COMBATE À DESINFORMAÇÃO COM FACTUAL: UMA SOLUÇÃO BASEADA EM MODELOS DE LINGUAGEM DE GRANDE ESCALA

Verônica Souza dos Santos[1], Paulo Henrique Batista Rodrigues[1], Geraldo Pereira Rocha Filho[2], Edna Dias Canedo[1], Frederico de Almeida Meirelles Palma[3] e Fábio Lúcio Lopes de Mendonça[1]
[1]Universidade de Brasília – UnB, Brasil
[2]Universidade Estadual do Sudoeste da Bahia (UESB), Brasil
[3]Coordenação Geral da Rede SUAS – MDS, Brasil

RESUMO

Com o crescimento das tecnologias digitais e das redes sociais, a disseminação de fake news tornou-se um problema significativo, afetando a sociedade em áreas como política, saúde e processos democráticos. Esta pesquisa aborda duas lacunas fundamentais: (i) a ausência de uma categorização sistemática das diferentes formas de desinformação; e (ii) a falta de modelos de grandes linguagens de modelos (LLMs) capazes de generalizar eficientemente em múltiplos domínios e conjuntos de dados heterogêneos. Diante disso, este trabalho apresenta o FACTUAL, uma solução baseada em LLMs para detectar e mitigar fake news. O FACTUAL utiliza o modelo LLaMA e frameworks assíncronos como Asyncio e AioSQLite para processar grandes volumes de dados textuais de forma eficiente. A solução inclui uma camada de banco de dados para armazenar os resultados e um middleware que integra a análise de IA oferecendo classificações confiáveis e justificativas claras. Os resultados demonstram que o FACTUAL é eficaz na identificação de padrões de desinformação, apresentando uma boa precisão e confiabilidade nos resultados.

PALAVRAS-CHAVE

Fake News, Desinformação, Redes Sociais, Verificação de Autenticidade, Manipulação de Informações

1. INTRODUÇÃO

A disseminação de fake news, embora não seja um fenômeno recente, aumentou com o crescimento das tecnologias digitais e a popularização das redes sociais. Aliado com o avanço da internet, as fakes news se tornaram mais frequentes, já que qualquer indivíduo agora pode disseminar informações sem controle, utilizando redes sociais, tais como Facebook e Twitter (Alnabhan and Branco, 2024). Um exemplo relevante no Brasil ocorreu durante as eleições presidenciais de 2018, quando uma série de informações falsas sobre urnas eletrônicas foi amplamente divulgada, gerando desconfiança pública e afetando o debate político nacional (Bernardi, 2021; iG São Paulo, 2018). As fakes news, caracterizadas pela propagação rápida de informações fabricadas ou distorcidas, alcançam e influenciam um público maior do que as notícias verdadeiras. Esse conteúdo, frequentemente apresentado como factual, tem sido utilizado para manipular opiniões e obter vantagens políticas, sociais e econômicas, com impactos diretos sobre a percepção pública e os processos decisórios em diversas áreas.

No Brasil, o Marco Civil da Internet, regulamentado pela Lei nº 12.965/2014, estabelece diretrizes para a liberdade de expressão e proteção de dados na internet. No entanto, essa legislação ainda apresenta limitações no combate às fake news, especialmente no que diz respeito à remoção ágil de informações falsas e à responsabilização de provedores de conteúdo. Além das questões locais, eventos globais como as eleições presidenciais dos EUA em 2016 demonstraram a gravidade global do problema, evidenciando como a desinformação afeta os processos democráticos e reforçando a necessidade de regulamentações e tecnologias mais eficazes (Alnabhan and Branco, 2024; Sharma et al., 2019). Esses cenários, tanto local quanto global, destacam a urgência de novas pesquisas e abordagens para mitigar a propagação de fake news, especialmente em situações em que a rapidez da disseminação pode gerar consequências graves.

Detectar fake news não é uma tarefa trivial devido a diversos fatores, tais como a ambiguidade e subjetividade do conteúdo, a rapidez com que essas informações se propagam, a criação de conteúdo sofisticado que dificulta a diferenciação entre verdadeiro e falso, e a falta de rótulos e dados confiáveis (Alnabhan and Branco, 2024; Sharma et al., 2019). Em razão disso, a Inteligência Artificial (IA) surge como uma solução promissora para o combate à desinformação. Utilizando Modelos de Linguagem de Grande Escala (LLMs), a IA consegue analisar grandes volumes de dados em tempo real, identificando padrões e sinais de fake news. A complexidade dos LLMs permite lidar com a ambiguidade e a subjetividade do conteúdo, facilitando a detecção de informações falsas mesmo em meio a textos sofisticados e rapidamente disseminados (Sun et al., 2024; Wu et al., 2024). LLMs, como o LLAMA, Mistral e Phi, possuem uma capacidade avançada de compreender nuances no conteúdo textual, ajudando a identificar informações potencialmente falsas e fornecendo uma análise aprofundada do contexto, mesmo em casos de ambiguidade e conteúdo sofisticado. Além disso, tecnologias como o processamento de linguagem natural permitem que modelos de IA compreendam o contexto e o conteúdo de um texto, auxiliando na verificação da veracidade das informações.

Diversos trabalhos têm explorado o problema da detecção de fake news em diferentes contextos e abordagens. Em (Hu et al., 2024) e (Wang et al., 2023) exploram o uso de LLMs e abordagens multiespecializadas para ampliar a detecção de fake news em múltiplos domínios. No entanto, essas abordagens enfrentam limitações relacionadas à complexidade computacional e à introdução de vieses. Por outro lado, os trabalhos de (Kaliyar et al., 2021) e (Segura-Bedmar and Alonso-Bartolome, 2022) focam em arquiteturas baseadas em redes neurais convolucionais (CNNs) e abordagens multimodais, combinando texto e imagem para aumentar a acurácia na classificação de notícias falsas, embora o desequilíbrio de classes nos datasets utilizados limite sua eficácia em alguns cenários. Em um contexto diferente, raja2023fake aborda a detecção de fake news em línguas de poucos recursos, utilizando aprendizado por transferência com modelos transformadores, enquanto dou2021user propõe um framework que combina preferências de usuários e redes sociais para melhorar a detecção de fake news. No entanto, a complexidade computacional e a dependência de dados históricos ainda são desafios comuns a essas abordagens, limitando sua aplicabilidade em cenários reais e escaláveis.

Vale destacar, no entanto, que esta pesquisa foca em duas lacunas principais: (i) a falta de uma categorização das diferentes formas de desinformação, como desinformação deliberada, informações falsas não intencionais e rumores; e (ii) a escassez de modelos de LLMs capazes de generalizar de maneira eficaz em múltiplos domínios e conjuntos de dados heterogêneos. Essas lacunas tornam-se essenciais para aprimorar as capacidades de detecção de fake news em cenários complexos e variados, e esta pesquisa explora tais lacunas.

Com isso em mente, este trabalho apresenta o FACTUAL - Fake News Analysis with Confidence and Trust Using AI and LLM - uma solução baseada em LLMs para detectar e mitigar a disseminação de fake news. O FACTUAL foi modelado com base no modelo LLaMA, que oferece uma capacidade de análise de dados textuais por meio do framework Ollama, permitindo identificar padrões e sinais característicos de desinformação de forma eficiente. Além disso, o FACTUAL foi desenvolvido sobre frameworks assíncronos como Asyncio e AioSQLite, otimizando a integração com bancos de dados e facilitando a execução paralela de tarefas, o que resulta em uma maior escalabilidade e eficiência no processamento de grandes volumes de dados.

O restante deste trabalho está organizado da seguinte forma. Na Seção 2, discutimos os trabalhos relacionados. A Seção 3 descreve o desenvolvimento do FACTUAL, enquanto a Seção 4 aborda a avaliação de desempenho. Por fim, a Seção 5 apresenta as conclusões e direções futuras.

2. TRABALHOS RELACIONADOS

Esta seção explora diferentes trabalhos que abordam o problema da detecção de fake news. O trabalho de (Ma et al., 2024) propõe o framework Event-Radar, focado na detecção de fake news em mídias multimodais. O Event-Radar combina aprendizado multi-view e inconsistências a nível de eventos entre texto e imagem, modelando grafos de eventos para capturar relações sujeito-predicado nas notícias. Utilizando distribuições Beta, o modelo ajusta a credibilidade das modalidades. Suas limitações incluem a falta de exploração de relações causais entre eventos e a dependência de ferramentas externas para reconhecimento de entidades nomeadas.

Os trabalhos de (Hu et al., 2024) e (Wang et al., 2023) compartilham abordagens centradas no uso de LLMs e técnicas avançadas de detecção de fake news em múltiplos domínios. Em (Hu et al., 2024), os autores exploram a utilização do GPT-3.5 para atuar como conselheiros no processo de detecção de fake news. Nesse contexto, os LLMs fornecem justificativas e insights que complementam o desempenho de modelos menores, como o BERT. A proposta inclui a criação de uma Rede de Orientação Adaptativa de Justificativas (ARG), que integra as percepções geradas pelos LLMs para melhorar a capacidade de julgamento dos modelos mais leves. Embora essa abordagem adicione interpretabilidade aos modelos menores, sua dependência de infraestrutura computacional robusta representa uma limitação significativa. Além disso, a ausência de um mecanismo explícito de verificação de fatos pode deixar lacunas na detecção em casos mais complexos.

De forma semelhante, o estudo de (Wang et al., 2023) explora um sistema que busca ampliar a capacidade de detecção de fake news em múltiplos domínios, como política, entretenimento e finanças. Os autores utilizam soft-labels, uma abordagem que captura melhor as nuances de diferentes áreas, e um mecanismo de LeapGRU, projetado para ignorar palavras irrelevantes. O modelo aplica uma análise mais especializada, dividindo as notícias por domínio e utilizando grupos de especialistas para extrair características específicas. Embora esse método forneça uma análise direcionada, as limitações incluem a complexidade no tratamento de textos longos e a possível introdução de vieses, devido a interpretações subjetivas dos especialistas.

Os trabalhos de (Kaliyar et al., 2021) e (Segura-Bedmar and Alonso-Bartolome, 2022) focam em modelos multimodais e no uso de redes neurais convolucionais (CNNs) para detecção de fake news, utilizando diferentes modalidades de entrada. Em (Kaliyar et al., 2021), os autores propõem uma combinação do BERT com uma CNN, que utiliza blocos paralelos com diferentes tamanhos de kernel para capturar dependências semânticas de longa distância. Essa combinação é eficaz para lidar com ambiguidades em textos de mídias sociais, mas sua complexidade limita a aplicabilidade em larga escala, especialmente em ambientes com recursos restritos. Além disso, o treinamento desses modelos requer muitos dados rotulados, o que pode ser um desafio em domínios com baixa disponibilidade.

De forma semelhante, o artigo de (Segura-Bedmar and Alonso-Bartolome, 2022) apresenta uma abordagem multimodal que combina texto e imagem para a detecção de fake news. Utilizando o dataset Fakeddit, que contém seis categorias de notícias, e uma arquitetura baseada em CNN, os autores combinam modalidades de entrada para enriquecer a representação da notícia e melhorar a acurácia. No entanto, uma limitação significativa está no desequilíbrio das classes no dataset utilizado, o que impacta diretamente o desempenho em detectar classes minoritárias, resultando em uma redução de precisão e recall.

O estudo de (Raja et al., 2023) propõe uma abordagem específica para a detecção de fake news em línguas de poucos recursos, utilizando aprendizado por transferência com modelos transformadores pré-treinados, como o mBERT e o XLM-RoBERTa, ajustados para línguas dravidianas. A metodologia envolve fine-tuning adaptativo, permitindo que o modelo ajuste suas representações de acordo com os dados dessas línguas, usando transfer learning a partir de corpora em inglês. Apesar de alcançar bons resultados, o desempenho do modelo é fortemente dependente de dados em inglês, o que representa uma barreira quando se trata de generalização para línguas menos exploradas.

Em (Dou et al., 2021), os autores propõem o framework UPFD (User Preference-aware Fake News Detection), que considera as preferências dos usuários, baseadas em históricos de postagens, e o contexto da propagação de notícias nas redes sociais. O framework utiliza uma Rede Neural Gráfica (GNN) para combinar essas informações e prever a credibilidade das notícias. No entanto, o desempenho do UPFD está diretamente relacionado à qualidade dos dados históricos dos usuários, que podem não estar sempre presentes. Além disso, a fusão de múltiplas fontes de informação torna o framework computacionalmente complexo, enfrentando desafios de escalabilidade.

3. FACTUAL - ANÁLISE DE FAKE NEWS COM CONFIANÇA UTILIZANDO IA E LLMS

Esta seção apresenta o FACTUAL, uma solução baseada em LLM para detectar e mitigar a disseminação de fake news. Desenvolvido com base em modelos de linguagem de grande escala, o FACTUAL visa fornecer uma análise confiável de conteúdos potencialmente falsos, permitindo que os usuários tomem decisões informadas e em tempo real. O objetivo geral do FACTUAL é oferecer uma estrutura eficiente e escalável para o treinamento de grandes modelos de linguagem, proporcionando uma solução eficaz para a detecção de fake

news. Especificamente, o FACTUAL busca maximizar o desempenho em tarefas de processamento de linguagem natural, como a geração, tradução e compreensão de texto, ao mesmo tempo que permite a detecção em tempo real e a mitigação da desinformação de forma eficiente.

3.1 Visão Geral do Funcionamento do FACTUAL

A Figura 1 apresenta o funcionalmento da arquitetura do FACTUAL. O FACTUAL é composto por quatro componentes principais: (i) a camada de banco de dados; (ii) os modelos de IA; (iii) o middleware; e (iv) os frameworks de execução. A Camada de Banco de Dados (Rótulo 1, Figura 1) é responsável pelo armazenamento e gerenciamento dos dados utilizados no processo de detecção. Essa camada no FACTUAL utiliza um banco de dados SQLite para armazenar os registros relacionados a fake news que mantém informações comparativas, em que são gravadas as saídas da análise realizada pelo FACTUAL. Já a Camada de Modelos e Frameworks de IA (Rótulo 2, Figura 1) concentra o núcleo da análise do FACTUAL, utilizando o modelo LLaMA 3.2 para processar grandes volumes de dados textuais. O framework Ollama é utilizado para integrar o modelo ao FACTUAL, permitindo a análise e a identificação de padrões característicos de desinformação. O uso de um modelo de linguagem avançado como o LLaMA oferece ao FACTUAL a capacidade de compreender nuances e ambiguidade no conteúdo textual, aumentando a precisão da detecção de fake news em cenários complexos.

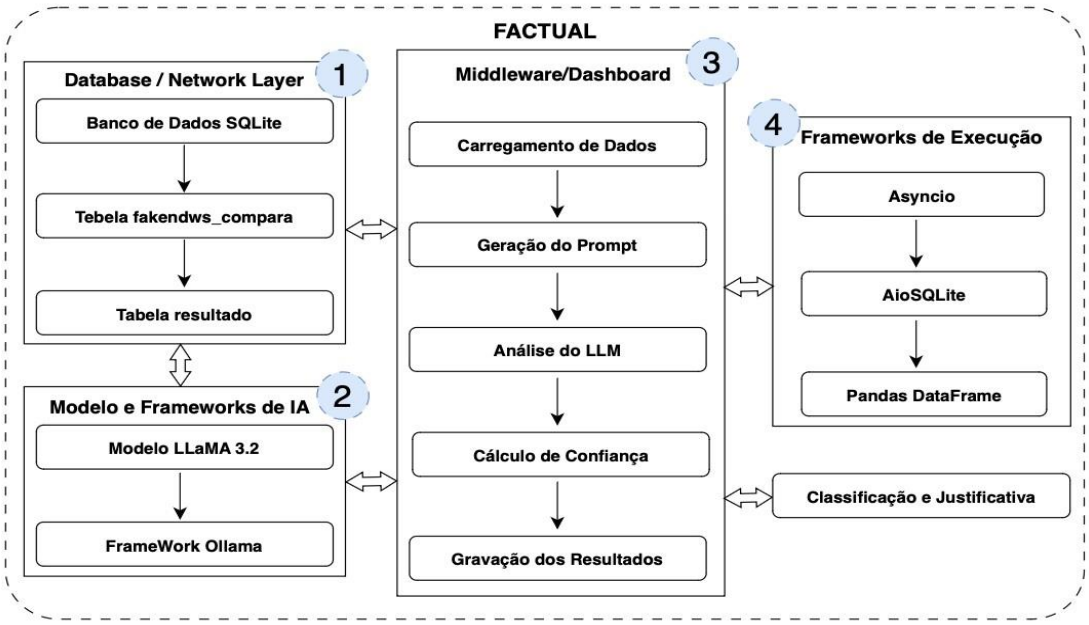


Figura 1. Visão geral do funcionamento do FACTUAL

O Middleware (Rótulo 3, Figura 1) desempenha um papel central na integração entre a camada de dados e os modelos de IA. Esse componente é responsável por tarefas como o carregamento de dados do banco de dados, a geração de prompts apropriados para o modelo de IA, a análise dos resultados obtidos e o cálculo de confiança, o que permite ao sistema avaliar a veracidade das informações analisadas. Além disso, o middleware grava os resultados no banco de dados para posterior consulta, fechando o ciclo de análise do FACTUAL. por fim, a camada de Frameworks de Execução (Rótulo 4, Figura 1) é desenvolvida para otimizar o desempenho do FACTUAL, utilizando frameworks assíncronos como Asyncio e AioSQLite para garantir a execução paralela das tarefas. Esses frameworks permitem que o FACTUAL processe múltiplos pedidos de análise simultaneamente, sem sobrecarregar o sistema, garantindo eficiência e escalabilidade. Os resultados são organizados em Pandas DataFrames, facilitando a visualização e a classificação dos dados, além de fornecer justificativas claras para as conclusões sobre a veracidade das informações analisadas.

3.2 Camada de Banco de Dados no FACTUAL

A camada de banco de dados do FACTUAL é responsável pela estruturação e armazenamento das informações necessárias para o processo de detecção de fake news. O sistema utiliza um banco de dados SQLite, escolhido pela sua leveza e integração com outras camadas do sistema. Dois componentes principais compõem essa camada: (i) a tabela "fakenews_compara"; e (ii) a tabela de resultados. A tabela "fakenews_compara" armazena os dados de entrada, que são informações textuais e outras variáveis coletadas de fontes que potencialmente disseminam fake news. Essa tabela mantém um registro das informações que serão analisadas pelo modelo de linguagem, proporcionando a base para o início do processo de detecção.

Após a análise realizada pelos modelos, os resultados são armazenados na tabela de resultados, que contém as saídas geradas pelo sistema após a detecção de fake news. Nessa tabela, são gravados dados como a classificação das notícias (verdadeiras ou falsas), as justificativas fornecidas pelos modelos e o nível de confiança atribuído a cada classificação. Esses registros são utilizados para a posterior visualização e análise, garantindo que o sistema mantenha um histórico das avaliações e suas justificativas. A escolha por uma base de dados SQLite está atrelada à necessidade de um sistema leve que permita o rápido acesso aos dados e que lide com operações de escrita e leitura simultâneas, sem comprometer o desempenho. Essa solução também facilita a integração com frameworks assíncronos, como o AioSQLite, que permitem o processamento de múltiplas requisições simultâneas, tornando o FACTUAL escalável e apto a processar grandes volumes de dados de maneira eficiente.

3.3 Processo de Classificação de Notícias na Camada de Modelos e Frameworks de IA

A camada de modelos e frameworks de IA é responsável pelo processamento e análise das informações textuais no FACTUAL. O principal componente dessa camada é o modelo de linguagem LLaMA 3.2, que realiza a classificação das notícias enviadas pelo usuário, determinando se são verdadeiras ou falsas. O primeiro passo no processo é a inicialização do FACTUAL, em que o modelo LLaMA 3.2 e o framework Ollama são carregados e conectados. Essa etapa é essencial para preparar o ambiente de execução, garantindo que o sistema esteja pronto para processar as notícias. Uma vez configurado, o FACTUAL pode começar a receber as notícias a serem analisadas. O processo de análise de notícias ocorre de maneira iterativa, passando por cada notícia individualmente. O texto da notícia é enviado ao modelo LLaMA 3.2, por meio do framework Ollama, que processa o conteúdo e gera uma resposta contendo a probabilidade de que a notícia seja falsa. A partir dessa resposta, o sistema calcula o nível de confiança da análise, utilizando as informações fornecidas pelo modelo para gerar uma métrica que reflete a certeza do sistema em sua classificação.

Com base na probabilidade retornada pelo modelo LLaMA 3.2, a notícia é classificada como fake, verídica ou indeterminada. O sistema compara a probabilidade de ser fake com um limiar previamente definido. Se a probabilidade ultrapassar esse valor, a notícia é marcada como fake; caso contrário, ela é classificada como verídica ou indeterminada. Além da classificação, o sistema armazena o resultado da análise, o nível de confiança associado à decisão e gera uma justificativa clara ao usuário. O framework Ollama facilita essa interação, fornecendo a interface para o envio dos prompts ao modelo e a recepção dos resultados. Esses resultados, que incluem a classificação e justificativa, são enfileirados em uma tarefa assíncrona, permitindo que o sistema processe múltiplas requisições de forma eficiente.

3.4 Fluxo de Processamento e Integração do Middleware/Dashboard no FACTUAL

A camada de Middleware/Dashboard do FACTUAL gerencia o fluxo de dados entre as diferentes partes do sistema, facilitando a interação entre o usuário, o modelo de IA e o banco de dados. A principal função dessa camada é carregar os dados, gerar prompts para o modelo LLaMA 3.2, processar os resultados e retornar as respostas ao usuário. O diagrama de sequência apresentado na Figura 2 apresenta o fluxo de processamento que ocorre dentro do FACTUAL. O processo começa com o usuário submetendo uma notícia, que é então carregada da tabela "fakenews_compara" no banco de dados SQLite. As notícias incluem informações como título, texto, URL e editor. Essas informações são organizadas e preparadas para o envio ao modelo LLaMA

3.2 por meio da geração de prompts. O middleware organiza e estrutura essas informações, gerando o prompt que será enviado ao modelo de IA para análise.

Após a geração do prompt, o modelo LLaMA 3.2 realiza a classificação da notícia como verdadeira (fato) ou falsa (fake) ou indeterminado e gera uma justificativa para a classificação realizada. Essa resposta é então enfileirada em uma tarefa assíncrona, utilizando o framework Asyncio. A capacidade de processar múltiplas tarefas de forma assíncrona é fundamental para que o FACTUAL consiga lidar com um fluxo constante de requisições, conforme apresentado na Figura 2.

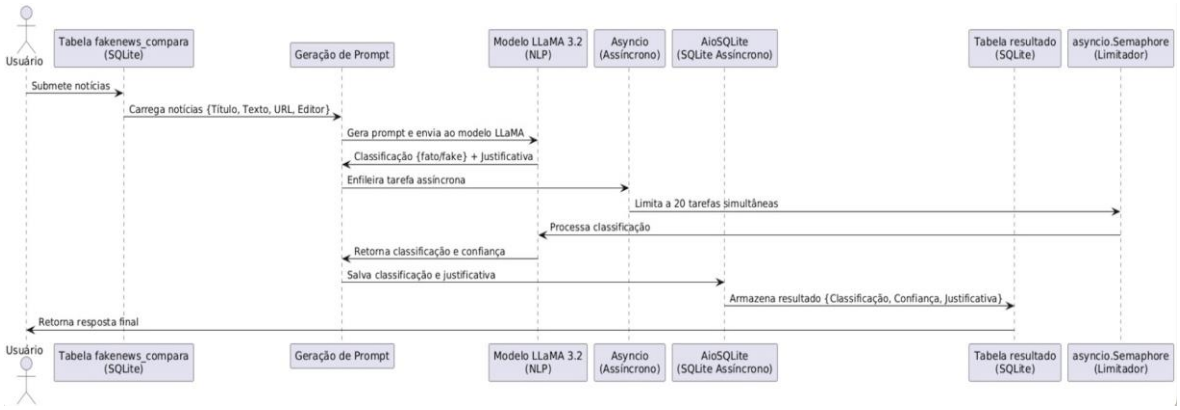


Figura 2. Diagrama de sequência para classificação de notícias

A seguir, o middleware recupera a classificação gerada pelo modelo, juntamente com a confiança, e armazena esses dados na tabela de resultados, retornando a resposta ao usuário. O sistema utiliza AioSQLite, uma extensão assíncrona do SQLite, que habilita operações de I/O não bloqueantes, otimizando latência e throughput em ambientes com alta concorrência. Isso permite que múltiplas requisições de leitura e escrita sejam processadas simultaneamente, maximizando a eficiência ao lidar com grandes volumes de dados. A tabela de resultados armazena a classe prevista, a confiança e uma justificativa gerada pelo modelo. O uso de índices no banco otimiza as consultas, garantindo rápido acesso ao histórico de análises, mesmo em cenários de escalabilidade. O sistema também implementa asyncio.Semaphore, que limita corrotinas concorrentes, prevenindo sobrecargas e controlando o uso de recursos de forma eficiente.

3.5 Execução Assíncrona e Escalabilidade na Camada de Frameworks de Execução do FACTUAL

A Camada de Frameworks de Execução do FACTUAL é responsável por otimizar o processamento e garantir a escalabilidade do sistema, gerenciando as operações de forma assíncrona. Esta camada utiliza frameworks como Asyncio e AioSQLite para processar múltiplas tarefas simultaneamente, sem sobrecarregar o sistema, e lidar com grandes volumes de dados que é uma das características desta pesquia. O Asyncio é um framework de execução assíncrona que permite o processamento de várias requisições em paralelo. No FACTUAL, o Asyncio é utilizado para gerenciar a execução de tarefas simultâneas, conforme apresentado no diagrama da Figura 2. Isso é necessário para garantir que o modelo de IA possa analisar diversas notícias de maneira eficiente, mesmo em momentos de alta demanda.

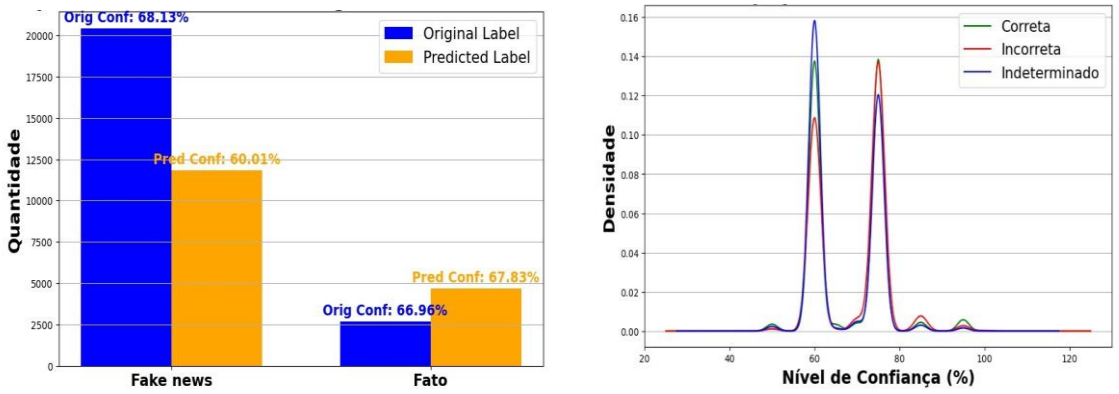
Além de processar múltiplas tarefas em paralelo, a camada de execução utiliza AioSQLite, uma extensão assíncrona do SQLite. O AioSQLite permite ao FACTUAL interagir com o banco de dados de forma eficiente, armazenando os resultados da classificação de fake news em tempo real, sem interrupções no fluxo de trabalho. O semáforo assíncrono controla o número de tarefas executadas simultaneamente. Conforme mostrado na Figura 2, o asyncio.Semaphore limita as tarefas, evitando sobrecarga e garantindo operações dentro da capacidade do servidor, equilibrando desempenho e eficiência no processamento de grandes volumes de notícias.

4. AVALIAÇÃO DE DESEMPENHO

4.1 Impacto dos Resultados Obtidos

Na Figura 3(a), são apresentados os resultados comparativos entre a quantidade de notícias e a média de confiança, analisando as previsões do sistema FACTUAL para os rótulos originais (Original Label) e os rótulos previstos (Predicted Label). Nota-se que, para o rótulo “Fake news”, o sistema apresenta uma média de confiança de 68.13% no rótulo original, enquanto a confiança para as previsões é ligeiramente inferior, com 68.07%. Esse comportamento também se repete para o rótulo “Fato”, em que a confiança média no rótulo original é de 66.96%, enquanto nas previsões é de 67.83%. Essa pequena variação nos níveis de confiança entre os rótulos originais e previstos reflete a consistência do modelo LLaMA 3.2 integrado ao sistema FACTUAL para detectar e classificar fake news e fatos. Embora haja uma redução na quantidade de notícias classificadas como “Fake news” nas previsões, o sistema mantém um nível de confiança próximo ao dos rótulos originais. Tal consistência é essencial para garantir a precisão e a robustez na identificação de desinformação, além de proporcionar maior credibilidade ao sistema.

A Figura 3(b) apresenta o gráfico de densidade que compara os níveis de confiança das previsões corretas, incorretas e indeterminadas no FACTUA. Observa-se uma concentração de previsões em torno de dois picos principais, localizados em aproximadamente 60% e 90% de confiança, isso ratifica o resultado da Figura 3(a). A similaridade entre as curvas de previsões incorretas e indeterminadas sugere uma necessidade de ajuste nos limiares de confiança, destacando a importância de ajustes finos na parametrização para melhorar a separabilidade entre previsões corretas e errôneas no sistema.



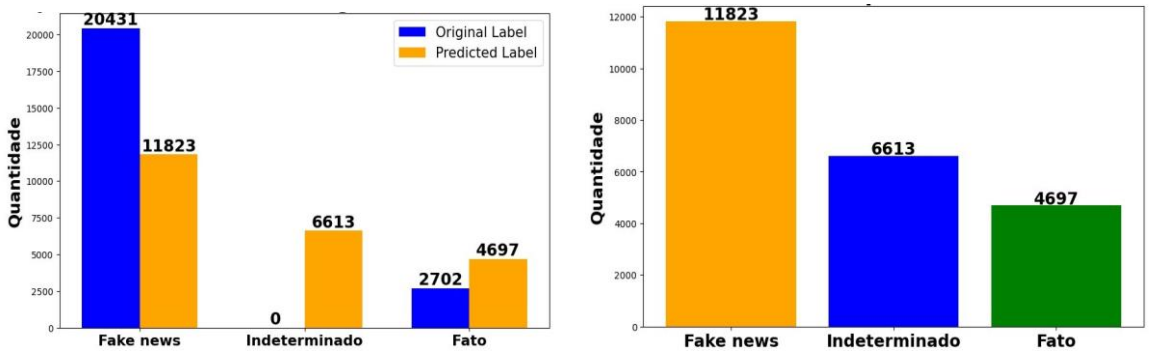
(a) Quantidade de amostras.

(b) Densidade.

Figura 3. Desempenho do FACTUAL para as métricas quantidades de amostras e densidade

Na Figura 4(a), é apresentada uma comparação entre a quantidade de rótulos originais (Original Label) e rótulos previstos (Predicted Label) no sistema FACTUAL para três categorias: “Fake news”, “Indeterminado” e “Fato”. Para o rótulo “Fake news”, o sistema classificou originalmente 20.431 instâncias, enquanto previu 11.823. No caso do rótulo “Fato”, houve uma discrepância, com 2.702 instâncias classificadas originalmente, mas 4.697 previstas. Para a categoria “Indeterminado”, o rótulo original não teve nenhuma instância, mas o sistema previu 6.613, indicando a existência de uma margem de incerteza em suas previsões. Já a Figura 4(b) complementa a análise, mostrando a distribuição dos rótulos previstos. Neste caso, é possível ver que o sistema previu 11.823 instâncias como “Fake news”, 6.613 como “Indeterminado” e 4.697 como “Fato”. Essa distribuição mostra uma predominância de previsões de “Fake news”, seguidas por uma quantidade considerável de previsões indeterminadas. Observa-se que o sistema tende a diminuir a quantidade de instâncias classificadas como “Fake news” nas previsões em comparação com os rótulos originais e aumenta significativamente as previsões na categoria “Indeterminado”, indicando que o sistema pode ter enfrentado dificuldades em classificar de forma assertiva um número considerável de notícias. O aumento nas previsões

de “Fato” em comparação com os rótulos originais também sugere uma possível tendência do sistema em superestimar a veracidade de algumas notícias.



(a) Comparação entre os rótulos.

(b) Distribuição dos rótulos.

Figura 4. Desempenho do FACTUAL para os rótulosFake News, Indeterminado e Fato

5. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou o FACTUAL, um sistema de detecção de fake news que foi modelado por LLM para processar e analisar grandes volumes de dados textuais. O FACTUAL foi projetado com o objetivo de oferecer não apenas um nível de acurácia nas previsões, mas também uma avaliação criteriosa da confiabilidade das fontes de informação. A capacidade de atribuir graus de confiança variados às previsões permite que o FACTUAL identifique padrões de desinformação com maior assertividade, comparando o conteúdo analisado com informações verdadeiras e falsas. Além disso, o sistema foi desenvolvido para ser escalável, adaptando-se a diferentes volumes e contextos de dados, o que o torna uma ferramenta eficaz no combate à disseminação de notícias falsas. Os resultados obtidos demonstram a eficácia do FACTUAL na identificação de fake news e na atribuição de níveis de confiança às fontes, garantindo uma classificação precisa e confiável que pode ser aplicada em diferentes cenários. Para trabalhos futuros, planeja-se expandir o FACTUAL para novos contextos, além de melhorar a calibração dos limiares de confiança, especialmente para reduzir a quantidade de previsões classificadas como “Indeterminado”.

AGRADECIMENTOS

Os autores agradecem o apoio técnico e computacional do Laboratório LATITUDE, da Universidade de Brasília, ao TED 01/2019 da Advocacia Geral da União (Outorga AGU 697.935/2019), ao TED 01/2021 da Secretaria Nacional de Assistência Social – SNAS/DGSUAS/CGRS, ao TED 01/2021 da Coordenação-Geral de Tecnologia da Informação (CGTI) da Procuradoria Geral da Fazenda Nacional – PGFN, ao Projeto SISTER City – Sistemas Inteligentes Seguros e em Tempo Efetivo Real para Cidades Inteligentes (Outorga 625/2022), ao Projeto “Sistema de Controle e Unificação de Projetos para o Governo Distrito Federal – Sispro-DF” (Outorga 497/2023), ao Decanato de Pesquisa e Inovação – DPI/UnB e a FAP/DF.

REFERÊNCIAS

Alnabhan, M. Q. and Branco, P. (2024). Fake news detection using deep learning: A systematic literature review. *IEEE Access*.

Bernardi, A. J. B. (2021). *Fake news e as eleições de 2018 no Brasil: como diminuir a desinformação?* Editora Appris.

Dou, Y., Shu, K., Xia, C., Yu, P. S., and Sun, L. (2021). User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2051–2055.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., and Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

iG São Paulo (2018). Fake news marcaram as eleições de 2018; relembre as 10 mais emblemáticas. Acessado em: 18 out. 2024.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Ma, Z., Luo, M., Guo, H., Zeng, Z., Hao, Y., and Zhao, X. (2024). Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Raja, E., Soni, B., and Borgohain, S. K. (2023). Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Segura-Bedmar, I. and Alonso-Bartolome, S. (2022). Multimodal fake news detection. *Information*, 13(6):284.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Sun, Y., He, J., Cui, L., Lei, S., and Lu, C.-T. (2024). Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Wang, D., Zhang, W., Wu, W., and Guo, X. (2023). Soft-label for multi-domain fake news detection. *IEEE Access*.

Wu, J., Guo, J., and Hooi, B. (2024). Fake news in sheep's clothing: Robust fake news detection against llmempowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378.