir o melhor dos dois paradigmas.

Braz and Digiampietri [2024] investiga o uso de modelos híbridos em domínios cruzados, demonstrando que o enriquecimento dos modelos garante maior robustez e adaptabilidade a diferentes contextos. Além do ganho em desempenho, a combinação de representações distintas permite capturar informações complementares, ampliando a capacidade dos modelos em generalizar entre diferentes tipos de dados e tarefas. Isso é particularmente relevante na detecção de desinformação, em que a diversidade de sinais linguísticos e contextuais exige uma representação mais abrangente e integrada.

Dessa forma, o presente trabalho se insere nesse contexto emergente, propondo uma abordagem híbrida que busca combinar informações linguísticas e semânticas para formar representações mais completas e eficazes na detecção de desinformação digital.

4 Desenvolvimento

Todos os scripts e conjunto de dados utilizados neste estudo estão disponíveis publicamente no repositório: `https://github.com/GuilhermeDex/Aprendizado-de-Maquina`

4.1 Base de Dados

Para o desenvolvimento deste estudo, foi utilizada a base de dados **WELFake**[2], amplamente reconhecida na literatura por sua abrangência e diversidade temática. A base é composta por **72.134 instâncias** de notícias, distribuídas entre textos *verdadeiros* e *falsos*, provenientes da união de quatro conjuntos consolidados: *Kaggle Fake News Dataset*, *McIntire Dataset*, *Reuters*, e *BuzzFeed Political*. Essa fusão resulta em um corpus equilibrado e heterogêneo, contendo exemplos que variam em estilo, fonte, vocabulário e complexidade discursiva — aspectos fundamentais para garantir que o modelo aprenda padrões generalizáveis.

A escolha dessa base de dados se justifica pela sua diversidade linguística e semântica, uma vez que a desinformação se manifesta de formas distintas conforme o domínio temático (político, econômico, social ou científico). Assim, o uso de uma base plural permite que os modelos treinados adquiram maior robustez frente a textos de diferentes naturezas.

4.2 Classificação

O processo de classificação desenvolvido neste trabalho foi estruturado em três etapas principais e interdependentes, descritas a seguir:

1. **Representação vetorial dos dados textuais**;
2. **Classificação Base**, utilizando apenas as representações vetoriais puras;
3. **Classificação Enriquecida**, incorporando características linguísticas extraídas com o *Linguistic Feature Toolkit* (LFTK).

A tarefa de detecção de notícias falsas foi formulada como um problema de classificação supervisionada binária, na qual cada instância textual é rotulada como `fake` ou `real`. Os modelos empregados foram o **Support Vector Machine (SVM)** ([Suthaharan, 2016]), o **RoBERTa-base** ([Liu *et al*., 2019]) e o **LLaMA 3.1** ([Grattafiori *et al*., 2024]), escolhidos de forma a representar três gerações distintas de técnicas em PLN: algoritmos clássicos de aprendizado de máquina, modelos de representação contextual e grandes modelos de linguagem (*Large Language Models*).

4.2.1 *Etapa 1: Representação Vetorial dos Dados Textuais*

A etapa de representação vetorial é crucial, pois transforma os textos — originalmente cadeias de caracteres — em vetores numéricos compreensíveis pelos algoritmos. Para o modelo

[2] `https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification`

SVM, foi necessário utilizar uma arquitetura externa para geração de *embeddings*, uma vez que este não possui mecanismo interno de codificação semântica. Assim, adotou-se o modelo *sentence-transformers/all-mpnet-base-v2*[3], um *transformer* pré-treinado amplamente utilizado para geração de embeddings de sentenças. Este modelo converte cada texto em um vetor denso de 768 dimensões, capaz de capturar relações semânticas e sintáticas complexas.

Para os modelos **RoBERTa** e **LLaMA 3.1**, as representações vetoriais foram extraídas diretamente das suas camadas internas (*hidden states*), mantendo a coerência semântica entre os tokens e o contexto global da sentença. Essa escolha possibilita comparar o desempenho dos modelos sob uma base representacional homogênea, permitindo analisar o ganho obtido pela inclusão de características linguísticas explícitas.

4.2.2 Etapa 2: Classificação Base

Na Classificação Base, conforme ilustrado pela Figura 1, foram utilizadas exclusivamente as representações vetoriais geradas pelos modelos para a predição da veracidade das notícias. Essa abordagem permitiu avaliar a capacidade intrínseca dos *embeddings* em distinguir padrões semânticos associados a notícias falsas, sem qualquer interferência de variáveis externas.

Figura 1. Representação abordagem Embedding

O modelo SVM foi configurado com o kernel *RBF*, ideal para tarefas de classificação não linear, enquanto o RoBERTa e o LLaMA 3.1 foram ajustados por meio de *fine-tuning* supervisionado, utilizando o otimizador *AdamW* e a função de perda *Cross-Entropy Loss*. As métricas de avaliação consideradas foram acurácia, precisão, recall e F1-Score, calculadas com base em uma validação cruzada em 5 dobras (5-Fold Cross-Validation). Essa técnica divide o conjunto de dados em cinco subconjuntos, treinando o modelo em quatro e testando em um, de forma rotativa, o que garante maior robustez estatística e minimiza o risco de sobreajuste (*overfitting*).

4.2.3 Etapa 3: Classificação Enriquecida com LFTK

Na terceira etapa, denominada classificação enriquecida, buscou-se aprimorar o desempenho dos modelos por meio da integração entre representações vetoriais e características linguísticas extraídas com a ferramenta *Linguistic Feature Toolkit* (LFTK). Essa biblioteca permite a extração automática de mais de 200 indicadores linguísticos, abrangendo múltiplas dimensões do uso da linguagem, incluindo:

- **Métricas de complexidade sintática**, como comprimento médio das sentenças e das palavras;
- **Indicadores de coerência e coesão**, incluindo densidade lexical, diversidade de vocabulário e uso de pronomes e conectivos;
- **Índices de legibilidade e idade de aquisição de palavras**, que refletem o nível de dificuldade e acessibilidade textual.

Dentre as variáveis disponibilizadas pelo LFTK, foram selecionadas uma de cada categoria, resultando em um conjunto final composto por: *t_n_ent_money*, *a_n_ent_pw*, *simp_adj_var*, *simp_ttr*, *n_adj*, *a_adj_pw*, *fkre* e *rt_fast*. Essas variáveis foram concatenadas aos vetores de representação gerados pelos modelos de linguagem, criando **representações híbridas** que combinam informação semântica e propriedades linguísticas explícitas. O processo está ilustrado pela Figura 2.

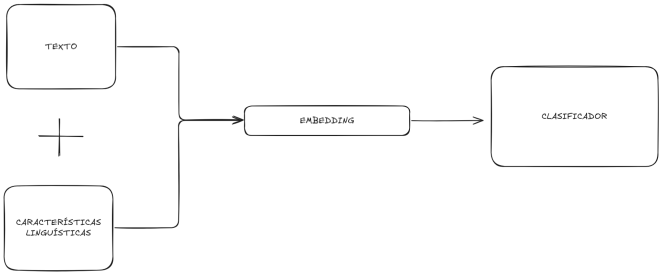

Figura 2. Representação abordagem Híbrida

A hipótese subjacente a essa etapa é que a combinação entre representações semânticas profundas e atributos linguísticos explícitos permite capturar de forma mais abrangente tanto o conteúdo quanto o estilo discursivo das notícias. Tal integração é particularmente relevante na detecção de desinformação, pois notícias falsas tendem a apresentar padrões linguísticos distintos, como maior simplicidade sintática, menor diversidade lexical ou uso excessivo de determinados conectivos e adjetivos, que nem sempre são totalmente capturados por representações baseadas apenas em contexto semântico.

Além de potencializar a eficácia preditiva dos modelos, essa abordagem contribui significativamente para a interpretabilidade e transparência dos resultados, aspectos fundamentais em aplicações sensíveis como a identificação de *fake news*. A análise das características linguísticas incorporadas permite compreender melhor quais dimensões da linguagem exercem maior influência nas decisões do modelo, favorecendo uma interpretação mais informada e confiável dos resultados obtidos. Em síntese, essa etapa busca não apenas otimizar o desempenho, mas também promover uma compreensão mais profunda sobre os mecanismos linguísticos subjacentes à propagação da desinformação.

4.3 Treinamento e Avaliação

Para a avaliação do desempenho dos modelos adotados neste trabalho, foi empregada a técnica de validação cruzada k-fold, com k=5 (5-fold cross-validation). Nessa abordagem, o conjunto de dados é dividido em cinco subconjuntos (folds) de tamanhos aproximadamente iguais. Em cada iteração, quatro folds são utilizados para o treinamento do modelo e o fold restante é reservado para teste. Esse processo é repetido cinco vezes, de modo que cada fold seja usado exatamente uma vez como conjunto de validação.

As métricas finais reportadas correspondem à média obtida ao longo das cinco iterações, proporcionando uma estimativa mais estável e confiável do desempenho do modelo.

[3] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

A escolha do valor k=5 busca equilibrar precisão na estimativa e custo computacional. Um número maior de folds tende a gerar estimativas mais precisas, porém a um custo computacional elevado, enquanto valores menores podem introduzir maior variabilidade nas métricas. Assim, o 5-fold é amplamente adotado na literatura como um compromisso adequado entre robustez e eficiência. Além disso, essa estratégia reduz a dependência de uma única partição treino/teste, utiliza de forma mais completa os dados disponíveis e fornece avaliações mais consistentes e generalizáveis para comparação entre diferentes modelos, pois permite que o modelo treine, aprenda e se teste utilizando diferentes dado e "pontos de vista"sobre eles em cada iteração.

5 Resultados

5.1 Resultados

Seguindo a metodologia e o desenvolvimento descritos na Seção 4, foram obtidos os resultados referentes à **Classificação Base**, apresentados na Tabela 2.

Tabela 2. Resultados de desempenho dos modelos na tarefa de Classificação Base.

Modelo	Precision	Recall	F1-score
SVM	93.87	93.34	93.60
RoBERTa	51.44	100.0	67.93
LLaMA 3.1	48.27	40.55	44.08
XGBoosting	90.48	90.58	90.53
Naive Bayes	76.47	77.78	77.12

Os resultados indicam que, mesmo utilizando representações vetoriais, o modelo SVM apresentou o melhor desempenho geral, superando os modelos baseados em arquiteturas transformer, seguido pelo XGBoosting, demonstrando a capacidade de desempenho dos métodos *ensemble*. Esse comportamento sugere que, para o conjunto de dados e a representação adotados, o SVM mostrou-se mais eficiente na separação das classes, possivelmente em razão de sua capacidade de generalização em espaços de alta dimensionalidade e de sua maior estabilidade frente a conjuntos de dados de tamanho moderado. E por fim, o Naive Bayes apresenta o terceiro melhor resultado, suprando, também, os modelos de linguagem.

Por outro lado, o modelo RoBERTa, embora não tenha obtido o maior valor de F1-score, alcançou recall máximo (1.00), o que indica que foi capaz de identificar corretamente todos os exemplos positivos. No entanto, o baixo valor de *precision* evidencia uma tendência à superclassificação da classe positiva, comprometendo o equilíbrio entre precisão e sensibilidade. Já o modelo LLaMA 3.1 apresentou desempenho inferior em todas as métricas, indicando possíveis dificuldades de adaptação ao domínio e à natureza do conjunto de dados empregado, o que pode estar relacionado à falta de ajuste fino (*fine-tuning*) mais direcionado à tarefa específica.

Em seguida, foram conduzidos novos experimentos aplicando a metodologia proposta neste trabalho, que incorpora características linguísticas extraídas automaticamente por meio da ferramenta LFTK. Os resultados obtidos estão apresentados na Tabela 3.

Tabela 3. Resultados de desempenho dos modelos utilizando a metodologia proposta.

Modelo	Precision	Recall	F1-score
SVM	92.21	91.61	92.21
RoBERTa	51.40	100.0	67.9
LLaMA 3.1	61.1	14.9	24.0
XGBoosting	89.05	89.50	89.30
Naive Bayes	74.77	76.24	75.50

Observa-se que a inclusão das características linguísticas extraídas pelo LFTK não resultou em melhorias expressivas no desempenho global dos modelos e, em alguns casos, até piorou os resultados. O SVM manteve o melhor desempenho entre os três avaliados, sendo novamente seguido pelo XGBoosting e pelo Naive Bayes, demonstrando robustez frente à introdução das novas variáveis. E o LLaMA obteve melhora no recall e piora no precision. Esse comportamento sugere que as características adicionadas, embora potencialmente informativas, não foram suficientemente discriminativas para impactar de forma significativa as decisões dos classificadores. Além disso, no caso do LLaMA, talvez as informações adicionadas tenham favorecido apenas uma das classes, enviesando o aprendizado do modelo.

Uma hipótese plausível é que, dentre as 220 características extraídas, as selecionadas para compor a solução não representaram as mais relevantes para o problema em questão. Além disso, o grande volume de atributos linguísticos pode ter introduzido redundância e ruído, diluindo o ganho esperado.

Dessa forma, trabalhos futuros podem explorar técnicas adicionais de seleção e redução de dimensionalidade, como *Recursive Feature Elimination (RFE)* ou *Principal Component Analysis (PCA)*, com o intuito de identificar subconjuntos de características mais representativos. A exploração de diferentes combinações de métricas linguísticas pode contribuir para uma avaliação mais precisa da eficácia da metodologia proposta, além de possibilitar o refinamento do processo de extração de atributos e o aprimoramento do desempenho dos modelos de classificação.

6 Conclusão

O presente estudo investigou o uso de técnicas de aprendizado supervisionado para a detecção de *fake news*, propondo uma abordagem que combina representações vetoriais semânticas, derivadas de modelos de linguagem, com classificadores tradicionais. Os experimentos realizados demonstraram que, mesmo em sua configuração base, isto é, utilizando apenas *embeddings* como representação textual, o modelo Support Vector Machine (SVM) apresentou o melhor desempenho geral entre as abordagens testadas.

Esse resultado evidencia que algoritmos clássicos de classificação, quando alimentados por representações contextuais ricas, ainda são altamente competitivos em tarefas complexas de Processamento de Linguagem Natural (PLN). A capacidade do SVM de generalizar em espaços de alta dimensionalidade mostrou-se particularmente eficaz na distinção entre notícias verdadeiras e falsas, reforçando seu potencial em cenários com conjuntos de dados limitados ou desbalanceados. Além disso, nota-se também o desempenho do XGBoosting e do Naive Bayes, que reforçam a capacidade dos

modelos clássicos de superarem o desempenho dos modelos linguagem nesse contexto, conseguindo melhores resultados a um menor custo, indo de encontro à literatura para a maioria dos outros cenários, que apontam as redes neurais como o SOTA (Estado da arte) para a maioria desses problemas.

O modelo RoBERTa apresentou recall máximo, indicando sensibilidade elevada à classe positiva, porém à custa de uma redução significativa na precision, o que evidencia uma tendência à classificação excessiva de instâncias como verdadeiras. Já o modelo LLaMA 3.1 apresentou desempenho inferior nas três métricas analisadas, sugerindo a necessidade de ajustes adicionais, como fine-tuning mais direcionado ou estratégias de adaptação ao domínio.

Embora a introdução das características linguísticas extraídas pelo LFTK não tenha produzido melhorias expressivas no desempenho, os resultados indicam a viabilidade da integração entre informações linguísticas e representações semânticas. Acredita-se que aprimoramentos na etapa de seleção de atributos, bem como a exploração de diferentes subconjuntos de métricas linguísticas, possam potencializar o impacto da metodologia proposta.

Como perspectivas futuras, destaca-se a aplicação de técnicas avançadas de seleção de características, o uso de abordagens de *ensemble learning* e a ampliação do corpus de treinamento com dados mais diversificados e balanceados. Além disso, pretende-se investigar o uso de representações multimodais e a adaptação da metodologia para outros domínios textuais, visando aprimorar a robustez e a generalização dos modelos na detecção automática de desinformação.

Declarações complementares

Agradecimentos

ESTA DECLARAÇÃO É OPCIONAL. Este é um texto de agradecimentos com várias linhas. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Financiamento

ESTA DECLARAÇÃO É OPCIONAL. Esta pesquisa foi financiada por lorem ipsum dolor sit amet, consectetur adipiscing elit.

Contribuições dos autores

ESTA DECLARAÇÃO É OBRIGATÓRIA. Sugerimos que os autores descrevam sua contribuição usando a Taxonomia CRediT (https://credit.niso.org/) como neste exemplo: JV contribuiu para a concepção deste estudo. CB, RP e CM realizaram os experimentos. JV é o principal contribuidor e escritor deste manuscrito. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

ESTA DECLARAÇÃO É OBRIGATÓRIA. Se não houver conflitos de interesse, os autores devem declarar: "Os autores declaram que não têm nenhum conflito de interesses". Caso contrário, a declaração deve ser: "Os autores declaram que têm os seguintes conflito de interesses: lorem ipsum dolor sit amet, consectetur adipiscing elit."

Disponibilidade de dados e materiais

ESTA DECLARAÇÃO É OBRIGATÓRIA. Se os autores estiverem disponibilizando seus dados e/ou códigos abertamente, a declaração deve ser: "Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual estão disponíveis em …". Caso contrário, a declaração deve ser: "Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual serão feitos mediante solicitação".

Outras informações relevantes

ESTA DECLARAÇÃO É DESEJÁVEL. Informações adicionais relevantes, como, por exemplo, a aprovação em comitê de ética ou o uso de ferramentas de IA generativa no desenvolvimento do artigo. Essa declaração é opcional, se não houver nada a ser acrescentado, pode ser deixada em branco

Referências

Almeida, R. F. S. d. (2023). *Building portuguese language resources for natural language processing tasks*. PhD thesis, Universidade Federal de Minas Gerais.

Braz, R. R. and Digiampietri, L. A. (2024). Detecção de fake news em domínios cruzados: Uma revisão sistemática. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. SBC.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. ACL.

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and De Oliveira, N. (2024). Detecção de fake news em português: Análise comparativa entre métodos de representação em português, inglês e multilíngues. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Ferreira, M. A. P. (2020). Fake news: emoções como estratégia discursiva. *Cadernos de Linguística*, 1(4):01–16.

Garg, S. and Suthar, D. K. (2022). Linguistic features based framework for automatic fake news detection. *Information Processing & Management*, 59(6):103053. DOI: 10.1016/j.ipm.2022.103053.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G.,

Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models.

Guarise, L. (2019). Detecção de notícias falsas usando técnicas de deep learning. Technical report, Universidade de São Paulo (USP).

Jawaher Alghamdi, Suhuai Luo, Y. L. (2023). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*.

Kai Shu, Amy Sliva, S. W. J. T. H. L. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36. DOI: 10.1145/3137597.3137600.

Lee, B. W. and Lee, J. H.-J. (2023). Lftk: Handcrafted features in computational linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Silva, F. S., Fernandes, J. F. S., and Correia, A. M. B. (2020). Towards automatically filtering fake news in portuguese. *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 227–230.

Suthaharan, S. (2016). *Support Vector Machine*, pages 207–235. Springer US, Boston, MA. DOI: 10.1007/978-1-4899-7641-3$_9$.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6000–6010. Curran Associates, Inc.

Wang, H., Zhang, Y., and Liu, Y. (2025). Svm, bert, or llm? a comparative study on multilingual instructed deception detection. *MDPI*, 6:239. Disponível no ResearchGate.

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., and Gadekallu, T. R. (2023). Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions.