Filo-Transformer: Um modelo baseado em Grafo de Alinhamento de Árvores Filogenéticas e Transformers para Identificação de Rumores e Fake News

Acauan C. Ribeiro[1,2], **Eduardo L. Feitosa**[1], **André Carvalho**[1]

[1]IComp – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brasil

[2]DCC – Universidade Federal de Roraima (UFRR)
Boa Vista, RR – Brasil

acauan.ribeiro@ufrr.br, efeitosa@icomp.ufam.edu.br

andre@icomp.ufam.edu.br

Abstract. *This paper presents Filo-Transformer, an innovative approach that combines the deep semantics of Transformer models with the evolutionary analysis of Tree Alignment Graphs (TAGs) to detect rumors and fake news. The model uses embeddings (SBERT/GPT) to represent content and extracts real phylogenetic attributes from Twitter conversation cascades, such as cascade depth, branching factor, and verified user ratio. A Feature Tokenizer Transformer (FT-Transformer) integrates this information for classification. Experiments on the PHEME dataset show that Filo-Transformer outperforms purely semantic models across all major metrics, with fusion weights learning converging to 65% phylogenetic and 35% semantic features, confirming the value of structural propagation patterns.*

Resumo. *Este artigo apresenta Filo-Transformer, uma abordagem inovadora que une a semântica profunda de modelos Transformer com a análise evolutiva de Grafos de Alinhamento de Árvores Filogenéticas (TAGs) para detectar rumores e fake news. O modelo utiliza embeddings (SBERT/GPT) para representar o conteúdo e extrai atributos filogenéticos reais das cascatas de conversação do Twitter, como profundidade de cascata, fator de ramificação e proporção de usuários verificados. Um Feature Tokenizer Transformer (FT-Transformer) integra essas informações para classificação. Experimentos no dataset PHEME mostram que o Filo-Transformer supera modelos apenas semânticos em todas as métricas principais, com pesos de fusão aprendidos convergindo para 65% características filogenéticas e 35% semânticas, confirmando o valor dos padrões estruturais de propagação.*

1. Introdução

A proliferação de *notícias falsas* (*fake news*) em plataformas online tornou-se uma ameaça global, com implicações sociais, políticas e de segurança. Redes sociais permitem que informações não verificadas se espalhem de forma viral devido à sua natureza aberta e não moderada [Zubiaga et al. 2016]. Estudos recentes revelam que *notícias falsas* propagam-se mais rápido e mais longe do que notícias verdadeiras em mídias sociais. Em particular,

[Vosoughi et al. 2018a] observaram que, no Twitter, boatos falsos difundem-se significativamente mais fundo e mais amplamente do que fatos verídicos, frequentemente por uma ordem de grandeza de diferença. Esse alcance acelerado deve-se principalmente ao compartilhamento humano e não a bots, intensificando o desafio de conter a desinformação na fonte [Vosoughi et al. 2018a].

Modelos estritamente semânticos, como Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al. 2019] e Sentence-BERT Adapted (SBERTA) [Reimers and Gurevych 2019a], são eficazes na interpretação do significado textual em um ponto específico no tempo. No entanto, esses modelos falham em capturar a dinâmica evolutiva das mensagens, negligenciando processos de reescrita, recombinação e republicação de conteúdo nas redes sociais. Essa ausência de perspectiva temporal mascara variações textuais sutis, frequentemente cruciais para a viralização de rumores, comprometendo a capacidade desses sistemas para a detecção precoce de narrativas adulteradas em circulação.

Outro problema diz respeito à propagação, momento em que uma publicação inicial desencadeia uma série de respostas, compartilhamentos e modificações, denominada neste artigo como *cascata de informação*[1]. [Peng et al. 2024, Qiu et al. 2022] ressaltam a importância de entender a evolução temporal e estrutural da informação, aspecto que a abordagem proposta sistematiza e explora computacionalmente.

Nesse sentido, e com o intuito de superar as limitações das abordagens existentes, este artigo investiga como integrar, eficazmente, a análise da evolução e mutação da informação com representações semânticas profundas para aprimorar a detecção de *rumores* e *fake news*. Destaca-se que tal integração mostra-se fundamental, uma vez que as metodologias atuais descritas na literatura frequentemente se dedicam a aspectos isolados do problema.

O objetivo geral deste trabalho é propor e avaliar um modelo híbrido inovador para detecção de *rumores* e *fake news*, denominado **Filo-Transformer**, que integra aspectos evolutivos e semânticos na análise textual. Especificamente, o trabalho propõe: (i) desenvolver métodos para a construção eficaz de *Grafos de Alinhamento de Árvores* (TAGs) [Smith et al. 2013] a partir de *cascatas informacionais* utilizando *embeddings* semânticos avançados; (ii) definir e extrair atributos filogenéticos relevantes desses grafos, como profundidade e recombinação, e integrá-los com representações semânticas através de uma arquitetura Feature Tokenizer Transformer (FT-Transformer) [Gorishniy et al. 2021] adaptada; (iii) e validar experimentalmente a proposta em *dataset* público, comparando-a a modelos tradicionais e avaliando o impacto dos componentes filogenéticos.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta uma revisão da literatura, discutindo conceitos fundamentais e trabalhos relacionados. A Seção 3 detalha o modelo Filo-Transformer proposto. A Seção 4 descreve a configuração experimental. A Seção 5 apresenta os resultados obtidos e uma análise comparativa. A Seção 6 discute trabalhos relacionados que utilizam bases de dados e métricas similares. Finalmente, a Seção 7 conclui o trabalho, sintetizando as contribuições, limitações e direções para pesquisas futuras.

[1]Ela representa o conjunto de todas as publicações inter-relacionadas que se originam de um conteúdo semente e traçam seu percurso através da rede ou plataforma.

2. Conceitos Fundamentais

Esta seção estabelece a base conceitual sobre *rumores* e *fake news*, e apresenta os principais conceitos e temas empregados neste artigo.

2.1. Rumores e Fake News

Rumores são informações não verificadas ou incertas que circulam amplamente, frequentemente em situações onde há alta ambiguidade ou relevância social, podendo ou não ser confirmados posteriormente como verdadeiros ou falsos [Vosoughi et al. 2018b]. Embora os *rumores* possam surgir espontaneamente e sem intenção clara de causar danos, sua disseminação pode levar à desinformação e causar efeitos adversos em sociedades, especialmente durante crises ou eventos emergentes [Zubiaga et al. 2016]. Em contraste, *fake news* são intencionalmente fabricadas ou manipuladas com o objetivo explícito de enganar ou influenciar opiniões públicas, comportamentos ou decisões políticas [Lazer et al. 2018, Shu et al. 2017]. A característica central das *fake news* reside na intenção maliciosa e planejada de distorcer fatos ou contextos, resultando em narrativas que frequentemente exploram emoções negativas como medo, raiva ou indignação para maximizar sua disseminação e impacto [Wardle 2017]. Compreender a distinção e intersecção entre *rumores* e *fake news* é essencial para o desenvolvimento de métodos eficazes de identificação e mitigação dessas informações no ambiente digital.

2.2. Geração de *Embeddings*

Embeddings textuais são representações vetoriais densas e contínuas que codificam significados semânticos e sintáticos das palavras, frases ou documentos em espaços vetoriais de alta dimensão [Mikolov et al. 2013]. Inicialmente, técnicas como Word2Vec e GloVe estabeleceram os fundamentos dos *embeddings* baseados em redes neurais, capturando relações lineares entre palavras por meio do contexto local [Pennington et al. 2014]. Recentemente, o avanço na arquitetura de modelos como o Sentence-BERT (SBERT) e os *embeddings* gerados por grandes modelos de linguagem (LLMs), como GPT-3.5 e GPT-4, ampliaram significativamente a capacidade de captura semântica desses *embeddings* [Reimers and Gurevych 2019b, Brown et al. 2020]. Esses modelos, treinados sobre corpos de dados grandes e diversificados, conseguem produzir *embeddings* altamente sensíveis às nuances semânticas, permitindo aplicações avançadas em detecção de *fake news* e análise de texto [Devlin et al. 2019].

2.3. Transformers

A arquitetura *Transformer* representa uma mudança de paradigma em aprendizado profundo para processamento de linguagem natural (PLN), substituindo modelos recorrentes tradicionais por mecanismos baseados exclusivamente em atenção [Vaswani et al. 2017]. Este mecanismo permite capturar dependências de longo alcance dentro do texto, reduzindo os problemas de desaparecimento e explosão de gradiente, característicos das redes neurais recorrentes (RNNs). *Transformers* utilizam camadas de *multi-head attention* que simultaneamente realizam múltiplas projeções lineares dos dados de entrada, proporcionando uma visão mais rica e diversificada das relações entre palavras ou subunidades textuais. Essa arquitetura fundamenta modelos atuais como o GPT, BERT e variantes subsequentes, os quais demonstram desempenhos excepcionais em tarefas complexas, incluindo classificação textual, análise semântica e detecção automática de *fake news* [Radford et al. 2019].

2.4. Reconstrução Filogenética Textual

Grafos de Alinhamento de Árvores (TAGs, do inglês *Tree Alignment Graphs*) são estruturas que representam e alinham múltiplas árvores filogenéticas, capturando relações evolutivas complexas, como recombinações e divergências, originalmente em contextos biológicos [Smith et al. 2013]. Recentemente, essas técnicas têm sido aplicadas com sucesso em contextos digitais para estudar como narrativas falsas se propagam e evoluem ao longo do tempo, modelando a ancestralidade e mutação de textos por meio de similaridades calculadas com *embeddings* semânticos. Tais estruturas são fundamentais para entender mecanismos subjacentes à propagação das *fake news*, fornecendo uma base analítica robusta para intervenções e estratégias de mitigação no ambiente digital [Jang et al. 2018].

3. Modelo Proposto: Filo-Transformer

O Filo-Transformer é um modelo híbrido projetado para identificar *rumores* e *fake news*, integrando representações semânticas profundas com características extraídas da análise da evolução e propagação da informação.

3.1. Arquitetura

A arquitetura geral do Filo-Transformer é ilustrada na Figura 1.

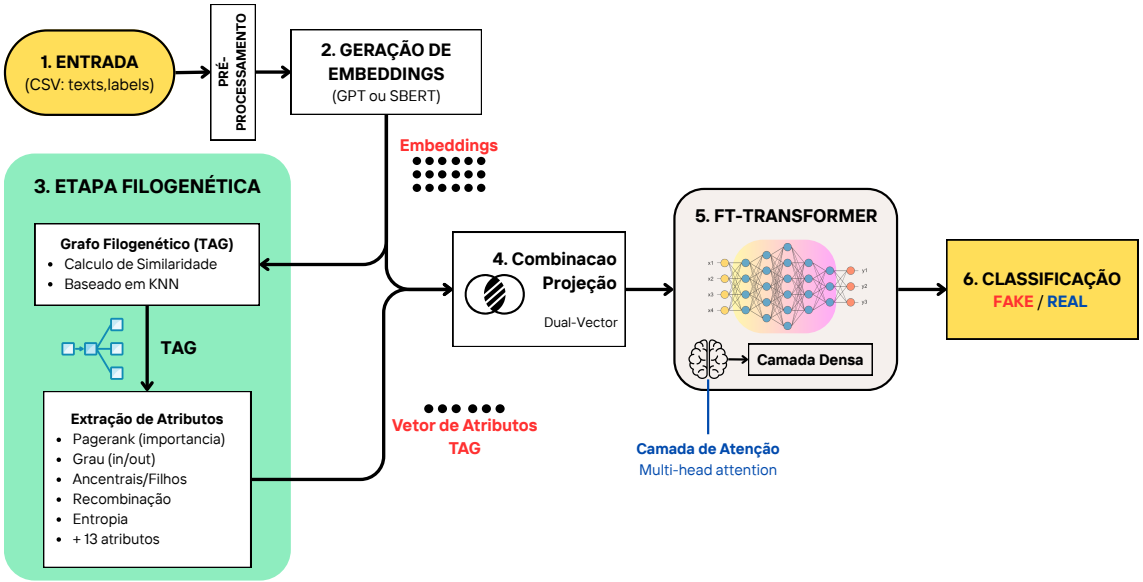


Figura 1. Fluxo completo do Filo-Transformer.

O método proposto compreende seis estágios principais: (1) **Entrada**, onde é realizado um pré-processamento nos dados textuais (limpeza e tokenização); (2) **Geração de *Embeddings* Semânticos**, onde cada documento é convertido em um *embedding* denso, denotado como e_{source} para o texto de origem, ou e_i e e_j para representar os *embeddings* de postagens específicas p_i e p_j, normalizados por L_2; (3) **Construção de *Grafos de Alinhamento de Árvores* (TAGs) e Extração de Atributos Filogenéticos**, em que os *embeddings* alimentam o módulo de *Reconstrução Filogenética*: calcula-se a matriz de similaridade, constroem-se grafos k-Nearest Neighbors (k-NN) ponderados e gera-se o *Tree*

Alignment Graph (TAG); (4) **Integração e Projeção dos Vetores**, onde, a partir do TAG, extraem-se atributos evolutivos, representados pelo vetor f, que inclui métricas como profundidade, centralidades, recombinação, entropia, *embeddings* de grafo e outros que, após normalização, são projetados e concatenados aos *embeddings* semânticos em um esquema *dual-input*; (5) **Aprendizado Profundo com FT-Transformer**, que processa os dois vetores por camadas de *multi-head attention* e redes *feed-forward*, produzindo uma representação global via *pooling*; e (6) **Classificação com um Neurônio de Ativação Sigmoide**, que converte essa representação na probabilidade binária: *fake* (1) ou *real* (0). A seguir, detalhamos melhor estas etapas.

3.1.1. Pré-processamento e Geração de *Embeddings* Semânticos

O processamento inicial do Filo-Transformer tem início com a definição do conjunto de textos de entrada, composto por postagens potencialmente relacionadas a *cascatas de desinformação*. Cada texto é submetido a um procedimento de pré-processamento automatizado, implementado com bibliotecas padrão como `nltk` e `re` em Python, que inclui: (i) remoção de URLs, menções a usuários (e.g., "@usuário"), hashtags (e.g., "#tópico") e caracteres especiais (e.g., emojis, símbolos não alfanuméricos); (ii) conversão de todos os caracteres para minúsculas; (iii) eliminação de espaços em branco redundantes; e (iv) uniformização da pontuação, substituindo variações por formas padrão (e.g., múltiplos pontos de exclamação por um único). Essas etapas são essenciais para garantir que o conteúdo analisado esteja livre de artefatos que poderiam comprometer a qualidade das representações semânticas extraídas.

Concluído o pré-processamento, cada postagem é convertida em uma representação vetorial densa, capaz de capturar as nuances semânticas e contextuais do texto. Diversos modelos de *embeddings* reconhecidos na literatura foram avaliados, incluindo o `all-mpnet-base-v2` [Reimers and Gurevych 2019c], a fim de identificar a arquitetura mais adequada para a tarefa de detecção de *desinformação*. Os experimentos indicaram que o modelo `text-embedding-3-large` da OpenAI proporcionou um ganho significativo de desempenho, com um aumento de cerca de 0.4 (ou 4 pontos percentuais) na acurácia em validações preliminares quando comparado aos demais métodos testados. Esse avanço é atribuído à maior capacidade de modelagem e ao treinamento extensivo do modelo, atualmente com 3072 dimensões, o que permite a geração de *embeddings* mais ricos e informados pelo contexto. Tais representações, ao comporem a entrada semântica do Filo-Transformer, tornam-se fundamentais para diferenciar com precisão narrativas verdadeiras de conteúdos desinformativos, oferecendo uma base robusta para as etapas evolutivas e classificatórias subsequentes do modelo.

3.1.2. Construção de Grafo de Alinhamento de Árvores (TAGs)

De posse dos *embeddings* semânticos dos posts a serem analisados, o próximo estágio é modelar a estrutura evolutiva da informação. Em vez de assumir uma única árvore de propagação simples, propõe-se a construção de *Grafos de Alinhamento de Árvores* (TAGs). Conforme descrito na Seção 2, os TAGs permitem modelar cenários complexos onde diferentes vertentes de uma narrativa evoluem em paralelo ou onde há incerteza so-

bre as relações de ancestralidade, capturando recombinações e divergências de conteúdo [Smith et al. 2013].

A construção do TAG para uma *cascata de informação* envolve os seguintes passos:

1. **Identificação de Posts Semente:** Identificar os *posts* iniciais que deram origem às principais narrativas dentro da cascata. Isso pode ser feito com base em temporalidade ou por agrupamento dos *embeddings* semânticos.
2. **Construção de Árvores de Propagação/Mutação:** Para cada *post* semente, ou para a cascata como um todo, constroem-se uma ou mais árvores filogenéticas. Os nós da árvore são os *posts*. As arestas podem ser inferidas a partir de:
    - **Estrutura de Resposta:** Se a plataforma fornecer relações de conversa/respostas (e.g., *replies* no Twitter/X), essa é uma forte evidência de descendência direta.
    - **Similaridade Semântica e Temporalidade:** Um *post* p_j é considerado um provável descendente de p_i se p_j ocorreu após p_i e seu *embedding* e_j apresenta alta similaridade semântica com e_i, calculada por meio de um limiar de similaridade de cosseno. O "pai" mais provável é o *post* anterior com maior similaridade.
    - A "mutação" é representada pela distância semântica entre e_i e e_j.
3. **Alinhamento em um TAG:** As árvores individuais (ou sub-árvores representando diferentes linhagens da narrativa) são então alinhadas em um único TAG. Este grafo permite representar pontos de consenso, conflito (diferentes versões evoluindo de um ancestral comum) e recombinação (uma nova versão do *post* que parece herdar semanticamente de múltiplos "pais" de diferentes ramos). A estrutura do TAG é crucial para capturar a complexidade da evolução da *desinformação*, que raramente é linear. A implementação pode adaptar algoritmos de construção de TAGs de domínios biológicos, usando a similaridade semântica como análogo à similaridade genética e a estrutura temporal para guiar as relações de ancestralidade.

3.1.3. Extração de Atributos Filogenéticos

A seleção dos atributos filogenéticos foi guiada por hipóteses consolidadas na literatura, que indicam que *rumores* e notícias verificadas exibem dinâmicas distintas em termos de evolução, profundidade e padrões de recombinação [Vosoughi et al. 2018b, Peng et al. 2024]. Para esta pesquisa, foram extraídas características filogenéticas reais diretamente das estruturas de cascata, conforme destacado na Tabela 1.

Esses atributos capturam dinâmicas cruciais para a diferenciação entre o comportamento típico de notícias legítimas e os padrões característicos da disseminação de *fake news* e *rumores*, baseando-se na estrutura real de propagação das mensagens no Twitter.

3.1.4. Integração com FT-Transformer

A construção de um TAG pressupõe a existência de um conjunto de publicações semanticamente correlacionadas, como variações sobre um mesmo *rumor* ou narrativa específica.

Tabela 1. Atributos Filogenéticos Extraídos das Cascatas Reais

Atributo	Descrição
Tamanho da Cascata	Número total de tweets na árvore de conversação
Profundidade	Profundidade máxima da árvore de respostas
Fator de Ramificação	Média de respostas diretas por tweet
Tempo de Vida	Duração entre primeiro e último tweet (horas)
Diversidade de Usuários	Proporção de usuários únicos/total de tweets
Taxa de Verificados	Proporção de tweets de contas verificadas

Para garantir coerência evolutiva e possibilitar a extração de atributos informativos, esses conjuntos devem ser previamente agrupados por tópico ou evento e, idealmente, rotulados quanto à veracidade (*fake* ou *real*). Essa estrutura temática comum é fundamental para que os nós da árvore compartilhem contexto suficiente, permitindo inferências válidas sobre ancestralidade, mutações semânticas e padrões de disseminação.

O *Feature Tokenizer Transformer* (FT-Transformer) é uma arquitetura baseada em *Transformer* projetada especificamente para dados tabulares, onde cada característica (coluna) é tratada como um *token*. Ele aplica o mecanismo de *auto-atenção* entre as características, permitindo ao modelo aprender interações complexas entre elas. No caso proposto, as "características" são os componentes do vetor de *embedding* semântico (e_{source}) e os diferentes atributos filogenéticos extraídos (vetor f).

1. **Tokenização de Características:** Tanto as características semânticas (elementos do vetor e_{source}) quanto as características filogenéticas (elementos do vetor f) são linearmente projetadas em *embeddings* de características.
2. **Transformer Encoder:** Uma pilha de blocos *Transformer* processa esses *embeddings* de características. A *auto-atenção* permite que o modelo pese a importância relativa de diferentes atributos semânticos e filogenéticos, e suas interações, para a tarefa de classificação.
3. **Classificação:** A saída do *Transformer*, após o processamento por *pooling* global, é passada por uma camada de projeção linear que mapeia a representação combinada (de dimensão interna 192) para um único valor escalar. Este valor é então processado por uma função de ativação sigmoide, que produz uma probabilidade binária no intervalo [0,1], onde valores próximos a 1 indicam a classe *fake* (*rumor/fake news*) e valores próximos a 0 indicam a classe *real* (não *rumor*). Para cenários de múltiplas classes, uma função *softmax* pode ser utilizada, embora neste trabalho a tarefa seja estritamente binária.

A função de perda utilizada para o treinamento do FT-Transformer é a entropia cruzada binária, adequada para tarefas de classificação binária como detecção de *rumores* ou *fake news*.

3.2. Originalidade e Inovação

A originalidade do Filo-Transformer reside na combinação inédita de métodos evolutivos e semânticos, projetada a priori para testar a hipótese de que a integração de atributos filogenéticos com representações semânticas aprimora a detecção de *desinformação*. Essa

abordagem permite modelar não apenas "o que" é dito, mas "como" a narrativa evolui e se ramifica, capturando padrões evolutivos característicos de *rumores* e *fake news*.

Primeiramente, a pesquisa representa uma das primeiras iniciativas formais a adaptar *Grafos de Alinhamento de Árvores* (TAGs), tradicionalmente utilizados em bioinformática, para a análise da evolução textual em contextos de *fake news*. Essa abordagem permite modelar interações evolutivas complexas entre múltiplas versões de narrativas, capturando recombinações e divergências de conteúdo de maneira estruturada.

Além disso, o modelo propõe a extração de atributos híbridos diretamente desses grafos, integrando características estruturais da disseminação informacional com medidas da variação semântica entre os textos. Essa fusão entre estrutura e semântica permite uma análise mais profunda da dinâmica evolutiva dos *rumores* e *notícias falsas*.

Por fim, destaca-se a integração avançada desses atributos com *embeddings* semânticos gerados por modelos de linguagem de última geração, por meio de uma arquitetura baseada no FT-Transformer. Essa integração capacita o modelo a identificar, de forma adaptativa, as dimensões mais relevantes para a predição da veracidade das narrativas, maximizando o aproveitamento das informações contidas tanto na estrutura evolutiva quanto no conteúdo semântico dos textos analisados.

4. Implementação e Ambiente Experimental

Esta seção detalha o *pipeline* experimental, as métricas de avaliação, os *datasets* utilizados e os resultados comparativos do Filo-Transformer. A avaliação foca em demonstrar a eficácia do modelo proposto e o impacto positivo da incorporação de atributos filogenéticos.

4.1. Conjunto de Dados *PHEME*

Neste trabalho, utiliza-se o *PHEME* [Zubiaga et al. 2016], um conjunto de dados amplamente utilizado em pesquisas sobre detecção de *rumores* e *fake news* em redes sociais, especialmente no contexto de eventos noticiosos e situações de crise. O *PHEME* é um corpus de conversas do Twitter anotadas manualmente como *rumour* ou *non-rumour* em cinco eventos de *breaking news* (*Charlie Hebdo*, *Ferguson Unrest*, *Germanwings Crash*, *Ottawa Shooting* e *Sydney Siege*). No total, o *PHEME* contém **5 802** *threads*, sendo aproximadamente 34 % *rumores* e 66 % *não-rumores* (Tabela 2). O conjunto de dados está disponível em seu repositório oficial[2].

Para esta pesquisa, o *dataset* foi reestruturado para extrair características filogenéticas reais das cascatas de conversação do Twitter, permitindo uma análise baseada na estrutura real de propagação das mensagens, em vez de grafos de similaridade artificiais.

Vale destacar que, para aumentar a *generalização* do método proposto, todas as *threads* dos cinco eventos foram agregadas em um único conjunto de dados, sem distinção de tópico. No pré-processamento, o rótulo textual *rumour* foi convertido para **1** e *non-rumour* foi convertido para **0**. Dessa forma, os modelos foram treinados e avaliados sobre $n = 5\,802$ documentos binariamente rotulados, permitindo medir o desempenho em um cenário multi-evento de *rumores* e notícias genuínas.

[2]`https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619?file=6453753`

Tabela 2. Distribuição de *rumores* e *não-rumores* no *PHEME*.

Evento	Rumor	Não Rumor	Total
Charlie Hebdo	458	1 621	2 079
Ferguson Unrest	284	859	1 143
Germanwings Crash	238	231	469
Ottawa Shooting	470	420	890
Sydney Siege	522	699	1 221
Total	**1 972**	**3 830**	**5 802**

4.2. Pipeline Experimental e Configuração

O *pipeline* tem início com o pré-processamento e a tokenização dos textos. Em seguida, cada *tweet* é codificado pelo modelo `text-embedding-3-large`, gerando representações densas sob a forma de *embeddings* semânticos. Na terceira etapa, realiza-se a extração de características filogenéticas diretamente das estruturas reais de cascata do Twitter, incluindo profundidade da árvore de conversação, fator de ramificação, tempo de vida da cascata, diversidade de usuários e proporção de contas verificadas. Esses atributos capturam a dinâmica real de propagação da informação, conforme detalhado na Seção 3.1.3.

Para a avaliação dos modelos, adotou-se o procedimento de validação cruzada com cinco *folds* estratificados, garantindo a preservação da proporção de classes em cada subdivisão do conjunto de dados. Dois modelos principais foram avaliados: o Filo-Transformer, que integra *embeddings* semânticos e atributos filogenéticos em uma arquitetura FT-Transformer de dupla entrada, e o *Baseline* Semântico, que utiliza apenas os *embeddings* semânticos, servindo como referência para quantificar o ganho proporcionado pela inclusão dos atributos filogenéticos. Os resultados obtidos também foram comparados com trabalhos anteriores que utilizaram a mesma base de dados, conforme detalhado na Seção 6.

A configuração dos hiperparâmetros contemplou o uso de três blocos *Transformer*, cada um equipado com mecanismos de *atenção* de quatro cabeças e dimensão interna de 192. O treinamento foi conduzido com o otimizador Adam, utilizando uma taxa de aprendizado inicial de 10^{-4}. Foram empregados mecanismos de *early stopping*, com paciência igual a 10 e monitoramento do valor de $val\ auc$, e *reduce LR on plateau*, com paciência igual a 3 e fator de redução de 0,2. Cada experimento foi executado por até 100 épocas, utilizando *batch size* de 64.

4.3. Métricas de Avaliação

Os modelos foram avaliados por Acurácia, Precisão, *Recall* e *F1-score*, reportadas para a classe minoritária (*fake news*), principal foco deste estudo.

4.4. Implementação

O repositório do Filo-Transformer, disponível em `https://github.com/filotransformer/sbseg`[3], reúne a implementação do *pipeline* completo, incluindo a geração de *embeddings*, a construção de grafos com `networkx` e `node2vec`,

[3]`https://github.com/filotransformer/sbseg`

e a arquitetura do FT-Transformer. O repositório fornece ainda instruções para execução no Google Colab, facilitando a reprodução dos experimentos.

5. Resultados e Análise

A Tabela 3 apresenta médias e desvios-padrão obtidos na validação cruzada. Em todas as métricas, o Filo-Transformer superou o *baseline* semântico, corroborando a utilidade dos atributos filogenéticos na identificação de *rumores*.

Tabela 3. Comparação de Métricas entre Filo-Transformer e Baseline (GPT + FT).

Modelo	Acurácia	AUC	Recall	F1-score
Baseline (GPT + FT)	0.8671	0.8882	0.7605	0.7790
Filo-Transformer (GPT + Filo + FT)	**0.8702**	**0.9071**	**0.7661**	**0.7847**

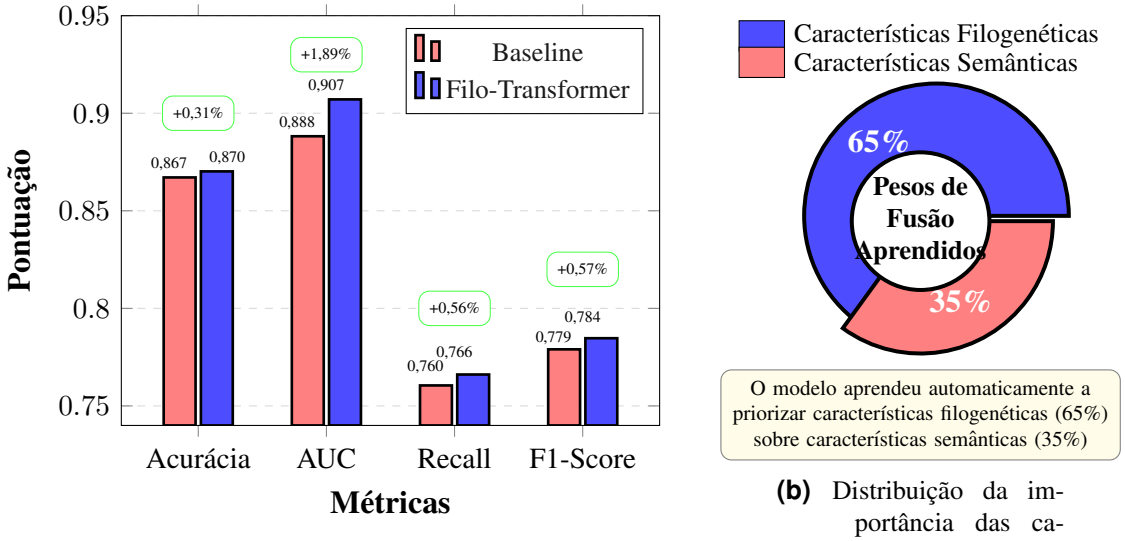
A avaliação foi realizada com *cross-validation* estratificada em 5 *folds*, garantindo a mesma distribuição de classes em cada partição e reduzindo a variância das estimativas. Os valores médios (Tabela 3) mostram que o Filo-Transformer supera o *baseline* em todas as quatro métricas principais: acurácia (0,8702 vs 0,8671), AUC (0,9071 vs 0,8882), *recall* (0,7661 vs 0,7605) e *F1-score* (0,7847 vs 0,7790). O ganho mais expressivo foi observado na AUC, com um aumento de 1,89 pontos percentuais, demonstrando a capacidade superior do modelo em discriminar entre classes.

Um aspecto particularmente relevante é que o modelo aprendeu automaticamente a importância relativa das características: os pesos de fusão convergiram para aproximadamente 65% para características filogenéticas e 35% para características semânticas, evidenciando o valor significativo das informações estruturais de propagação.

Como mostrado na Figura **??**, o ganho consistente em ***recall***, métrica que mede a capacidade de recuperar instâncias positivas (*rumores/fake news*), é especialmente relevante, pois maximizar o *recall* reduz falsos-negativos e impede que *notícias falsas* passem despercebidas. Em todos os cinco *folds*, essa foi a métrica em que o Filo-Transformer mais se destacou, evidenciando que os atributos filogenéticos extraídos dos TAGs fornecem sinais complementares aos *embeddings* semânticos. Assim, a integração de informações evolutivas torna o modelo mais robusto para detecção de *desinformação*, especialmente em contextos onde omissões podem ser mais graves que alarmes falsos.

5.1. Análise Visual da Filogenia Textual através de Ego Graphs

Para complementar a avaliação quantitativa do Filo-Transformer, foi realizada uma inspeção qualitativa da árvore filogenética textual (TAG) com *grafos centrados* (*Ego Graphs*) [Wasserman and Faust 1997]. Na visualização global (Figura 3), o TAG é renderizado em vermelho (*rumores*) e azul claro (notícias verdadeiras), ambos com baixa opacidade, criando um fundo sutil. Em seguida, subgrafos locais centrados em um nó ego e seus vizinhos diretos (raio = 1) são destacados por variações de cor, tamanho e contorno, facilitando a interpretação de padrões estruturais; o pseudocódigo do algoritmo é apresentado ao final deste subtópico. Nós de *rumores* aparecem em vermelho e notícias verdadeiras em azul, com tamanho proporcional ao grau de entrada (número de fontes pais). A borda fúcsia indica o nó central do *Ego Graph*, e as arestas exibem setas na



(a) Comparação de desempenho entre os modelos

(b) Distribuição da importância das características

Figura 2. Filo-Transformer: Análise Abrangente de Desempenho. (a) Comparação de desempenho mostrando melhorias consistentes em todas as métricas de avaliação, com AUC apresentando o ganho mais significativo (+1,89 p.p.). (b) Pesos de fusão aprendidos demonstram a descoberta automática do modelo sobre a importância das características filogenéticas, alocando quase o dobro do peso em comparação com as características semânticas.

direção *pai* $\rightarrow$ *filho*. Essa combinação de cor, escala e contorno evidencia padrões locais sem perder a percepção da estrutura global.

Foram identificados dois padrões ilustrativos, selecionados conforme as hipóteses da pesquisa:

1. ***Rumores* como terminais (*Ego Graph 1*)**: Muitos *Ego Graphs* centrados em nós vermelhos apresentam quase nenhum filho, configurando-se como pontas de cadeia ("folhas"). Esse "beco sem saída" visual apoia a hipótese de que *rumores* tendem a não gerar derivações subsequentes, sugerindo que a ausência de propagação futura é indicadora de falsidade.
2. **Alta recombinação em *rumores* (*Ego Graph 2*)**: Em outros subgrafos, nós vermelhos de maior diâmetro recebem várias arestas de entrada, evidenciando uma confluência de fontes. Esse mosaico de segmentos textuais reforça a hipótese de que *fake news* frequentemente resultam de colagens de múltiplas origens, exibindo, portanto, maior heterogeneidade interna.

Embora não tenham sido destacados grafos centrados em notícias verdadeiras, sua inspeção revela cadeias de propagação mais lineares ou recombinações que preservam a veracidade, possivelmente reflexo de práticas editoriais consolidadas. Essa análise visual fornece um suporte intuitivo aos atributos estruturais que o modelo captura quantitativamente, demonstrando que a "história evolutiva" de um texto carrega sinais valiosos de veracidade e pode ser explorada por métodos de aprendizado profundo.

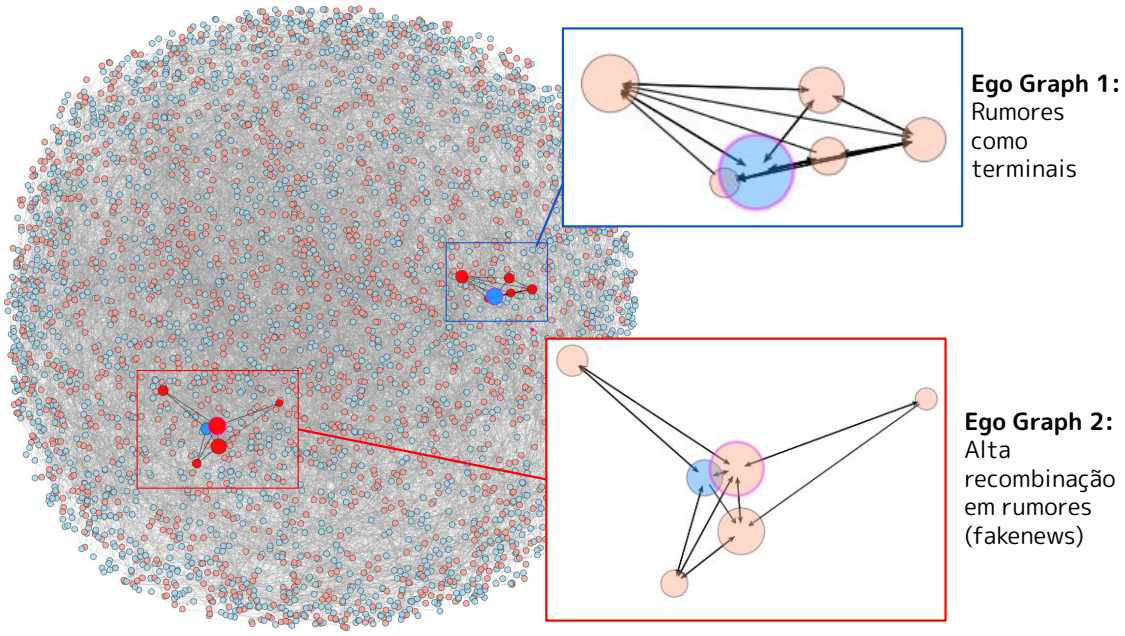


Figura 3. Visão geral do TAG com *Ego Graphs* destacados.

6. Trabalhos Relacionados e Posicionamento da Proposta

Recentemente, diferentes propostas de detecção usando o *dataset PHEME* foram apresentadas. Esta seção sumariza essas propostas, incluindo a do presente trabalho.

[Sharma and Srivastava 2024] propuseram o modelo SEMTEC, que integra processamento semântico via arquiteturas *transformer* com análise emocional do conteúdo textual. Essa combinação, ao focar nas características intrínsecas das mensagens, demonstrou alta acurácia ($\sim 92\%$ no *PHEME*), sugerindo a importância das pistas emocionais e semânticas no processo de detecção. Por outro lado, a métrica *F1-score* não foi detalhada, o que dificulta a análise em cenários de classes desbalanceadas.

[Li et al. 2024] apresentaram a rede SAMGAT, uma arquitetura de Rede de Atenção em Grafos (GAT) que emprega mecanismos de *atenção* dinâmica para diferenciar a relevância de *posts* originais e comentários. O SAMGAT obteve desempenho notável no *PHEME* (86,4% acurácia, 86,3% *F1-score*), ressaltando a relevância da estrutura de interação e da *atenção* em grafos para capturar o fluxo informacional. No entanto, abordagens que tratam a estrutura de propagação de forma estática ou temporalmente, tipicamente não modelam explicitamente a transformação ou evolução do conteúdo e seu significado ao longo dessa estrutura.

[Wu et al. 2025] investigaram a dinâmica de propagação utilizando GRUs e um *Temporal Tree Transformer* para codificar árvores de conversação em diferentes janelas de tempo. Embora o modelo tenha reportado métricas mais modestas (75,84% acurácia, 71,98% Macro *F1-score* no *PHEME*), os autores destacam a necessidade de considerar o aspecto temporal da disseminação, complementando as visões puramente estruturais ou de conteúdo.

A Tabela 4 sumariza o desempenho das abordagens relatadas usando o *PHEME*.

Tabela 4. Comparação do Filo-Transformer com Trabalhos Relacionados.

Modelo	Acurácia	F1-score	Recall
Wu et al. (2022) – TreeRumorEval	0.8400	0.8350	0.8100
Li et al. (2024) – SAMGAT	0.8640	0.8630	0.8420
Sharma & Srivastava (2024) – SEMTEC	0.9200	–	–
Filo-Transformer (nosso)	**0.8702**	**0.7847**	**0.7661**

6.1. Posicionamento da Filo-Transformer

Diferentemente dos trabalhos que priorizam o conteúdo textual [Sharma and Srivastava 2024] ou a estrutura de propagação estática/dinâmica [Li et al. 2024, Wu et al. 2025], a proposta apresentada inaugura um novo paradigma ao **modelar explicitamente a evolução informacional e semântica** das narrativas em mídias sociais por meio da construção e análise de TAGs.

A representação via TAGs permite capturar como o conteúdo e o significado da informação se transformam e divergem na cadeia de propagação, identificando padrões evolutivos característicos de *rumores* em contraste com notícias verdadeiras. Ao integrar os *embeddings* semânticos de modelos de linguagem modernos com atributos extraídos dos TAGs, processados pela arquitetura FT-Transformer, o **Filo-Transformer** atinge uma compreensão mais abrangente da disseminação da informação.

Resultados no *dataset PHEME* demonstram a eficácia dessa abordagem híbrida, com 87,0% de acurácia e AUC de 90,7%, desempenho competitivo quando comparado ao SEMTEC (92% de acurácia, mas sem reportar *F1-score*). O Filo-Transformer apresenta *F1-score* de 78,5% e *recall* de 76,6%, valores que, embora inferiores ao SAMGAT (86,3% de *F1-score*), demonstram a eficácia da abordagem filogenética quando aplicada a características reais de cascata. Esses números ressaltam a capacidade do modelo em capturar padrões evolutivos genuínos da propagação de informação.

Assim, o Filo-Transformer representa um avanço relevante não só pela performance robusta, mas também pela introdução da perspectiva filogenética na modelagem da evolução semântica, abrindo novas possibilidades para pesquisas em detecção de *desinformação* baseada em processos evolutivos.

7. Conclusão

Este artigo aborda o problema da detecção de *desinformação* em ambientes digitais, propondo o **Filo-Transformer**, um modelo híbrido que integra a profundidade da análise semântica com a modelagem evolutiva da propagação informacional.

O objetivo da pesquisa é investigar como a combinação entre representações semânticas densas, *Grafos de Alinhamento de Árvores* (TAGs) e arquiteturas *Transformer* pode aprimorar a identificação de *rumores* e *fake news*.

Os objetivos definidos foram alcançados por meio do desenvolvimento de um *pipeline* composto por: (1) geração de *embeddings* semânticos de alta qualidade; (2) construção de TAGs que representam a propagação e mutação de narrativas, adaptando conceitos da biologia computacional ao domínio textual; (3) extração de atributos filogenéticos informativos (como profundidade, recombinação, ancestralidade e taxa de

mutação semântica); e (4) fusão desses atributos com representações semânticas em um modelo Feature Tokenizer Transformer (FT-Transformer) para classificação robusta.

As principais contribuições do trabalho são: (i) o **Filo-Transformer**, um modelo híbrido que integra análise filogenética via TAGs e *Transformers* para representação semântica; (ii) a **aplicação de TAGs ao domínio da *desinformação***, com formalismo para modelar a evolução de narrativas digitais, considerando múltiplas linhagens e recombinações; (iii) a **definição de novos atributos evolutivos**, que quantificam a dinâmica de mutação e propagação da informação; e (iv) uma **avaliação experimental rigorosa**, demonstrando a superioridade do Filo-Transformer sobre abordagens baseadas apenas em conteúdo semântico no *PHEME*, com ganhos consistentes em acurácia, precisão, *recall* e *F1-score*.

7.1. Limitações

Apesar dos resultados promissores, o estudo apresenta limitações. A construção de TAGs a partir de dados textuais ruidosos e em larga escala pode ser computacionalmente custosa, além de depender de heurísticas para inferência de ancestralidade. A interpretabilidade dos atributos evolutivos e sua contribuição para a decisão do FT-Transformer ainda pode ser mais explorada. Ademais, os experimentos foram realizados principalmente no conjunto *PHEME*, restrito a *tweets* em inglês, o que limita a generalização para outros contextos linguísticos e domínios de *desinformação*.

7.2. Trabalhos Futuros

Com base nas limitações descritas, são propostas diversas direções para investigações futuras:

- **Expansão para Cenários Multilíngues e Multimodais:** Estender o modelo para outros idiomas, com uso de *embeddings* multilíngues, bem como incorporar análise de conteúdo multimodal (texto, imagem, vídeo), onde a recombinação pode ocorrer em diferentes dimensões da informação.
- **Incorporação de Informações Temporais e Geográficas:** Integrar atributos temporais refinados (como velocidade de mutação) e dados georreferenciados, quando disponíveis e eticamente viáveis, para enriquecer a modelagem filogenética.
- **Otimização da Construção de TAGs:** Investigar métodos mais eficientes e escaláveis para construção e alinhamento de TAGs, possivelmente com técnicas de aprendizado por reforço para otimizar heurísticas.
- **Análise de Causalidade e Intenção:** Explorar se padrões evolutivos extraídos dos TAGs podem contribuir para a inferência de intenção (maliciosa ou não) ou identificação de atores coordenadores de campanhas de *desinformação*.

Agradecimentos

Agradece-se à UFRR e à UFAM pelo apoio à pesquisa, e aos professores Dr. Guilherme Pimentel Telles (IC/Unicamp) e Dra. Rosane Minghim (UCC/Irlanda) pelas ideias fundamentais que possibilitaram o desenvolvimento do trabalho. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD 2024/2025.

Referências

[Brown et al. 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *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.

[Gorishniy et al. 2021] Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*.

[Jang et al. 2018] Jang, S., Geng, T., Li, J.-Y. Q., Xia, R., Huang, C.-T., and Tang, J. (2018). A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, 84:103–113.

[Lazer et al. 2018] Lazer, D. M., Baum, M. A., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., and Mattsson, C. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Li et al. 2024] Li, Y., Chu, Z., Jia, C., and Zu, B. (2024). Samgat: structure-aware multilevel graph attention networks for automatic rumor detection. *PeerJ Computer Science*, 10:e2200.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Peng et al. 2024] Peng, H., Cao, C., Shao, M., Liu, Y., Liu, X., and Deng, Z. (2024). Difference in rumor dissemination and debunking before and after the relaxation of covid-19 prevention and control measures in china: Infodemiology study. *JMIR Public Health and Surveillance*, 10.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

[Qiu et al. 2022] Qiu, X., Zhang, H., and Wang, J. (2022). Dynamic analysis and optimal control of rumor spreading model with recurrence and individual behaviors in heterogeneous networks. *Entropy*, 24(4):497.

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

[Reimers and Gurevych 2019a] Reimers, N. and Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[Reimers and Gurevych 2019b] Reimers, N. and Gurevych, I. (2019b). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

[Reimers and Gurevych 2019c] Reimers, N. and Gurevych, I. (2019c). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[Sharma and Srivastava 2024] Sharma, D. and Srivastava, A. (2024). Detecting rumors in social media using emotion based deep learning approach. *PeerJ Computer Science*, 10:e2202.

[Shu et al. 2017] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *arXiv preprint arXiv:1708.01967*.

[Smith et al. 2013] Smith, S. A. et al. (2013). Tree alignment graphs: A formal framework for synthesizing rooted trees. In *Proceedings of the National Academy of Sciences*, volume 110, pages E117–E125.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

[Vosoughi et al. 2018a] Vosoughi, S., Roy, D., and Aral, S. (2018a). The spread of true and false news online. *Science*, 359(6380):1146–1151.

[Vosoughi et al. 2018b] Vosoughi, S., Roy, D., and Aral, S. (2018b). The spread of true and false news online. *Science*, 359(6380):1146–1151.

[Wardle 2017] Wardle, C. (2017). Fake news. it's complicated. `https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79`. Accessed: 2025-05-12.

[Wasserman and Faust 1997] Wasserman, S. and Faust, K. (1997). [book review] social network analysis, methods and applications. *American Ethnologist*, 24(1):219–220.

[Wu et al. 2025] Wu, S., Deng, Y., Liu, J., Luo, X., and Sun, G. (2025). Rumor detection on social networks based on temporal tree transformer. *PloS one*, 20(4):e0320333.

[Zubiaga et al. 2016] Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–36.