Revista Eletrônica de Iniciação Científica em Computação, 2025

Research Paper

Aprendizado Supervisionado na Era da Desinformação: Estratégias para Detecção de 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.

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.

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

Received: 17 October 2025

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.

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. Essas medidas capturam nuances e estruturas textuais que podem ser discriminativos para diferenciar notícias falsas de verdadeiras, complementando informações puramente semânticas.

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 é que a fusão dessas representações enriquece a base de aprendizado do modelo, resultando em maior precisão e capacidade de generalização na classificação de desinformação. O objetivo é gerar uma solução híbrida que enriqueça os dados para treinamento dos modelos e traga ganhos na eficácia.

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**, serve de base para esse processo. 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*.

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.

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**??**. 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

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* de menor porte. 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.

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 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* e *Naïve Bayes* 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.

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 impulsi-

onou o surgimento das abordagens baseadas 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, RoBERTa, DistilBERT e GPT, 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, 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**[1], 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.

Antes do treinamento, os dados foram submetidos a um processo de pré-processamento textual, que incluiu as etapas de:

- **Normalização textual**, com a conversão para letras minúsculas e remoção de caracteres especiais e URLs;
- **Tokenização**, realizada por meio das bibliotecas *Transformers* e *NLTK*;
- **Remoção de stopwords**, a fim de eliminar termos sem carga semântica significativa;
- **Lematização**, para reduzir as palavras à sua forma canônica;
- **Balanceamento das classes**, garantindo distribuição proporcional entre instâncias de notícias verdadeiras e falsas durante o processo de treinamento.

Essas etapas são essenciais para mitigar ruídos e redundâncias no texto, reduzindo a dimensionalidade dos dados e

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

aprimorando a qualidade das representações vetoriais. Além disso, buscou-se preservar o conteúdo semântico das sentenças, de modo que o processo de limpeza não interferisse na análise dos padrões discursivos relevantes para a classificação.

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)**, o **RoBERTa-base** e o **LLaMA 3.1**, 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 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*[2], 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, 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.

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 LFTK (Linguistic Feature Toolkit). Essa biblioteca permite a extração automática de mais de 200 indicadores linguísticos, incluindo:

- Métricas de complexidade sintática (comprimento médio das sentenças, palavras);
- Indicadores de coerência e coesão (densidade lexical, diversidade de vocabulário, uso de pronomes e conectivos);
- índices de legibilidade e idade de aprendizado de palavras.

Essas variáveis foram concatenadas aos vetores gerados pelos modelos de linguagem, criando representações híbridas mais ricas e informativas. A hipótese é que a junção de características semânticas profundas e atributos linguísticos explícitos permite capturar tanto o *conteúdo* quanto o *estilo* discursivo das notícias, o que é particularmente relevante na detecção de desinformação.

Além de potencializar a eficácia, essa integração contribui para a interpretabilidade dos resultados, pois permite compreender quais aspectos linguísticos mais influenciaram a decisão do modelo, o que contribui em termos de transparência e compreensão dos resultados.

4.3 Treinamento e Avaliação

Para avaliar o desempenho dos modelos adotados neste trabalho utilizou-se validação cruzada do tipo k-fold com k=5 (5-fold). Nessa abordagem os dados são particionados em cinco subconjuntos (folds) de tamanho aproximadamente igual; em cada iteração, quatro folds são usados para treinamento e o fold restante para teste, repetindo-se o processo até que cada fold tenha sido usado uma vez como conjunto de validação. As métricas finais reportadas correspondem à média (e ao desvio-padrão) obtidos nas cinco iterações.

A escolha da 5-fold visa balancear a necessidade de um estimador confiável do desempenho com o custo computacional. Esse procedimento reduz a variância da avaliação em relação a uma única divisão treino/teste, aproveita de forma mais eficiente toda a amostra disponível e fornece estimativas mais robustas para comparação entre modelos.

5 Resultados

Seguindo a metodologia e o desenvolvimento descritos na Seção 3, foram obtidos os resultados parciais referentes à Classificação Base, apresentados na Tabela 2.

Os resultados indicam que, mesmo utilizando representações vetoriais, o modelo SVM apresentou o melhor desempenho geral no primeiro experimento de classificação, superando

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

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	1.00	67.93
LLaMA 3.1	48.27	40.55	44.08

os modelos baseados em arquiteturas transformer. Esse resultado sugere que, para o conjunto de dados e a representação adotados, o SVM foi mais eficiente na separação das classes, possivelmente devido à sua capacidade de generalização em espaços vetoriais de alta dimensionalidade.

Em contrapartida, o modelo RoBERTa, embora não tenha alcançado o maior valor de F1-Score, apresentou recall máximo (1.00), o que indica que foi capaz de identificar corretamente todos os exemplos positivos. No entanto, o baixo valor de precision revela uma tendência à superclassificação da classe positiva, o que impactou negativamente o equilíbrio global do modelo.

6 Conclusão

O presente estudo investigou o uso do aprendizado supervisionado na detecção de fake news, propondo uma abordagem que combina representações vetoriais semânticas de modelos de linguagem com técnicas tradicionais de classificação. Os experimentos evidenciaram 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. Essa performance demonstra que classificadores clássicos, quando alimentados por representações contextuais ricas, são capazes de generalizar eficientemente em espaços de alta dimensionalidade, alcançando resultados competitivos em tarefas de compreensão textual complexas como a identificação de desinformação. O modelo RoBERTa obteve recall máximo, mas com queda de precisão, indicando tendência à classificação excessiva de instâncias como reais; já o LLaMA 3.1 apresentou desempenho inferior.

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.

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

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

Jasraj Singh, Fang Liu, H. X. B. C. N. W. Z. (2024). Lingml: Linguistic-informed machine learning for enhanced fake news detection. *arXiv preprint arXiv:2405.04165*.

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.

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.

Sonal Garg, D. K. S. (2022). Linguistic features based framework for automatic fake news detection. *Information Processing & Management*.

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.

Yanping Shen, Qingjie Liu, N. G. J. Y. Y. Y. (2023). Fake news detection on social networks: A survey. *Applied Sciences*, 13(21):11877.