UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Sousa dos Anjos

Análise Experimental do Desempenho de Grandes Modelos de Linguagens na Detecção de Notícias Falsas

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Sousa dos Anjos

Análise Experimental do Desempenho de Grandes Modelos de Linguagens na Detecção de Notícias Falsas

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Adriano Mendonça Rocha

Coorientador: Prof. Dr. Silvio Ereno Quincozes

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2023

Lucas Sousa dos Anjos

Análise Experimental do Desempenho de Grandes Modelos de Linguagens na Detecção de Notícias Falsas

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Prof. Dr. Adriano Mendonça Rocha
Orientador

Prof. Dr. Silvio Ereno Quincozes
Coorientador

Prof. Dr. Rodrigo Sanches Miani

Prof. Dr. Murillo G. Carneiro

Uberlândia, Brasil
2023

"É como se todo mundo contasse uma história sobre si mesmo dentro da própria cabeça. Sempre. O tempo todo. Essa história faz o sujeito ser quem é. Nós nos construímos a partir desta história."

(Patrick Rothfuss)

Agradecimentos

Agradeço primeiramente aos meus pais, pela dedicação e educação que me proporcionaram ao longo da vida. Um agradecimento especial para minha mãe, pelo amor, esforço e por sempre valorizar o estudo como um dos pilares mais importantes em nossas vidas.

Às minhas irmãs pelo companheirismo, encorajamento e pela compreensão nas horas que me dediquei intensamente a este trabalho, fizeram toda a diferença.

Aos meus amigos, que estiveram ao meu lado ao longo desse caminho, agradeço pelo apoio e motivação nos momentos mais difíceis.

Agradeço aos professores Adriano e Silvio, pela orientação, dedicação incansável, sabedoria e ensinamentos ao longo dessa trajetória. Estendo meus agradecimentos a todos os professores da Faculdade de Computação da UFU, pelo comprometimento em ensinar ao longo desses anos.

Enfim, agradeço a todos aqueles que, de uma forma ou de outra, contribuíram na realização deste trabalho.

Resumo

A disseminação de notícias falsas tornou-se uma preocupação significativa na sociedade atual. Esse problema é evidente em plataformas de mídia social, onde a propagação de desinformação se tornou uma presença constante na vida diária de muitos indivíduos. Neste trabalho, é investigado o desempenho das ferramentas GPT e Bard na classificação de notícias falsas e reais, considerando 200 artigos de jornal e duas estratégias de formulação de perguntas. Os resultados revelam que o uso de uma pergunta bem formulada é crucial para obter respostas mais precisas. Em particular, foi observado uma melhoria de 46.6% na métrica F1-Score no primeiro modelo do GPT direcionando a pergunta para focar nas características de um texto falso. No segundo modelo o F1-Score registrou uma melhoria de 22.22% para 76.68% quando o foco do *prompt* era nas características de uma notícia falsa. Já no Bard os resultados foram mais timídos, com um F1-Score de 40.45% já na pergunta mais específica. As descobertas apresentadas nesse estudo indicam a superioridade do GPT em relação ao Bard. Ao realizar a detecção de *fake news* em todos os testes a ferramenta da OpenAI foi superior à ferramenta da Google na métrica de F1-Score.

Palavras-chave: ChatGPT, Bard, Inteligência Artifical, Detecção de Fake News, Processamento de Linguagem Natural.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	*Application Programming Interface* (Interface de Programação de Aplicação)
CSV	*Comma-Separated Values*
GPT	Generative Pre-trained Transformer
GRU	*Gated Recurrent Unit*
GWO	*Grey Wolf Optimization*
IA	Inteligência Artificial
LSTM	*Long Short-Term Memory*
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
RNN	Rede Neural Recorrente
SDNF	Sistema de Detecção de Notícias Falsas
SVM	*Support Vector Machine*
SSO	*Salp Swarm Optimization Algorithm*
TF-IDF	*Term Frequency–Inverse Document Frequency*

Sumário

1 Introdução

O avanço acelerado da disseminação de notícias falsas, conhecidas como *fake news* tem se tornado uma preocupação cada vez mais presente no dia a dia da sociedade. Essa preocupação atinge principalmente as redes sociais, onde conteúdos falsos se espalham em velocidades alarmantes. De acordo com o Tribunal Superior Eleitoral (2022), as *fake news* circularam 70% mais rápido que as verdadeiras no ano de 2022. Tais conteúdos, têm o potencial de causar danos sérios à sociedade e à saúde pública, como desencadear medo e estresse no indivíduo afetado (ROCHA et al., 2021). Isso se deve ao fato de que essas notícias atuam como narrativas que omitem ou adicionam informações aos fatos. Tais fatos, por sua vez, muitas das vezes são baseadas em teorias conspiratórias sem fundamento. Um exemplo é a falsa teoria de que a China testou o coronavírus como arma biológica (Redação Estadão, 2021). Outro exemplo é a falsa notícia de que a vacina contra a COVID-19 irá modificar o DNA humano (PINHEIRO, 2021).

Para contribuir no combate à disseminação de *fake news*, é importante entender esse problema e desenvolver ferramentas precisas para a detecção desse tipo de conteúdo. Nesse contexto, o presente trabalho visa estudar e entender o funcionamento do GPT—um modelo de linguagem treinado pela OpenAI que se tornou recentemente popular—e do Bard—um modelo semelhante desenvolvido pela Google—e avaliar o potencial de ambos os modelos para detecção de notícias falsas.

A importância desse trabalho reside na necessidade de reduzir os efeitos negativos das *fake news* na sociedade, seja no contexto da saúde pública, política, economia ou seja qual for o contexto. A criação de ferramentas eficientes para detecção de *fake news* pode contribuir para o desenvolvimento de um ambiente de informação mais saudável, levando à diminuição do impacto causado por essas notícias e possibilitando o aumento da confiabilidade da sociedade nos meios de comunicação. Isso permite que os usuários tomem decisões baseadas em fontes confiáveis e evitem serem influenciados por informações enganosas. A detecção de *fake news* é essencial para promover a confiabilidade dos meios de informação e para fortalecer a capacidade da sociedade em distinguir entre o que é verdadeiro e o que é falso.

1.1 Objetivos

O objetivo geral deste trabalho consiste em explorar a aplicação das ferramentas de Inteligência Artificial (IA) generativas conhecidas como ChatGPT e Google Bard, para enfrentar o desafio da detecção de notícias falsas. O principal objetivo é avaliar a viabilidade de utilizar os modelos IA providos por essas ferramentas como componentes centrais de um Sistema de Detecção de Notícias Falsas (SDNF). Uma aplicação de modelos avançados de linguagem, como o ChatGPT e o Google Bard, podem fornecer uma nova perspectiva sobre a detecção de notícias falsas e contribuir para mitigar esse problema.

Os objetivos específicos são listados a seguir:

- Implementar uma prova de conceito da arquitetura proposta por meio do modelo de linguagem de inteligência artificial GPT-3.5;
- Repetir o mesmo processo realizado com o GPT-3.5 com o Google Bard;
- Realizar uma análise comparativa entre ambos os modelos para determinar qual o mais eficiente;
- Investigar o impacto de diferentes *prompts* —texto enviado para o modelo de IA, nesse caso, solicitando a análise da notícia— na classificação de texto fornecida pelas ferramentas;
- Documentar todo o processo de desenvolvimento para contribuir com o avanço do conhecimento na área proposta, publicando-os através de artigos científicos.

1.2 Justificativa

Segundo levantamento feito pelo Poynter Institute, no Brasil, 4 em cada 10 pessoas afirmam receber *fake news* todos os dias (Pedro Guimarães; Cleber Rodrigues, 2022). Esse número cresce para 65% entre os brasileiros que se preocupam em confiar nesse tipo de conteúdo ou que seus parentes confiem. A quantidade de informações disponíveis atualmente dificulta o trabalho de identificar, avaliar e rotular esse tipo de conteúdo. Levando em consideração essas informações é de extrema importância que a sociedade tenha acesso à ferramentas que ajudem a identificar notícias enganosas.

Com a recente explosão de popularidade das ferramentas de Inteligência Artificial, principalmente as ferramentas focadas em texto, como o ChatGPT e o Google Bard, surge a importância de avaliar a capacidade dessas ferramentas em detectar *fake news*. Com isso, determinando se tais ferramentas são confiáveis ou não na detecção desse tipo de conteúdo.

Apesar de já existir algumas ferramentas de detecção de fake news, como o *Fake news detector* (DEVELOPERS; FAVA, 2017), *Detektor dezinformačných webov* (DESIGN, 2020) e o *Fake News Detector* (DEREK, 2018), a maioria não suporta múltiplas línguas. As ferramentas *FakeNewsBR* (HENRIQUES; NETO, 2022) e *FakeCheck* (MONTEIRO, 2018) aceitam texto em português mas não contemplam um alto nível de usabilidade, ou seja, não são intuitivas. No primeiro, é necessário acessar o site da ferramenta e adicionar o link ou texto da notícia para receber a análise, porém, em textos muito pequenos o resultado pode não ser muito preciso, como destacado no próprio site e, no segundo, é necessário que todo o texto seja copiado e colado na ferramenta para ser realizada a análise.

Existem também aplicações móveis, tais como o *Fake News Detector* (LazerLikeFocus, 2020), *Oigetit Fake News Filter* (Oigetit, Inc., 2019-2022) e o *Fake news agregador* (SANTOS, 2020-2022), com o propósito de detectar *fake news*. No entanto, tais aplicações são limitadas, pois somente suportam textos na língua inglesa ou funcionam como um agregador dos principais sites especialistas em *fake news* do Brasil, ou seja, caso a notícia desejada pelo usuário não esteja disponível nesta listagem o usuário acaba ficando sem resposta.

Para auxiliar na resolução desse problema, este trabalho tem como objetivo realizar uma análise de desempenho das ferramentas ChatGPT e Google Bard na detecção de notícias falsas. Utilizar tais ferramentas têm a vantagem da não necessidade de treinamento, já que as desenvolvedoras por trás são responsáveis por esta tarefa. Em particular, conforme detalhado na Seção 1.1, será proposta uma metodologia para análise de desempenho de ambas as ferramentas. Ao avaliar esses modelos, espera-se contribuir com o avanço das pesquisas em detecção de *fake news* e fornecer informações que podem auxiliar no desenvolvimento e aprimoramento de ferramentas disponibilizadas ao público geral.

2 Fundamentação Teórica

Este capítulo tem como objetivo apresentar alguns dos conceitos básicos que são essenciais para o entendimento do presente trabalho. Além disso, também serão explorados os trabalhos da literatura que, de alguma forma, foram relevantes para o desenvolvimento.

2.1 Conceitos Básicos

Com o intuito de obter uma compreensão mais aprofundada deste trabalho, é fundamental a compreensão de alguns conceitos, como Aprendizado de Máquina (*Machine Learning*), *Generative Pre-trained Transformer* (GPT), entre outros. Além disso, é importante estar familiarizado com o conceito de *Fake News*.

2.1.1 Aprendizado de Máquina

Segundo Naqa e Murphy (2015), o aprendizado de máquina é um ramo em constante evolução de algoritmos computacionais projetados para emular a inteligência humana, aprimorando gradativamente sua precisão. Os autores também destacam que técnicas baseadas em aprendizado de máquina têm sido aplicadas com sucesso em diversos campos, que vão desde o reconhecimento de padrões, visão computacional, engenharia espacial, finanças, entretenimento e biologia computacional até aplicações biomédicas e médicas. Os avanços tecnológicos no armazenamento e poder de processamento conseguidos nos últimos anos têm viabilizado o desenvolvimento de produtos baseados em aprendizagem de máquina, como o mecanismo de recomendação da Netflix e carros autônomos.

Através do uso de métodos estatísticos, os algoritmos são treinados para realizar classificações, previsões e apresentar resultados relevantes em projetos de mineração de dados. Esses resultados podem impulsionar a tomada de decisões em aplicativos e empresas, impactando positivamente as métricas essenciais de crescimento (IBM, 2021a).

2.1.2 Processamento de Linguagem Natural

Feldman (1999) faz uma analogia interessante entre Processamento de Linguagem Natural e o aprendizado de crianças, onde aprendem a linguagem descobrindo padrões e modelos, em como expressar o plural ou o singular, como combinar essas formas em verbos e substantivos e como montar uma frase, uma pergunta ou um comando. Segundo a autora, O Processamento de Linguagem Natural (NLP) parte do pressuposto de que,

se pudermos definir esses padrões e descrevê-los para um computador, então podemos ensinar a uma máquina algo sobre como falamos e nos entendemos.

Por outro lado, a definição trazida pela IBM diz que NLP refere-se ao ramo da IA, que tem como objetivo capacitar os computadores a compreender o texto e a fala de maneira similar aos seres humanos. Para alcançar esse objetivo, o NLP combina a linguística computacional, que se baseia em regras para modelar a linguagem humana, com modelos estatísticos e de aprendizado de máquina. Essas tecnologias combinadas permitem que os computadores processem dados de texto ou voz em linguagem humana e compreendam o seu significado completo, incluindo a intenção e o sentimento do falante ou escritor.

É possível encontrar diversos exemplos de aplicações práticas do NLP, tais como sistemas de GPS ativados por voz, assistentes digitais, *software* de conversão de fala em texto, *chatbots* de atendimento ao cliente, entre outros. Essas tecnologias trazem conveniência para os consumidores e demonstram o potencial do NLP na interação entre humanos e computadores (IBM, 2021b).

2.1.3 Arquitetura Transformer

Em 2017, a arquitetura Transformer foi introduzida como um modelo de aprendizado de máquina no artigo *Attention Is All You Need* (VASWANI et al., 2017), até então, os modelos de linguagem eram baseados em Redes Neurais Recorrentes (RNN), *Long Short-Term Memory* (LSTM) e *Gated Recurrent Units* (GRUs), que são um aprimoramento das RNNs comuns. Apesar das melhorias geradas pelo LSTM, esse modelo têm dificuldades em compreender sequências longas de texto, devido ao fato de que todo o histórico conhecido como contexto é manipulado por um único vetor de estado. A arquitetura Transformer superou essa limitação por meio do mecanismo de atenção (*Attention*), além disso ela proporcionou um maior rendimento, pois as entradas são processadas em paralelo, sem a dependência sequencial. O GPT é um exemplo de ferramenta baseada nessa arquitetura (SINGH; MAHMOOD, 2021).

2.1.4 GPT

O GPT (*Generative Pre-trained Transformer*) é um modelo de linguagem baseado na arquitetura Transformer que utiliza técnicas de pré-treinamento para aprender a estrutura e a semântica da linguagem. Desenvolvido pela OpenAI, ele é reconhecido por sua habilidade em gerar texto coerente e natural. O GPT tem sido amplamente empregado em tarefas de processamento de linguagem natural, como geração de texto, tradução automática e resposta a perguntas, entre outras.

A arquitetura de transformadores utilizada pelos modelos GPT representa um

avanço significativo na pesquisa de Inteligência Artificial. A ascensão dos modelos GPT marca um marco na adoção generalizada de Aprendizado de Máquina, uma vez que essa tecnologia pode agora automatizar e aprimorar diversas tarefas. Desde tradução de idiomas e resumos de documentos até redação de postagens em blogs, criação de *sites*, desenvolvimento de elementos visuais e animações, escrita de código, pesquisa em tópicos complexos e até mesmo composição de poemas (Amazon AWS, 2023).

2.1.5 Modelos GPT

De acordo com a OpenAI (2023a), os modelos são uma família de modelos de linguagem projetados para realizar tarefas de processamento de linguagem natural, como tradução de idiomas, geração de texto e chat. A seguir, na Tabela 1, exemplos de alguns dos modelos disponíeis do GPT-3.5.

Tabela 1 – Modelos GPT-3.5

Modelo	**Descrição**
text-davinci-003	Pode realizar qualquer tarefa de linguagem com melhor qualidade, saída mais longa e seguir instruções de forma mais consistente do que os modelos curie, babbage ou ada. Também oferece suporte a recursos adicionais, como inserção de texto.
text-davinci-002	Capacidades semelhantes às do text-davinci-003, mas treinado com ajuste fino supervisionado em vez de aprendizado por reforço
gpt-3.5-turbo	Modelo GPT-3.5 mais capaz e otimizado para chat a 1/10 do custo do text-davinci-003.
gpt-3.5-turbo-16k	Mesmas capacidades do modelo gpt-3.5-turbo padrão, mas com 4 vezes o contexto.
gpt-3.5-turbo-instruct	Capacidades semelhantes às do text-davinci-003, mas compatível com o endpoint de Completions legado e não com Chat Completions.

Lembrando que a Tabela 1 não exibe todos os modelos disponíveis do GPT-3.5, apenas o que são mais relevantes para este trabalho.

2.1.6 Bard

Bard é um *chatbot* de inteligência artificial baseado em texto desenvolvido pela Google. Ele gera respostas em tempo real usando NLP (2.1.2) (AYDIN, 2023b). Segundo AYDIN (2023a), o Bard é baseado no *Pathways Language Model 2* (PaLM 2), um modelo de linguagem lançado no final de 2022. O PaLM e seu antecessor, a tecnologia *Language Model for Dialogue Applications* (LaMDA) do Google, são um modelo de linguagem lançado pelo Google em 2017. A arquitetura da rede neural é baseada no Transformer. De

acordo com a Google, o Bard pode auxiliar em tarefas criativas, explicar assuntos complexos e extrair informações de diversas fontes da internet, bem como traduzir idiomas, fazer resumos de documentos, gerar texto e códigos

2.1.7 *Fake News*

A definição clara de *fake news* é fundamental para combater esse fenômeno, pois compreender sua natureza é essencial para desenvolver estratégias eficazes de combate. Na literatura, diversas definições de *fake news* são apresentadas, porém, é importante buscar uma definição clara e precisa, pois qualquer ambiguidade pode criar brechas perigosas para a prática de censura ou perseguição política, por exemplo (ALVES; MACIEL, 2020).

Nesse sentido, o presente trabalho adota a definição de *fake news* proposta por Alves e Maciel (2020): "informações divulgadas com o objetivo de persuadir e fortalecer uma posição específica dentro de uma disputa narrativa em um ambiente altamente polarizado". Os autores também ressaltam que a disseminação de *fake news* não necessariamente envolve a intenção de enganar ou manipular o receptor de forma fraudulenta. Em vez disso, ela é concebida como uma estratégia para influenciar as pessoas e reforçar uma perspectiva particular em meio a debates acirrados e divisões sociais.

Essa definição nos permite compreender a complexidade do fenômeno das *fake news*, e assim, é possível elaborar estratégias efetivas para combatê-las.

2.2 Trabalhos Relacionados

Na literatura, vários autores abordam a criação de sistemas de detecção de *fake news*, onde apresentam abordagens de desenvolvimento e resultados diferentes. O artigo *Fake News Detection System using Web-Extension* (KHIVASARA; KHARE; BHADANE, 2020) tem como proposta uma extensão web que capacita os usuários a diferenciar entre conteúdos de notícias falsas e verdadeiras. Tal sistema utiliza de dois modelos de aprendizado profundo: um deles é um modelo LSTM que foi treinado especificamente para a detecção de notícias falsas, enquanto o outro utiliza o GPT-2 com o objetivo de classificar se um texto foi ou não gerado por IA. É importante destacar que o sistema apresentou uma acurácia de 98,6%, o que destaca um alto desempenho na detecção de *fake news*. Com uma taxa de acertos tão alta, o sistema proposto demonstra uma habilidade significativa na detecção de notícias. Entretanto, também é importante observar que o sistema se baseia na utilização do GPT-2 apenas para determinar se o texto foi ou não gerado por IA, além de utilizar uma versão mais antiga do modelo, considerando que já existem versões mais recentes da tecnologia. Levando em conta a relevância e eficácia do GPT, seria interessante explorar também sua aplicação na detecção de notícias falsas, aprimorando ainda mais o sistema.

O estudo apresentado por Raza e Ding (2022): *Fake news detection based on news content and social contexts: a transformer-based approach*, propõe a criação de um *framework*, baseado na arquitetura *Transformer*, no qual as informações das notícias e os contextos sociais são explorados para realizar a detecção. É interessante notar que, essa pesquisa, propõe a detecção de notícias falsas antes que ela seja amplamente propagada nas redes sociais, e também conscientizar os usuários sobre o tipo de conteúdo que estão consumindo. No entanto, no próprio artigo, o autor reconhece algumas limitações no modelo. O *framework*, por exemplo, não é multilíngue, isso restringe sua eficácia em países onde a língua inglesa não é amplamente falada e o *framework* também pode apresentar vieses devido à possibilidade na mudança de estratégias por parte dos criadores de notícias falsas ao longo do tempo, essa mudança pode dificultar a detecção, uma vez que o modelo foi treinado com uma base de dados em um determinado momento no tempo.

Os autores Baarir e Djeffal (2021), propuseram em *Fake News detection Using Machine Learning* a criação de um sistema de detecção de *fake news* que utiliza de técnicas de *Machine Learning* para a classificação do texto. Neste estudo, foram utilizadas as técnicas TF-IDF (*Term Frequency–Inverse Document Frequency*) e SVM (*Support Vector Machine*). O TF-IDF foi utilizado para extrair as características relevantes dos textos, tais como autor, data e emoção transmitida. Como observado pelos autores, tais características desempenham um papel fundamental ao ensinar o modelo a reconhecer os padrões e características que distinguem um texto falso de um verdadeiro. O SVM, por sua vez, é um algoritmo de aprendizado de máquina que foi empregado para classificar as entradas em categorias predefinidas, ou notícia falsa ou notícia verdadeira. Os autores também destacam que, caso a acurácia do modelo não atingisse um nível aceitável, os parâmetros do algoritmo de aprendizado eram revisados visando melhorar essa métrica. Apesar disso, também foi pontuado que seria benéfico utilizar uma carga de dados maior para os testes, a fim de obter resultados mais confiáveis. Também é válido mencionar que existem ferramentas eficientes disponíveis para auxiliar na detecção, como é o caso do GPT.

Em outro estudo, realizado por Özbay e Alatas (2019), é apresentada uma abordagem diferente das demais para o problema da detecção de notícias falsas, foi utilizado dois algoritmos meta-heurísticos, o GWO (*Grey Wolf Optimization*) e SSO (*Salp Swarm Optimization Algorithm*) para auxiliar na resolução do problema. Tal estudo utilizou três conjuntos diferentes de dados. Os resultados apontaram que o algoritmo GWO teve um desempenho superior ao SSO, onde foi obtida uma acurácia melhor em todos os conjuntos de dados utilizados, a precisão e *F1-Score* foram mais elevados em dois dos três conjuntos de dados. Como proposto pelos autores, seria interessante desenvolver versões adaptativas e híbridas dos algoritmos SSO e GWO, com o intuito de obter resultados melhores.

Por sua vez, o estudo apresentado por Aslam et al. (2021) visa acabar com a disseminação de rumores e notícias falsas e ajudar a população a identificar a fonte de notícias

como confiável ou não, classificando tais notícias automaticamente através de técnicas de aprendizado profundo, técnicas de NLP como tokenização, lematização e remoção de palavras irrelevantes. Os testes realizados com o modelo proposto, apresentaram resultados consistentes, com acurácia, precisão e recall sempre acima de 80%. Apesar de terem obtido um desempenho satisfatório, os autores ressaltaram a importância de aplicar o modelo em mais conjuntos de dados, inclusive mais recentes, tendo em vista que a detecção de notícias falsas é um campo novo que está em constante desenvolvimento.

O trabalho realizado por Jain e Kasbe (2018) apresenta um método beseado em dos dos algoritmos de aprendizado de máquina Naïve Bayes com o objetivo de apoiar ou não a ideia de utilizar IA para detecção de *fake news*. Segundo os autores, os classificadores Naïve Bayes são uma família de classificadores probabilísticos simples baseados na aplicação do teorema de Bayes. Eles podem prever as probabilidades de relacionamento para cada classe, como a probabilidade de que um registro ou ponto de dados dado pertença a uma classe específica. O estudo utilizou uma base de dados contendo 11.000 artigos de notícias etiquetadas como reais ou falsos. Após a rodada de testes, os autores calcularam a Precisão e a Revocação do modelo e combinaram esses valores em uma única métrica, plotando um gráfico com os valores de Revocação no eixo das abscissas (eixo horizontal) e os valores de Precisão no eixo das ordenadas (eixo vertical). Com isso tiveram uma curva denominada de curva ROC, e consideraram a métrica AUC (Área sob a curva), sendo assim, quanto mais próximo o valor de AUC estiver de 1, mais eficaz o modelo é em prever se uma notícia é verdadeira ou não. Como resultado, os autores tiveram a melhor pontuação de 0.931. Apesar dos bons resultados, os autores destacaram a importância do uso de mais dados para fins de treinamento e, além disso, é válido pontuar que é importante o estudo de ferramentas já disponíveis, como o Bard e GPT, para avaliação de desempenho para esse tipo de tarefa.

No estudo *Automatic Fake News Detection with Pre-trained Transformer Models* (SCHÜTZ et al., 2021) é proposta uma metodologia para detecção de *fake news* baseada nos seguintes modelos *Transformer*:

- BERT;
- RoBERTa;
- ALBERT;
- DistilBERT, e;
- XLNet.

Para conduzir o estudo, os autores utilizaram uma base de dados com 21.041 artigos, sendo 5.053 artigos falsos e 15.998 reais. Após o pré-processamento dos textos, o

conjunto de dados foi divido em 80% para treinamento e os 20% restantes para testes. Como resultado, os experimentos apontaram que o RoBERTa atingiu os melhores resultados com uma taxa de precisão de 0.87. Apesar da superioridade do RoBERTa o estudo apontou que todos os modelos têm uma boa capacidade de previsão. Os autores também apontaram que diferentes etapas de pré-processamento não têm um impacto significativo na previsão dos modelos. A partir desses resultados, o estudo aponta que uma abordagem baseada em modelos *Transformer* é uma promissora linha de detecção de notícias falsas, quase todos os experimentos apontados obtiveram uma precisão acima de 80%. Apesar dos bons resultados, os autores apontam que é importante aprofundar a pesquisa nessa área, principalmente no quesito explicabilidade para ajudar a entender as diferenças nos conceitos de notícias falsas e obter *insights* sobre os modelos e quais palavras têm o maior impacto na previsão.

Os autores Samadi, Mousavian e Momtazi (2021) propuseram uma metodologia pra responder a seguinte questão: "para uma tarefa complexa como a detecção de notícias falsas, qual combinação de modelos pré-treinados e classificadores neurais pode ter um desempenho preciso?". Para isso, criaram um *framework* para detecção de *fake news* com base em representação contextualizada de texto e classificação neural profunda. Além disso, compararam o desempenho de diferentes combinações de modelos pré-treinados e classificadores neurais conectando três classificadores diferentes a modelos de representação contextualizada de texto, como o BERT, RoBERTa, GPT-2 e Funnel Transformer. Os pesquisadores utilizaram um conjunto de dados de notícias falsas em inglês, sendo a maioria deles coletada de redes sociais e o restante consistem em artigos de notícias. Para comparar o desempenho dos modelos, utilizaram a métrica de avaliação Acurácia (definida em 3.1). Com o melhor resultado de todo o estudo, o Funnel Transformer obteve 0.9738 de acurácia, já o GPT obteve seu melhor resultado com uma acurácia no valor de 0.9673. Os autores apontam que novas análises textuais no processamento de linguagem natural, como o reconhecimento de entidades nomeadas, podem ser úteis para detectar notícias falsas, e deve ser aplicado em pesquisas futuras. É válido salientar que, nessa pesquisa, os usuários não usaram os modelos de linguagem pré-treinados (GPT, RoBERTa, etc) como o meio principal de detecção e sim combinados com outros classificadores neurais.

Uma das motivações do trabalho *"A Study of Algorithm-Based Detection of Fake News in Brazilian Election: Is BERT the Best?"* (MOREIRA et al., 2023) se dá pela escassez de estudos dedicados à análise do desempenho do BERT na detecção de notícias falsas em português. Nesse sentido, os autores propõem responder a seguinte questão: "o BERTimbau apresenta maior precisão quando comparado a algoritmos tradicionais de aprendizado de máquina?". No estudo, foi utilizado um dataset em português com um total de 7200 notícias, 50% falsas e 50% verdadeiras. Os textos do dataset foram pré-processados e submetidos a uma etapa de transformação de dados para serem interpretados pelo classificador. Como citado no texto, além de avaliar a eficácia do BERTimbau na detecção de

notícias falsas em português, foi realizada uma comparação de desempenho com o modelo BERT original usando uma versão traduzida do conjunto de dados em inglês. Para testar o desempenho, os autores utilizaram uma técnica de validação cruzada. Segundo eles, essa técnica divide o conjunto em 10 partes iguais e altera essas subdivisões entre os conjuntos de treinamento e teste. Como resultado, o estudo apontou que a maioria dos algoritmos, com exceção do Naive Bayes (83,21%), obtiveram uma precisão média superior a 93%. O estudo aponta que o desempenho do BERTimbau superou os algoritmos tradicionais de aprendizado de máquina em todas as métricas avaliadas. O desempenho que o modelo obteve nas métricas de precisão, revocação e *F1-Score* foi de 98,7%, 98,19% e 98,74%, respectivamente. Apesar dos bons resultados, os autores destacam que tais resultados devem ser considerados com cautela, levando em conta que o BERTimbau possui limitações, como o alto custo computacional e o seu potencial para viés, por exemplo.

Na Tabela 2, é apresentada uma comparação entre os artigos citados e o presente estudo, destacando as principais características que serão abordadas neste trabalho.

Tabela 2 – Comparação de Trabalhos

Referência	**Escopo**	**Algoritmos**	
		GPT (versão)	**Bard (versão)**
(KHIVASARA; KHARE; BHADANE, 2020)	Fake News	✓(2.0)	✗
(RAZA; DING, 2022)	Fake News	✗	✗
(BAARIR; DJEFFAL, 2021)	Fake News	✗	✗
(ÖZBAY; ALATAS, 2019)	Fake News	✗	✗
(ASLAM et al., 2021)	Fake News	✗	✗
(JAIN; KASBE, 2018)	Fake News	✗	✗
(SCHÜTZ et al., 2021)	Fake News	✗	✗
(SAMADI; MOUSAVIAN; MOMTAZI, 2021)	Fake News	✓(2.0)	✗
(MOREIRA et al., 2023)	Fake News	✗	✗
Este Trabalho	Fake News	✓(3.5)	✓(PaLM 2)

3 Desenvolvimento

De modo a implementar a análise proposta, será utilizada a API da OpenAI, o GPT para fazer a análise do texto e classificá-lo como *fake news* ou não. Para isso é necessário o estudo e entendimento da documentação da API do GPT. Para obter os resultados do Google Bard, foi necessária a utilização de uma biblioteca desenvolvida por terceiros (Daniel Park, 2023) pois, no momento em que este trabalho foi realizado, a Google ainda não oferecia acesso amplo à API do Bard, sendo necessário solicitar participar de uma lista de espera no *MakerSuite*—uma ferramenta para criação de aplicações usando os modelos de linguagem generativos da Google—para receber acesso à API.

O passo-a-passo da criação de resultados para a análise proposta pode ser visualizado na Figura 1 e detalhado com mais clareza a seguir.

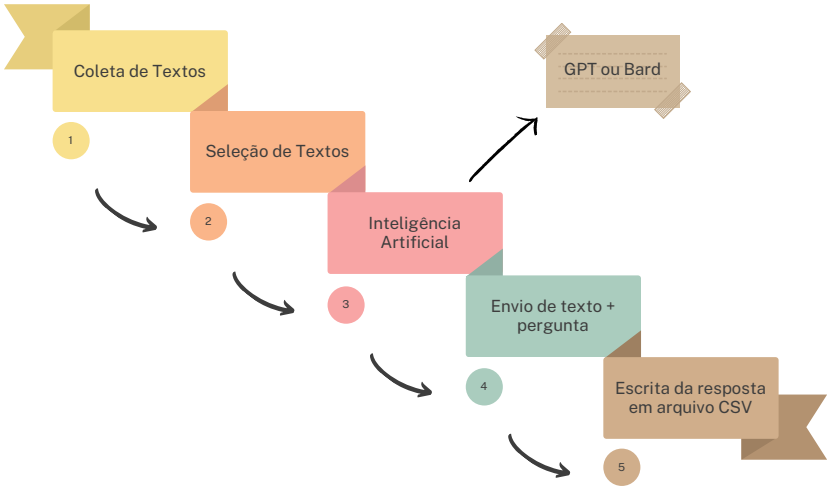


Figura 1 – Fluxograma de funcionamento.

Para realização do processo apresentado na Figura 1 foi desenvolvido um script *Python*, com o auxílio da biblioteca OpenAI, que pode ser aplicada a praticamente qualquer tarefa que exija compreensão ou geração de linguagem e código natural (OpenAI, s.d.). Para gerar resultados para o Google Bard, foi utilizada a biblioteca desenvolvida por Daniel Park (2023), que através do valor de *cookie* do navegador gera respostas assim como a ferramenta de *chat* disponibilizada pela própria Google (Google, 2023).

Para abordar as questões mencionadas anteriormente, foi adotado um processo

composto por três etapas concebidas para atingir os objetivos de classificar notícias como informações falsas ou verdadeiras:

3.1 Seleção e Preparação de Dados

Foi selecionado um conjunto de dados contendo notícias falsas em inglês, denominado *ISOT Fake News Dataset* (BOZKUş, 2023). O Conjunto de dados inclui dois tipos de artigos: notícias genuínas e notícias falsas, coletadas de fontes do mundo real. Os artigos de notícias falsas foram obtidos de sites não confiáveis que foram identificados pelo Politifact, uma organização de verificação de fatos nos Estados Unidos. Este conjunto de dados abrange 21.417 artigos autênticos e mais de 23.481 artigos falsos, a maioria dos artigos dos anos de 2016 e 2017. Conforme a descrição do conjunto de dados, os dados passaram por um processo de limpeza e pré-processamento, embora a pontuação e erros gramaticais nas notícias falsas tenham sido retidos no texto. Neste estudo foi selecionado um conjuto de 200 notícias, automatizados por um script Python, onde as 100 primeiras notícias verdadeiras e as 100 primeiras notícias falsas do dataset foram salvas em um arquivo csv. O dataset contém:

- título da notícia;
- conteúdo da notícia;
- assunto, e;
- data da notícia.

Vale ressaltar que o uso desses dados foi exclusivamente para fins de teste, uma vez que as ferramentas utilizadas são responsáveis pelo treinamento contínuo e validação de seus modelos de linguagem.

3.2 Comunicação com o ChatGPT

Para incorporar a comunicação com o Chat-GPT no código Python, os passos necessários foram os seguintes:

1. Criação de uma conta na plataforma OpenAI: visitando o site oficial e posteriormente fazendo login;
2. Uma vez logado, a seção de Chaves de API foi acessada. Dentro desta seção, uma nova chave de API foi gerada e copiada;

3. No código Python, a biblioteca OpenAI foi importada para habilitar suas funcionalidades. Por fim, a chave de API foi definida usando o método openai.api key, conforme detalhado no Algoritmo 1, linha 2.

Esses passos permitiram a integração do GPT no código Python.

3.3 Comunicação com o Google Bard

Para incorporar a comunicação com o Google Bard no código Python, os passos necessários foram os seguintes:

1. Criação de uma conta Google: visitando o site oficial e posteriormente fazendo login;
2. Uma vez logado, é necessário entrar no site oficial do Google Bard em um navegador e pressionar a tecla F12 para abrir as Ferramentas de Desenvolvedor;
3. Acessar a aba *Application* → *Cookies* e, depois copiar o valor *___Secure-1PSID*. Esse valor funciona como uma espécie de API KEY onde é permitido acesso às respostas do Bard.
4. No código Python, a biblioteca Bard-API foi importada para habilitar suas funcionalidades. Por fim, o valor copiado do *cookie* foi definido no método *get_answer* da biblioteca Bard-API, conforme detalhado no Algoritmo 2, linha 9.

Esses passos permitiram a integração do Bard no código Python.

3.4 Classificação de Texto

3.4.1 GPT

Um *script* Python foi utilizado para realizar a classificação dos textos selecionados. Esse *script* utilizou a API GPT fornecida pela OpenAI, permitindo que o modelo GPT-3.5 classificasse os textos. A OpenAI desenvolveu uma solução para melhorar a legibilidade ao processar a linguagem natural, levando em consideração o *feedback* humano. Essa solução é chamada InstructGPT. Com base no InstructGPT, eles criaram o modelo text-davinci-002, que é treinado com ajuste fino supervisionado. Por fim, a OpenAI aprimorou esse modelo substituindo essa abordagem por aprendizado por reforço. O modelo aprimorado foi chamado text-davinci-003. Como resultado, este último pode lidar com qualquer tarefa de linguagem com melhor qualidade, saída mais longa e seguimento consistente de instruções em comparação com os modelos curie, babbage ou ada (outros modelos disponíveis para uso) (OpenAI, 2023). Como o ChatGPT-3.5 não possui uma API dedicada

especificamente projetada para tarefas de classificação de texto, esta metodologia envolve a transmissão de dois elementos distintos:

1. o texto que requer classificação e;
2. instruções explícitas formuladas em uma pergunta cuidadosamente elaborada (*prompt*).

Ao fazer isso, o GPT-3.5 gera uma resposta pertinente para a classificação de texto, com base em sua base de conhecimento preexistente. Como resultado, um arquivo CSV é gerado, que inclui as classificações reais do conjunto de dados original, juntamente com as classificações atribuídas pelo GPT. Os passos desse script estão detalhados no pseudocódigo apresentado no Algoritmo 1.

3.4.2 Bard

A classificação de texto no Bard funciona de maneira semelhante à no GPT, porém, com algumas desavantagens. Utilizando a API do GPT, é possível definir o grau de criatividade das respostas e escolher diferentes tipos de modelos. Essas desvantagens se devem ao fato de, até o momento da escrita deste trabalho, não haver uma API do Bard amplamente disponível para usuários e pesquisadores, como apontado no início do Capítulo 3, sendo assim, a resposta que é retornada é exatamente a mesma resposta que o Bard daria se estivesse sendo utilizado o chat no navegador. Da mesma forma em que foi feita a transmissão de elementos no GPT foi feita no Bard, com exatamente a mesma base de dados:

1. o texto que requer classificação e;
2. instruções explícitas formuladas em uma pergunta cuidadosamente elaborada.

Após isso, o Bard gera uma resposta com base no seu conhecimento preexistente. E então, um arquivo CSV é gerado com as classificações reais do conjunto de dados original, juntamente com a classificação atribuída pela ferramenta. O passo-a-passo dessa análise está detalhada no Algoritmo 2

Algorithm 1 Classificação de Texto com GPT

Inicialização:
api_*key* = "*sua*_chave_*da*_api"
modelo = "modelo_*escolhido*"
function GERARCLASSIFICACAO(prompt)
    while verdadeiro **do**
        try:
        resposta = openai.Completion.create(
        engine=modelo,
        prompt=prompt,
        temperature=0.2,
        max_*tokens* = 1024,
        top_p = 1,
        frequency$_{penalty}$=0,
        presence_*penalty*=0
)
        Retorne resposta.choices[0].text.strip()
        catch openai.error.RateLimitError **as** e:
        imprima("Limite de taxa atingido. Aguardando 60 segundos...")
        esperar(60)
        catch openai.error.APIError **as** e:
        if o status for 402 ou 403 **then**
            imprima("Limite máximo de uso atingido. Salvando resultados e saindo...")
            Quebre o loop
        else
            Levante uma exceção
        end if
    end while
end function
csvFinalResult = abrir('results/classification_*result.csv*',' w', *nova*_linha=")
writerResult = csv.writer(csvFinalResult)
dados = []
abrir("results/dataset_*label*_shuffled.csv", 'rt') como fileFake:
leitor = csv.reader(fileFake)
próximaLinha(leitor)
for row **do**
    prompt = row[1] + "Pergunta"
    classificação = gerarClassificação(prompt)
    dados.append([row[1], row[3], classificação])
end for
writerResult.writerow(['texto', 'isFakeNews', 'classificaçãoGpt'])
writerResult.escrever_*linhas*(*dados*)
Feche o arquivo

Algorithm 2 Classificação de Texto com Bard

Inicialização:
csvFinalResult = abrir_*arquivo*('*results*/*classification*_result1.csv', 'escrita', nova_*linha* =")
writerResult = csv.escrever(csvFinalResult)
dados = []
abrir_*arquivo*("*dataset*/*dataset*_label.csv", 'ler_*texto*')*comoarquivoFalso* :
leitor = csv.ler(arquivoFalso)
próximaLinha(leitor)
for row **do**
    prompt = row[1] + "pergunta"
    classificação = bardapi.core.Bard(cookie).obter_*resposta*(*prompt*)
    dados.append([row[1], row[3], classificação['conteúdo']])
end for
writerResult.writerow(['texto', 'isFakeNews', 'classificaçãoBard'])
writerResult.escrever_*linhas*(*dados*)
Feche o arquivo

3.5 Avaliação e Comparação de Resultados

Para avaliar a eficácia da abordagem proposta, foram empregadas métricas de avaliação amplamente reconhecidas no domínio da classificação de texto, como Precisão, Revocação, Acurácia e *F1-Score*. A Acurácia estima as previsões corretas do modelo em comparação com o número total de instâncias. A Precisão, por outro lado, quantifica a proporção de instâncias corretamente identificadas como positivas entre todas as instâncias previstas como positivas. A revocação, por sua vez, captura a fração de instâncias positivas reais que foram identificadas com precisão pelo modelo. Por fim, o *F1-Score* combina os valores de precisão e revocação, fornecendo assim uma avaliação abrangente do desempenho do modelo. Os cálculos para essas métricas são apresentados nas Equações 3.1, 3.2, 3.3 e 3.4, respectivamente.

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{3.1}$$

$$Precisão = \frac{VP}{VP + FP} \tag{3.2}$$

$$Revocação = \frac{VP}{VP + FN} \tag{3.3}$$

$$F1 - Score = \frac{2 \cdot Precisão \cdot Revocação}{Precisão + Revocação} \tag{3.4}$$

Nas equações, VP se refere a Verdadeiros Positivos, que são os exemplos positivos classificados corretamente; VN se refere a Verdadeiros Negativos, que são os exemplos negativos classificados corretamente; FP se refere a Falsos Positivos, que são os exemplos positivos classificados incorretamente; e FN se refere a Falsos Negativos, que são os exemplos negativos classificados incorretamente.

Por meio destas métricas, o objetivo é avaliar a capacidade das ferramentas GPT e Bard de classificar com precisão notícias falsas, comparando seus resultados com a verdade básica do conjunto de dados e também comparando os resultados de ambas ferramentas. A utilização dessas métricas fornece *insights* sobre o desempenho do modelo e seu potencial para a detecção de notícias falsas.

4 Resultados

Este capítulo apresenta os resultados obtidos pela metodologia e testes propostos no Capítulo 3.

4.1 GPT-3.5

O modelo text-davinci-003 foi utilizado em uma tentativa inicial. Os resultados não foram encorajadores, já que o modelo obteve uma acurácia de 50%. O modelo classificou todos os textos como notícias verdadeiras, mesmo que metade deles fosse falsa. A baixa precisão (0%) sugere um desempenho fraco na identificação correta de notícias falsas. Após o resultado desanimador, foi setado o modelo text-davinci-002 para avaliar seu desempenho. O modelo foi questionado com o seguinte prompt juntamente com o texto da notícia Pergunta 1: **"Does the given text is fake news? Does it Spread misinformation? Answer only with yes or no."**, em tradução direta: *"O texto fornecido é uma notícia falsa? Espalha desinformação? Responda apenas com sim ou não."*. Foi solicitado que a resposta fosse apenas 'sim' ou 'não' para simplificar o processo de cálculo das métricas de avaliação. Os resultados revelaram uma melhoria em comparação com o modelo anterior.

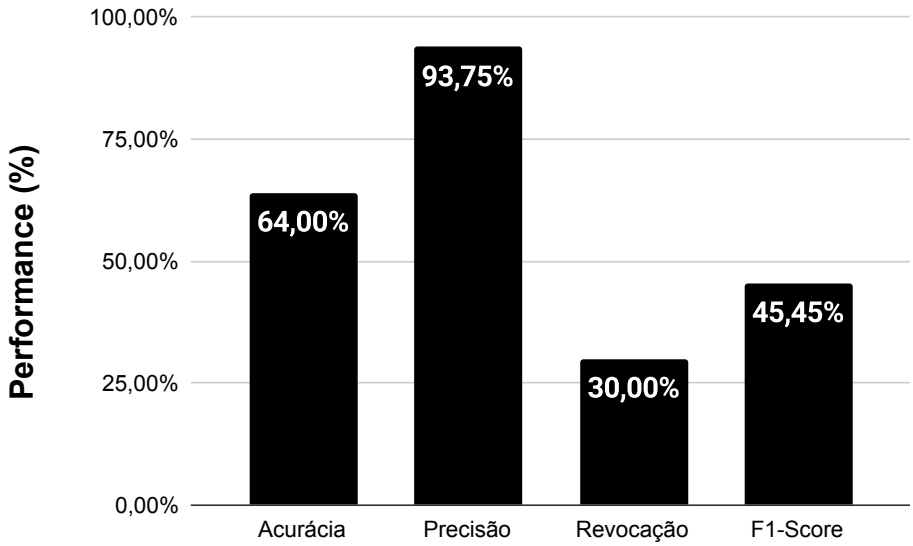


Figura 2 – Resultados - Pergunta 1

A Tabela 3 mostra a matriz de confusão detalhada para os resultados dos testes realizados com o modelo text-davinci-002 e utilizando a Pergunta 1.

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	30	70
	Verdadeira	2	98

Tabela 3 – Matriz de Confusão — Modelo: text-davinci-002 — Pergunta: 1

Os resultados mostrados na Figura 2 indicam uma precisão relativamente boa (93,8%), sugerindo que, quando o modelo classificou um texto como notícia falsa, estava provavelmente correto. No entanto, os baixos valores de revocação (30%), acurácia (64%) e F1-Score (45,5%) revelam a dificuldade do modelo em identificar corretamente um número significativo de casos de notícias falsas. Embora tenha havido uma melhoria na acurácia em comparação com o primeiro modelo, o desempenho geral não atendeu às expectativas. Nota-se que o GPT apresenta insegurança em identificar notícias falsas, portanto, foi executado novamente o modelo text-davinci-002, mas com um prompt diferente, agora pedindo para que o modelo identifique as características de uma notícia falsa. Uma tentativa para analisar se o modelo responde a pergunta com mais segurança. A Pergunta 2 foi a seguinte: **"Does the given text contain characteristics of fake news? Does it spread misinformation? Answer only with yes or no."**, em tradução direta: *"O texto fornecido contém características de notícias falsas? Espalha desinformação? Responda apenas com sim ou não."*. Os resultados estão exemplificados na Figura 3.

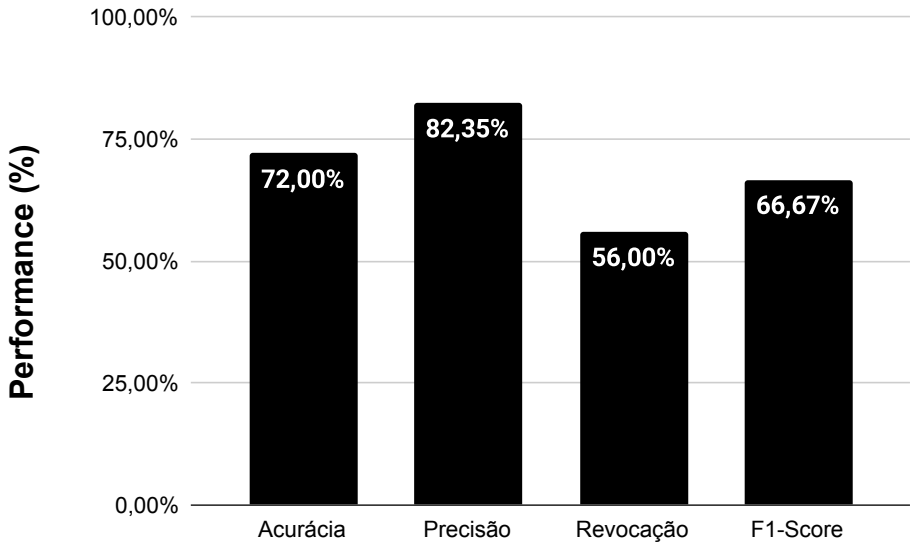


Figura 3 – Métricas de avaliação - Pergunta 2

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	56	44
	Verdadeira	12	88

Tabela 4 – Matriz de Confusão — Modelo: text-davinci-002 — Pergunta: 2

Comparando os resultados apresentados nas Figuras 2 e 3, é possível afirmar que o modelo teve um desempenho melhor quando a pergunta se concentrou na presença de características de notícias falsas. Em geral, houve melhorias na acurácia (72%), revocação (56%) e F1-Score (66,6%), indicando uma detecção mais confiável de notícias falsas em comparação com a primeira pergunta. A precisão atingiu 82,3%. Na Tabela 4 é apresentada a matriz de confusão do modelo quando utilizada a pergunta 2. Esses resultados demonstram que a solução atual é dependente da formulação da pergunta. Para detectar notícias falsas com precisão, são necessárias melhorias adicionais.

Em mais uma tentativa para melhorar os resultados, a ferramenta foi executada novamente, com as mesmas duas perguntas e a mesma base de dados, porém com o modelo gpt-3.5-turbo-instruct, uma versão atualizada do text-davinci-003.

Como indicado na Figura 4, os resultados não foram animadores ao utilizar a Pergunta 1 para obter resultados, o modelo não foi eficaz ao classificar notícias falsas como indica a taxa de 13% de revocação e um F1-Score de apenas 22.22% indicando a baixa performance do modelo. Na Tabela 5 é mostrada a matriz de confusão desta tentativa.

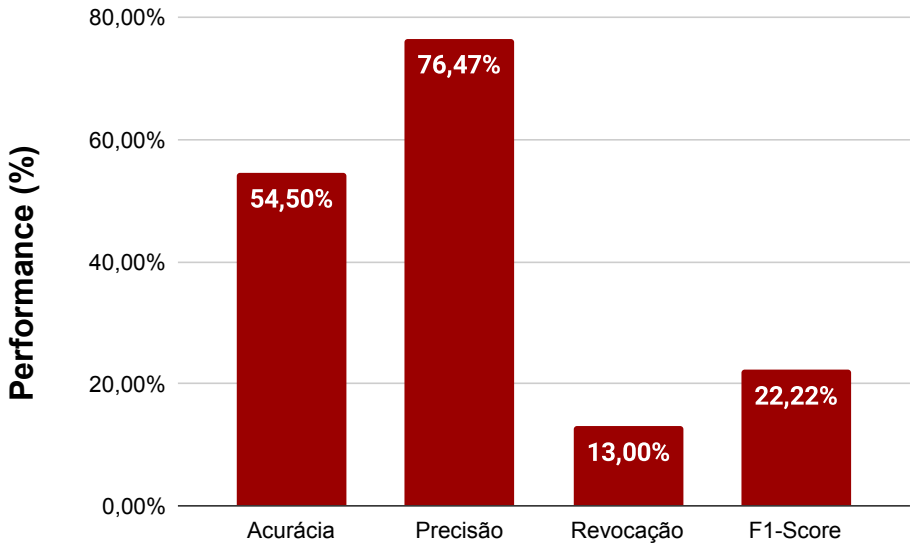


Figura 4 – Métricas de avaliação - Pergunta 1 - Modelo 2

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	13	87
	Verdadeira	4	96

Tabela 5 – Matriz de Confusão — Modelo: gpt-3.5-turbo-instruct — Pergunta: 1

Já ao utilizar a Pergunta 2 no novo modelo, podemos observar na Figura 5 que obtivemos um desempenho mais consistente. Esses resultados destacam uma melhoria substancial em comparação à Pergunta 1, e como observado anteriormente, quando o foco da pergunta são as características de *fake news* a ferramenta apresenta uma melhoria de desempenho.

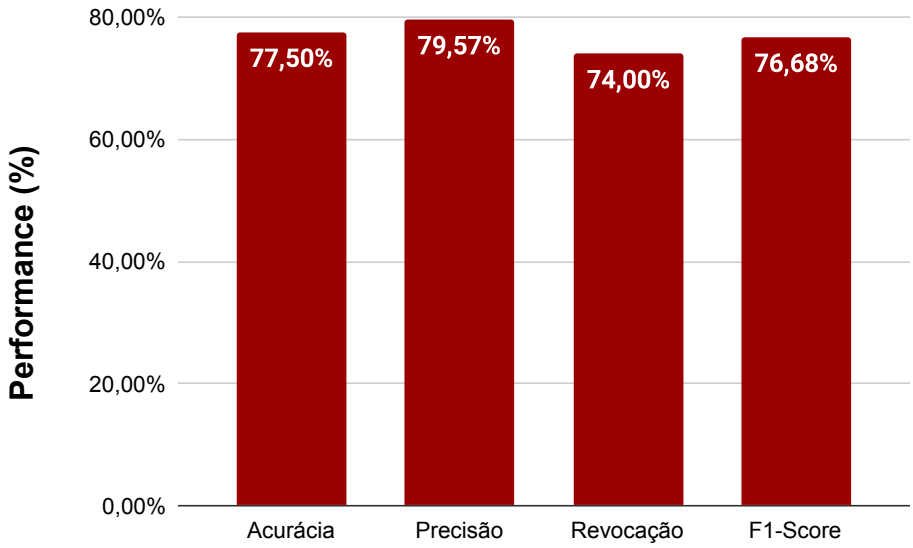


Figura 5 – Métricas de avaliação - Pergunta 2 - Modelo 2

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	74	26
	Verdadeira	19	81

Tabela 6 – Matriz de Confusão — Modelo: gpt-3.5-turbo-instruct — Pergunta: 2

4.2 Bard

Como não é possível selecionar um modelo que não o de chat para o Bard, foram realizadas rodadas de testes com a mesma base de dados utilizada no GPT e também com os mesmos *prompts*.

Com a Pergunta 1, os seguintes resultados exemplificados na Figura 6 foram observados.

Observa-se que, a ferramenta, atingiu uma precisão de 100%, ou seja, todos as vezes que ela classificou o texto como falso, ela estava correta, porém, a revocação foi muito baixa, apenas 7,25%, acurácia e F1-Score também obtiveram resultados abaixo do esperado, isso revela a dificuldade do modelo em identificar corretamente um número significativo de casos de notícias falsas.

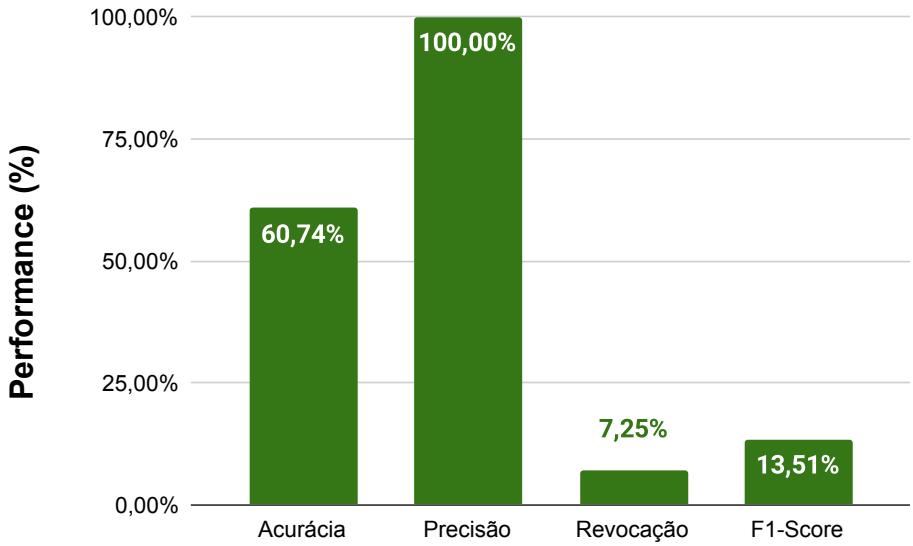


Figura 6 – Métricas de avaliação - Pergunta 1 - Bard

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	5	64
	Verdadeira	0	94

Tabela 7 – Matriz de Confusão — Modelo: bard — Pergunta: 1

Quando enviada a Pergunta 2 como *prompt*, os resultados indicaram uma melhoria na perfomance geral do sistema, evidenciada pelo aumento do F1-Score. Houve também uma melhoria na acurácia, indicando que, dentre todas as classificações, o modelo classificou corretamente 67,28% dos textos. A revocação também registrou uma melhoria, com a nova pergunta, o modelo classificou corretamente 26,47% dos textos que eram para

serem classificados como positivo. Por outro lado, a precisão caiu, se comparado com a pergunta anterior, indicando que a ferramenta acertou 85,71% das vezes em que indicou a classificação como uma notícia falsa. A Figura 7 mostra um gráfico com os resutados obtidos pelo Bard com a Pergunta 2.

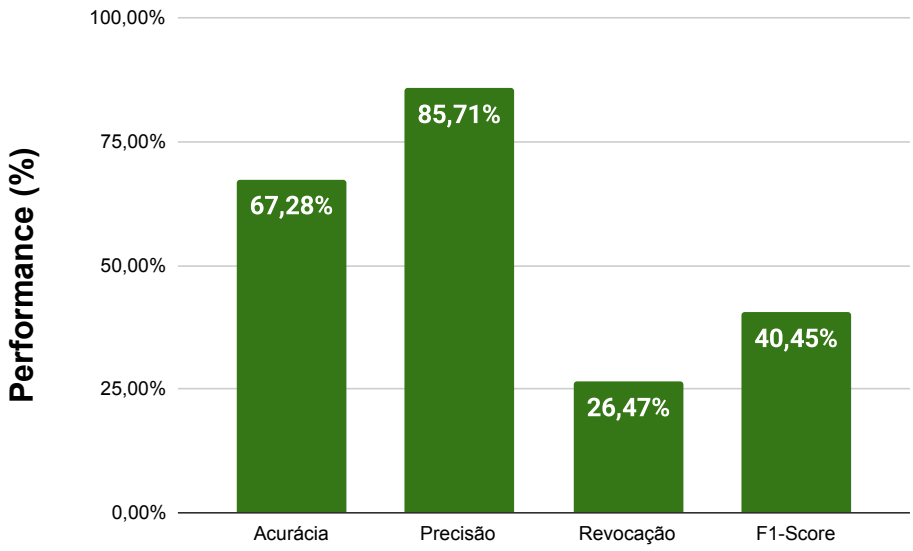


Figura 7 – Métricas de avaliação - Pergunta 2 - Bard

		Valores Preditos	
		Falsa	**Verdadeira**
Valores Reais	**Falsa**	18	50
	Verdadeira	3	91

Tabela 8 – Matriz de Confusão — Modelo: bard — Pergunta: 1

As Tabelas 7 e 8 apresentam as matrizes de confusão das Perguntas 1 e 2 respectivamente para o modelo Bard.

Apesar da melhoria em comparação com a Pergunta 1, os resultados ainda não foram satisfatórios. Vale ressaltar também que, o Bard não foi capaz de classificar alguns textos em ambas as perguntas, como resposta a ferramenta dizia que não foi desenvolvida para realizar esse tipo de tarefa.

A Figura 8 mostra que, dos 200 textos, a ferramenta deixou de classificar 37 na pergunta 1 e, na pergunta 2, o Bard não conseguiu indicar uma categoria para 38 textos, como exemplificado na Figura 9. Lembrando que, os resultados apresentados dados pelo Bard, são considerando apenas os textos que receberam classificação pela ferramenta.

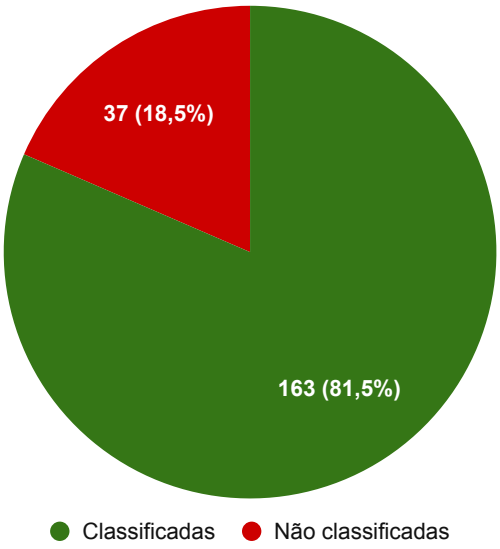


Figura 8 – Notícias Classificadas e não Classificadas - Pergunta 1 - Bard

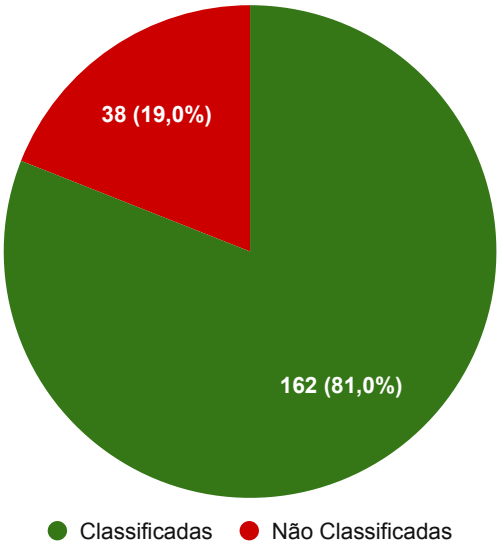


Figura 9 – Notícias Classificadas e não Classificadas - Pergunta 2 - Bard

4.3 Comparações

Baseado na Pergunta 1, como mostrado no gráfico apresentado na Figura 10, as ferramentas apresentaram algumas semelhanças em seus desempenhos.

O Primeiro modelo do GPT demonstrou uma alta precisão de 93,75%, indicando que o modelo é eficaz quando classifica um texto como falso. No entanto, a baixa taxa de revocação indica que a ferramenta tem uma tendência a não identificar um texto como notícia falsa.

Por outro lado, o Bard indicou uma taxa de precisão de 100%, o que pode parecer

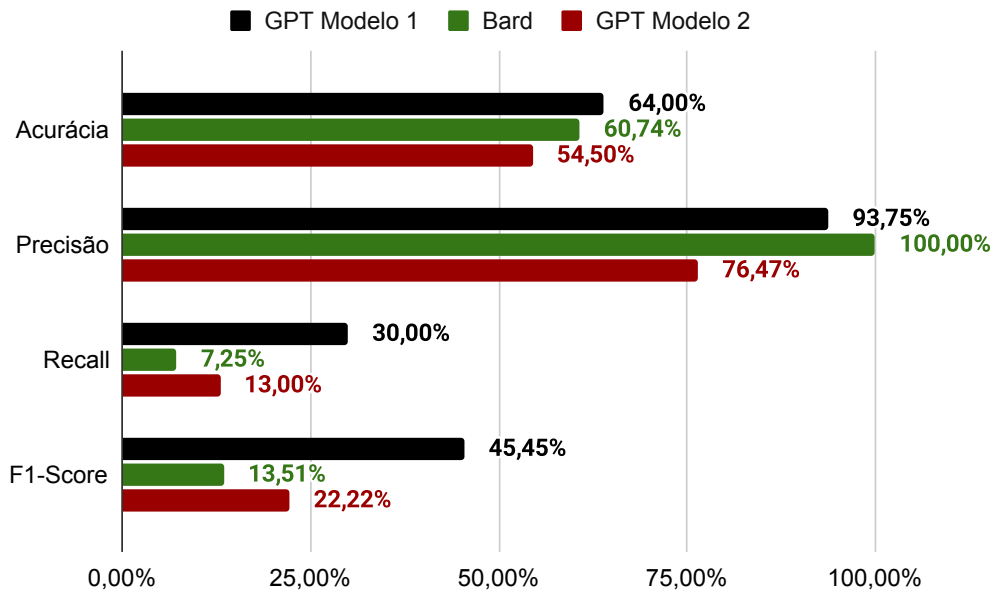


Figura 10 – Comparação Ferramentas - Pergunta 1

impressionante, porém, ao olhar para a baixa taxa de revocação (7,25%) é possível concluir que o Bard quase não identificou textos como falsos.

Por fim, o segundo modelo do GPT apresentou uma taxa de precisão mais baixa que dos testes anteriores, de aproximadamente 77% e também apresentou uma baixa taxa de revocação (13%).

Essa comparação evidencia que uma alta taxa de precisão não necessariamente indica que o modelo é eficaz ao realizar uma tarefa, destacando a importância das outras métricas de avaliação. Também, pode-se destacar a importância das melhorias e otimizações nos 3 modelos apresentados, bem como, a otimização da pergunta que é feita pros modelos. Com base nisso é possível observar uma melhoria nos resultados quando a pergunta é focada nas características das notícias falsas, como evidenciado na Figura 11, que apresenta os resultados das ferramentas com base na Pergunta 2.

Comparando os resultados com a primeira pergunta, é possível observar que, as ferramentas apresentaram uma melhoria na perfomance geral, observada na taxa de acurácia, que subiu para todos os modelos. Vale destacar que, o Modelo gpt-3.5-turbo-instruct do GPT, foi o que apresentou o melhor desempenho dentre os 3 para a tarefa de detecção de notícias falsas.

Essa comparação reforça que, o desempenho dos modelos pode variar significativamente dependendo da forma em que a pergunta foi formulada. Ela também destaca a superioridade do Modelo 2 do GPT nessa tarefa, apresentando uma sólida pontuação de 76,68% no F1-Score, a melhor observada neste estudo.

É vital destacar que o desempenho de cada modelo pode variar significativamente com base no conjunto de dados fornecido e nos parâmetros do modelo, destacando a necessidade contínua de pesquisa para entender a capacidade dos modelos de linguagem em tarefas de detecção de notícias falsas.

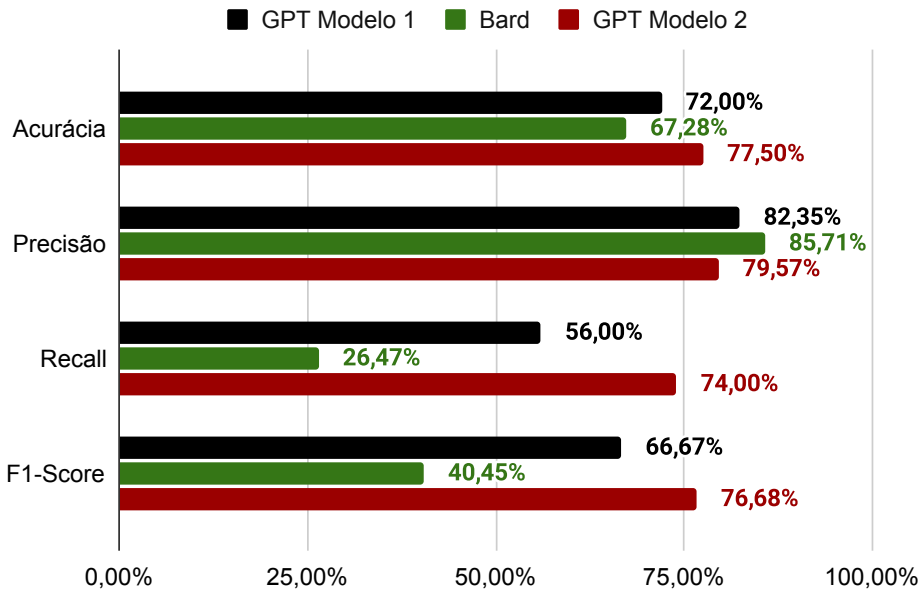


Figura 11 – Comparação Ferramentas - Pergunta 2

4.4 Aplicabilidade

Como discutido no Capítulo 3, o GPT contempla uma API oficial com documentação vasta e de fácil acesso que pode ser aplicada em qualquer tarefa que envolva compreensão ou geração de linguagem natural (OpenAI, 2023b). Os resultados envolvendo o GPT-3.5 são promissores, apontando que quanto melhor formulada a pergunta, melhor o desempenho do modelo na tarefa de detecção de *fake news*. Isso evidencia que o GPT tem potencial para ser o motor para detecção de notícias falsas em outras ferramentas, como uma extensão web, por exemplo.

No entanto, o Bard ainda não disponibiliza uma API oficial de maneira ampla, como citado em 3.4.2, isso impede a criação de ferramentas baseadas no Bard para detecção de *fake news*. Apesar disso, os resultados demonstraram que houve uma melhoria quando a pergunta foi focada nas características de notícias falsas, ainda que os resultados tenham sido baixos, isso indica que há potencial de melhoria na ferramenta, principalmente quando houver uma API semelhante à do GPT.

Uma vantagem de utilizar essas ferramentas na classificação de notícias falsas é a não necessidade em treinar o modelo visto que, tais ferramentas são treinadas previamente pelas empresas responsáveis por elas.

O código fonte das tarefas realizadas que levou a estes resultados estão disponíveis em:

- GitHub - Lucas Sousa dos Anjos - GPT (2023)
- GitHub - Lucas Sousa dos Anjos - Bard (2023)

5 Conclusão

Neste trabalho, foi investigado o desempenho das grandes ferramentas de IA disponíveis atualmente na detecção de notícias falsas. Através de uma análise experimental de um conjunto de artigos de notícias, tanto verdadeiras quanto falsas, foi observado que a forma como o *prompt* é formulado e apresentado às ferramentas influencia significativamente a qualidade e a precisão das respostas. Especificamente, ao solicitar que as ferramentas identificassem se o texto fornecido continha características de notícias falsas, é possível observar uma melhoria significativa nos resultados. Isso sugere que direcionar a pergunta de maneira mais específica e focada nas características das notícias falsas aprimora a eficácia dos modelos na detecção desse tipo de conteúdo.

Os resultados preliminares indicam a superioridade do GPT em comparação com o Bard em realizar a detecção de *fake news*. Essa superioridade pode ser explicada, em parte, ao fato do Bard não possuir uma API oficial disponível para o público geral onde é possível indicar modelos e o nível de criatividade de respostas como evidenciado com mais detalhes na subseção 3.4.2.

Em síntese, as descobertas apresentadas neste estudo, indicam a eficácia do GPT-3.5 na classificação de notícias falsas quando o *prompt* enviado contempla perguntas específicas sobre suas características. Pesquisas futuras podem explorar como esse modelo pode ser ainda mais otimizado para detectar notícias falsas em plataformas de mídia social e outras fontes online onde a desinformação é prevalente. Além disso, uma promissora linha de investigação seria avaliar os resultados do Bard quando houver uma API equivalente à API do GPT disponível de forma mais ampla. Bem como realizar a avaliação de como outros modelos de linguagem se comparam com os modelos presentes neste trabalho em termos de eficácia na detecção de notícias falsas.

Finalmente, em trabalhos futuros, é importante considerar uma base de textos maior para análise, incluir textos em diversas línguas e também desenvolver ferramentas utilizando desses modelos para a detecção de notícias falsas, como uma extensão web e um aplicativo móvel, por exemplo. Também é importante aplicar os testes feitos aqui, e mais amplos, com uma API oficial do Google Bard, quando disponível, além de realizar aprimoramentos no *prompt* enviado às ferramentas afim de obter resultados mais relevantes bem como realizar pesquisas nas versões mais recentes do GPT. Outro caminho a se seguir é comparar os resultados deste trabalho com ferramentas desenvolvidas especificamente para classificação de texto, como o BERT, por exemplo.

Referências

ALVES, M. A. S.; MACIEL, E. R. H. O fenômeno das fake news: definição, combate e contexto. **InternetLab**, UFMG, 2020. Disponível em: <http://hdl.handle.net/1843/44432>. Citado na página 17.

Amazon AWS. **What is GPT?** 2023. Disponível em: <https://aws.amazon.com/what-is/gpt/?nc1=h_ls>. Acesso em: 30 de maio 2023. Citado na página 16.

ASLAM, N.; KHAN, I. U.; ALOTAIBI, F. S.; ALDAEJ, L. A.; ALDUBAIKIL, A. K. Fake detect: A deep learning ensemble model for fake news detection. **Complexity**, Hindawi, 2021. Disponível em: <https://doi.org/10.1155/2021/5557784>. Citado 2 vezes nas páginas 18 e 21.

AYDIN, Google bard generated literature review: Metaverse. **Journal of AI**, İzmir Academy Association, v. 7, n. 1, p. 1–14, 2023. Citado na página 16.

AYDIN Ömer. Google bard generated literature review: Metaverse. **Journal of AI**, Journal of AI, 2023. Citado na página 16.

BAARIR, N. F.; DJEFFAL, A. Fake news detection using machine learning. **2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)**, IEEE, 2021. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9378748>. Citado 2 vezes nas páginas 18 e 21.

BOZKUş, E. **ISOT Fake News Dataset**. 2023. Disponível em: <https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets>. Acesso em: 28 de outubro 2023. Citado na página 23.

Daniel Park. **Google Bard API**. 2023. Disponível em: <https://github.com/dsdanielpark/Bard-API>. Acesso em: 28 de outubro 2023. Citado na página 22.

DEREK. **Fake News Detector**. 2018. Disponível em: <https://chrome.google.com/webstore/detail/fake-news-detector/ijfgnjaoiknhapbpafkehcngdnmgfnmf?hl=pt-br>. Acesso em: 18 de abril 2023. Citado na página 13.

DESIGN, R. **Detektor dezinformačných webov**. 2020. Disponível em: <https://chrome.google.com/webstore/detail/detektor-dezinforma%C4%8Dn%C3%BDch/ajfhmidimnkpbhnkcckllicmhhdipmoo?hl=pt-br>. Acesso em: 18 de abril 2023. Citado na página 13.

DEVELOPERS, D.; FAVA, G. **Fake news detector**. 2017. Disponível em: <https://chrome.google.com/webstore/detail/fake-news-detector/aebaikmeedenaijgjcfmndfknoobahep?hl=pt-br>. Acesso em: 18 de abril 2023. Citado na página 13.

FELDMAN, S. Nlp meets the jabberwocky: Natural language processing in information retrieval. **Information Today, Inc**, Information Today, Inc, 1999. Citado na página 14.

GitHub - Lucas Sousa dos Anjos - Bard. **Projeto Lucas Bard - Python**. 2023. Disponível em: <https://github.com/lucassousaan/bard>. Acesso em: 10 de janeiro de 2024. Citado na página 38.

GitHub - Lucas Sousa dos Anjos - GPT. **Projeto Lucas GPT - Python**. 2023. Disponível em: <https://github.com/lucassousaan/fake_news_project>. Acesso em: 10 de janeiro de 2024. Citado na página 38.

Google. **Google Bard**. 2023. Disponível em: <https://bard.google.com/?hl=pt>. Acesso em: 28 de outubro de 2023. Citado na página 22.

HENRIQUES, M. J.; NETO, F. L. **FakeNewsBR**. 2022. Disponível em: <https:/fakenewsbr.com>. Para mais informações: <https://cemeai.icmc.usp.br/fato-ou-fake-descobrir-se-uma-noticia-e-real-tambem-e-papel-da-matematica/>. Acesso em: 18 de abril 2023. Citado na página 13.

IBM. **What is machine learning?** 2021. Disponível em: <https://www.ibm.com/topics/machine-learning>. Acesso em: 30 de maio 2023. Citado na página 14.

_____. **What is natural language processing?** 2021. Disponível em: <https://www.ibm.com/topics/natural-language-processing>. Acesso em: 30 de maio 2023. Citado na página 15.

JAIN, A.; KASBE, A. Fake news detection. **2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences**, IEEE, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8546944>. Citado 2 vezes nas páginas 19 e 21.

KHIVASARA, Y.; KHARE, Y.; BHADANE, T. Fake news detection system using web-extension. **2020 IEEE Pune Section International Conference**, IEEE, 2020. Citado 2 vezes nas páginas 17 e 21.

LazerLikeFocus. **Fake News Detector**. 2020. Disponível em: <https://play.google.com/store/apps/details?id=com.lazerlikefoucs.whatsappfakenewsdetector3>. Acesso em: 18 de abril 2023. Citado na página 13.

MONTEIRO, R. A. **FakeCheck**. 2018. Disponível em: <https://nilc-fakenews.herokuapp.com/>. Acesso em: 18 de abril 2023. Citado na página 13.

MOREIRA, L. S.; LUNARDI, G. M.; RIBEIRO, M. de O.; SILVA, W.; BASSO, F. P. A study of algorithm-based detection of fake news in brazilian election: Is bert the best? **IEEE Latin America Transactions**, Hindawi, v. 21, n. 8, p. 897–903, Sep. 2023. Disponível em: <https://latamt.ieeer9.org/index.php/transactions/article/view/7900>. Citado 2 vezes nas páginas 20 e 21.

NAQA, I. E.; MURPHY, M. J. What is machine learning? In: _____. **Machine Learning in Radiation Oncology: Theory and Applications**. Cham: Springer International Publishing, 2015. p. 3–11. ISBN 978-3-319-18305-3. Disponível em: <https://doi.org/10.1007/978-3-319-18305-3_1>. Citado na página 14.

Oigetit, Inc. **Oigetit Fake News Filter**. 2019–2022. Disponível em: https://play.google.com/store/apps/details?id=io.scal.oigetithl=ptgl=US. Acesso em: 18 de abril 2023. Citado na página 13.

OpenAI. **Models**. 2023. Disponível em: <https://platform.openai.com/docs/models/gpt-3-5>. Acesso em: 20 de julho de 2023. Citado na página 24.

OpenAI. **Models - OpenAI**. 2023. Disponível em: <https://platform.openai.com/docs/models/gpt-3-5>. Acesso em: 04 de novembro de 2023. Citado na página 16.

_____. **OpenAI API**. 2023. Disponível em: <https://platform.openai.com/>. Acesso em: 05 de novembro de 2023. Citado na página 38.

_____. **OpenAI Libraries**. s.d. Disponível em: <https://platform.openai.com/docs/libraries/python-library>. Acesso em: 28 de outubro de 2023. Citado na página 22.

Pedro Guimarães; Cleber Rodrigues. **4 em cada 10 brasileiros afirmam receber fake news diariamente**. 2022. Disponível em: <https://www.cnnbrasil.com.br/nacional/4-em-cada-10-brasileiros-afirmam-receber-fake-news-diariamente/>. Acesso em: 17 de abril 2023. Citado na página 12.

PINHEIRO, C. **Vacinas de DNA e RNA contra coronavírus não causam alterações nos genes**. 2021. Disponível em: <https://saude.abril.com.br/medicina/vacinas-de-dna-e-rna-contra-coronavirus-nao-causam-alteracoes-nos-genes/>. Acesso em: 27 de abril 2023. Citado na página 11.

RAZA, S.; DING, C. Fake news detection based on news content and social contexts: a transformer-based approach. **International Journal of Data Science and Analytics**, Springer Nature, 2022. Disponível em: <https://doi.org/10.1007/s41060-021-00302-z>. Citado 2 vezes nas páginas 18 e 21.

Redação Estadão. **Site distorce notícia de TV australiana para alegar que China testou o coronavírus como arma biológica**. 2021. Disponível em: https://www.estadao.com.br/estadao-verifica/site-distorce-noticia-de-tv-australiana-para-alegar-que-china-testou-o-coronavirus-como-arma-biologica/. Acesso em: 27 de abril 2023. Citado na página 11.

ROCHA, Y. M.; MOURA, G. A. de; DESIDéRIO, G. A.; OLIVEIRA, C. H. de; LOURENçO, F. D.; NICOLETE, L. D. de F. The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. **Journal of Public Health**, Springer Nature, 2021. Citado na página 11.

SAMADI, M.; MOUSAVIAN, M.; MOMTAZI, S. Deep contextualized text representation and learning for fake news detection. **Information Processing Management**, v. 58, n. 6, p. 102723, 2021. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457321002077>. Citado 2 vezes nas páginas 20 e 21.

SANTOS, A. A. dos. **Fake news agregador**. 2020–2022. Disponível em: <https://play.google.com/store/apps/details?id=atila.dev.check_fake_news&hl=pt>. Acesso em: 18 de abril 2023. Citado na página 13.

SCHÜTZ, M.; SCHINDLER, A.; SIEGEL, M.; NAZEMI", K. Automatic fake news detection with pre-trained transformer models. **Pattern Recognition. ICPR International Workshops and Challenges**, Springer International Publishing, p. 627–641, 2021. Citado 2 vezes nas páginas 19 e 21.

SINGH, S.; MAHMOOD, A. The nlp cookbook: Modern recipes for transformer based deep learning architectures. **IEEE Access**, v. 9, p. 68675–68702, 2021. Citado na página 15.

Tribunal Superior Eleitoral. **Pílulas contra a desinformação: notícias falsas circulam 70% mais rápido do que as verdadeiras**. 2022. Disponível em: https://www.tse.jus.br/comunicacao/noticias/2022/Junho/pilulas-contra-a-desinformacao-noticias-falsas-circulam-70-mais-rapido-do-que-as-verdadeiras. Acesso em: 27 de abril 2023. Acesso em: 18 de abril 2023. Citado na página 11.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. **Advances in Neural Information Processing Systems**, Curran Associates, Inc., 2017. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado na página 15.

ÖZBAY, F.; ALATAS, B. A novel approach for detection of fake news on social media using metaheuristic optimization algorithms. **Elektronika ir Elektrotechnika**, Elektronika ir Elektrotechnika, 2019. Disponível em: <https://doi.org/10.5755/j01.eie.25.4.23972>. Citado 2 vezes nas páginas 18 e 21.

Apêndices

APÊNDICE A – Artigo Publicado (SBSeg 2023)

O presente trabalho foi publicado no XXIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg 2023).

Investigating the Performance of the GPT-3.5 Model in Fake News Detection: An Experimental Analysis

Lucas S. Anjos[1], **Silvio E. Quincozes**[2], **Juliano F. Kazienko**[3] **e Vagner E. Quincozes**[4]

[1]Universidade Federal de Uberlândia (FACOM)

[2] Universidade Federal do Pampa (UNIPAMPA)

[3] Universidade Federal de Santa Maria (UFSM)

[4] Universidade Federal Fluminense (UFF).

lucassousaanjos@ufu.br, silvioquincozes@unipampa.edu.br,

kazienko@redes.ufsm.br, vequincozes@id.uff.br

Abstract. *The dissemination of fake news has become a significant concern in the current society. This problem is evident on social media platforms, where the spread of misinformation has become a constant presence in the daily lives of many individuals. In this work, we investigate the performance of the GPT-3.5 model in classifying fake and real news, considering 200 newspaper articles and two strategies for question formulation. Our results reveal that using a well-formulated question is crucial to obtain more precise responses. In particular, we observed an improvement of 21.1% in the F1-Score metric by directing the question to focus on the characteristics of a fake text.*

1. Introduction

The intensive propagation of false news, known as "fake news", has caused concern in society's daily life. It mainly affects social media, where false content spreads at alarming speeds, as pointed out by the TSE [Tribunal Superior Eleitoral 2022]. In 2022, fake news circulated 70% faster than true news. Such content has the potential to cause serious harm to society (*e.g.* in public health, where malicious personnel trigger fear and stress in the affected individuals [Rocha et al. 2021]). Also, the spreading of theories such as the claim that the COVID-19 vaccine alters human DNA [Government 2023] has contributed to the propagation of misinformation and, accordingly, people's refusal to take vaccines.

To combat the dissemination of fake news and strengthen the reliability of information sources, it is crucial to address this problem and develop accurate and user-friendly tools for fake news detection. In this context, ChatGPT – a language model trained by OpenAI that became popular recently – is capable of providing responses and information in text, addressing various areas of knowledge based on its training. Therefore, it has the potential to be used in the analysis of false texts [Khivasara et al. 2020].

In this work, we explore the application of ChatGPT to tackle the challenge of fake news detection. Our main objective is to propose and evaluate the feasibility of using the ChatGPT-3.5 model as a central component of a fake news detection system. We believe that the application of advanced language models such as ChatGPT can provide a new perspective on fake news detection and contribute to mitigating this problem. To evaluate the feasibility, we conducted experiments to investigate the effectiveness and limitations of this approach, as well as its potential for future enhancements. Our results show that

the way the questions are formulated influences the quality and accuracy of the answers, reaching approximately 93.8% in the accuracy metric.

2. Related Works

In this section, we present relevant works. For that, we start by summarizing a comparison among these works and their main characteristics in Table 1.

Reference	Scope	Use GPT	GPT Version
[Khivasara et al. 2020]	Fake News	Yes	2.0
[Raza and Ding 2022]	Fake News	No	*
[Baarir and Djeffal 2021]	Fake News	No	*
[Özbay and Alatas 2019]	Fake News	No	*
[Aslam et al. 2021]	Fake News	No	*
This work	**Fake News**	**Yes**	**3.5**

Tabela 1. Comparison of Academic Works.

Different approaches have been proposed to detect fake news and improve news credibility. Some studies employed deep learning techniques, such as Long Short-Term Memory (LSTM) and GPT-2 models [Khivasara et al. 2020], whereas others explored the Transformer architecture to leverage news information and social contexts [Raza and Ding 2022]. With respect to [Khivasara et al. 2020], the GPT-2 model was utilized to determine whether the content of purported fake news originated from an Artificial Intelligence (AI) generator, rather than employing it to verify the authenticity of the news itself. Also, some studies utilized machine learning techniques, such as Term Frequency – Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM), to extract relevant features and classify texts as fake or genuine [Baarir and Djeffal 2021]. Additionally, metaheuristic algorithms, such as Grey Wolf Optimization (GWO) and Salp Swarm Optimization Algorithm (SSO), showed promise in fake news detection [Özbay and Alatas 2019]. Although these approaches achieved satisfactory results, there are challenges, such as bias and lack of adaptability to different languages and datasets.

It is also important to highlight the existence of tools that help identify misleading news, such as *Fake news detectors*[1][2], however, most of them do not support multiple languages. Other tools, such as *FakeNewsBR*[3] and *FakeCheck*[4] accept text in Portuguese but lack usability, meaning they are not intuitive. There are also mobile applications such as *Fake News Detector*[5], *Oigetit Fake News Filter*[6], and *Fake news aggregator*[7], which aim to detect fake news. However, such applications are limited as they only support texts in the English language or function as aggregators of the main Brazilian fake news

[1]https://chrome.google.com/webstore/detail/fake-news-detector/aebaikmeedenaijgjcfmndfknoobahep
[2]https:// chrome.google.com/webstore/detail/fake-news-detector/ ijfgnjaoiknhapbpafkehcngdnmgfnmf
[3]https:/fakenewsbr.com
[4]http://nilc-fakenews.herokuapp.com/
[5]https://play.google.com/store/apps/details?id=com.lazerlikefoucs.whatsappfakenewsdetector3
[6]https://play.google.com/store/apps/details?id=io.scal.oigetit
[7]https://play.google.com/store/apps/details?id=atila.dev.check_fake_news

websites. Consequently, if the desired news is not included in their listings, users are left without a satisfactory answer.

Therefore, it is evident that there is a gap in the existing tools: none of them offer a good level of usability and support for texts in all languages. Furthermore, only one academic paper uses GPT, albeit an outdated one (*i.e.*, 2.0) to detect fake news. Based on this, we intend to investigate the reliability of ChatGPT-3.5 in detecting fake news, seeking to directly or indirectly impact the development of solutions for this purpose.

3. Proposed Methodology

To address the aforementioned issues, in this work, we propose a novel methodology based on the Chat-GPT-3.5 platform. We adopted a process composed of three steps conceived to fulfill the goals of classifying news as either false or true information:

1. ***Data Selection and Preparation.*** We selected a dataset containing fake news, named ISOT Fake News Dataset[8]. The dataset comprises two types of articles: genuine news and fake news, collected from real-world sources. This dataset encompasses 21,417 authentic articles and than 23,481 fake articles. As per the dataset description, the data underwent a cleaning and pre-processing process, although punctuation and errors in the fake news were retained in the text. In this study, 200 texts were selected, with 100 of them being genuine news and the remaining 100 being fake news, all automated through a Python script. It's worth noting that the use of these data is solely for testing purposes, as OpenAI is responsible for the ongoing training and validation of its language models.
2. ***ChatGPT Communication*** To incorporate the communication with Chat-GPT into our Python code, the steps necessary were: i) account creation into the OpenAI platform; by visiting the official website and subsequently logging in. Once logged in, the API Keys section was accessed. Within this section, a new API key was generated and copied. In the Python code, the OpenAI library was imported to enable its functionalities. Finally, the API key was set using the `openai.api_key method`, as detailed in the Algorithm 1, line 2. These steps allowed for the integration of GPT into the Python code.
3. ***Text Classification.*** Subsequently, we employed a Python script to perform the classification of the selected texts. This script utilized the GPT API provided by OpenAI, enabling the GPT-3.5 model to classify the texts. OpenIA developed a solution to improve the readability when processing natural language by taking human feedback into account. This solution is called InstructGPT. Based on InstructGPT, they created the `text-davinci-002` model, which is trained with supervised fine-tuning. Lastly, OpenIA improved that model by replacing such an approach with reinforcement learning. The improved model was called `text-davinci-003`. As a result, the latter can process any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models (other available models for use) [OpenAI 2023]. Since ChatGPT-3.5 lacks a dedicated API specifically designed for text classification tasks, our methodology involves transmitting two distinct elements: (i) the text that requires classification, and (ii) explicit instructions articulated in a carefully

[8] `www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets`

formulated question. By doing so, we enable GPT-3.5 to generate a pertinent response for text classification, drawing on its pre-existing knowledge base. As a result, a CSV file was generated, which includes the actual classifications from the original dataset, along with the classifications assigned by GPT. The steps of this script are outlined in the pseudocode denoted in Algorithm 1.

Algorithm 1 Text Classification using GPT

Initialization:
  Set `openai.api_key` to the provided API key value
  Set `model` to "text-davinci-002"
Function `generate_classification(prompt)`:
  While true:
    Try:
      Get the response from GPT API using `openai.Completion.create()`
      Return the GPT model response without whitespace
    Catch `openai.error.RateLimitError` **as** `e`:
      Print "Rate limit reached. Waiting for 60 seconds..."
      Sleep for 60 seconds
    Catch `openai.error.APIError` **as** `e`:
      If the status is 402 or 403:
        Print "Maximum usage limit reached."
      Break the loop
      Otherwise:
        Raise an exception
Open `csvFinalResult` file in write mode (`'w'`) with `newline=''`
Create a writer `writerResult` for the `csvFinalResult` file
Create an empty list called `data`
Open the `file` file in read mode with `'rt'`
Read the next line from the file to skip the header
For each row **in** the file reader:
  Construct the prompt by concatenating the text from the second column (`row[1]`) with the English question
  Call the `generate_classification` function with the prompt to get the GPT classification
  Append `[row[1],row[3],classification]` to the `data` list
Write the header line `['text', 'is_fake_news', 'gpt_classification']` using `writerResult`
Write the data rows from `data` using `writerResult`
Close the `file`

4. Experiments

In order to assess the efficacy of the proposed approach, we employed widely recognized evaluation metrics in the domain of text classification, such as Accuracy, Precision, Recall, and F1-Score. Accuracy estimates of the model's correct predictions in comparison to the total number of instances. Precision, on the other hand, quantifies the proportion of instances correctly identified as positive amongst all instances predicted as positive. Recall, alternatively, captures the fraction of actual positive instances that were accurately identified by the model. Lastly, the F1-Score amalgamates the values of precision and recall, thereby yielding a comprehensive evaluation of the model's performance. The computations for these metrics are presented in Equations 1, 2, 3, and 4, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

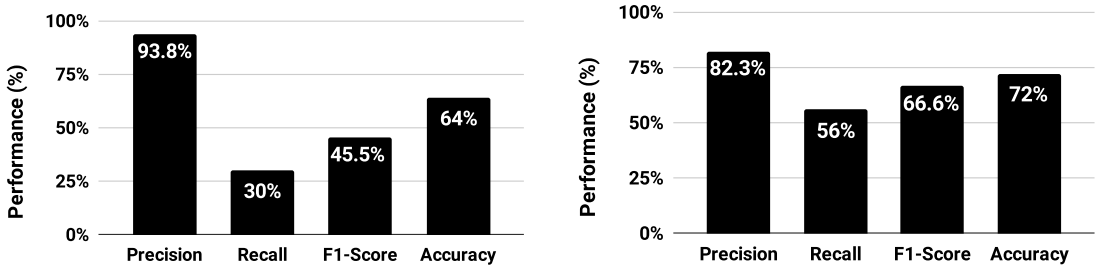
$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

In equations, `TP` refers to True Positives, which are the positive examples correctly classified; `TN` refers to True Negatives, which are the negative examples correctly classified; `FP` refers to False Positives, which are the positive examples incorrectly classified; and `FN` refers to False Negatives, which are the negative examples incorrectly classified.

Through this methodology, we aim to evaluate the GPT model's capacity for accurately classifying fake news, comparing its results against the dataset's ground truth. The utilization of these metrics provides insight into the model's performance and its potential for fake news detection.

5. Results

The model `text-davinci-003` was used in an initial attempt. The results were not encouraging as the model achieved an accuracy of only 48.66%. The model classified all texts as true news, even though half of them were false. The low accuracy suggests poor performance in correctly identifying fake news. After the discouraging result, we switched to the `text-davinci-002` model to evaluate its performance. The model was questioned with the following prompt along with the news text **Question 1: "Does the given text is fake news? Does it Spread misinformation? Answer only with yes or no."**. The results revealed an improvement compared to the previous model. Figure 1(a) depicts the obtained results for Question 1.



(a) Question 1. (b) Question 2.

Figura 1. Comparison of Performance Metrics in Fake News Detection.

These results indicate relatively good precision (93.8%), suggesting that when the model classified a text as fake news, it was likely correct. However, the low values of recall (30%), accuracy (64%), and F1-Score (45.5%) reveal the model's difficulty in correctly identifying a significant number of fake news cases. Although there was an improvement in accuracy compared to the first model, the overall performance did not meet our expectations. Therefore, we decided to run the `text-davinci-002` model again, but with a different prompt: **Question 2: "Does the given text contain characteristics of fake news? Does it spread misinformation? Answer only with yes or no."**. Figure 1(b) shows the result when applied to Question 2.

Comparing the results presented in Figures 1 and 2, it is possible to state that the model performed better when the question focused on the presence of characteristics of fake news. In general, there were improvements in accuracy (72%), recall (56%), and F1-Score (66.6%), indicating a more reliable detection of fake news compared to the first question. Precision reached 82.3%. These results demonstrate that the current solution is dependent on the question formulation. To accurately detect fake news, further improvements are necessary. In particular, as the GPT-3.5 model does not have a dedicated API for text classification, the proposed method in this work yielded promising but suboptimal results. This approach was an essential first step, highlighting both the potential and the limitations of current technology and future works on the implementation of a specific API for text classification to enhance future outcomes.

6. Conclusion and Future Work

In this work, we investigated the performance of the GPT-3.5 model in detecting fake news. Through an experimental analysis of a set of news articles, both true and false, we found that the way the prompt is formulated and presented to the GPT API significantly influences the quality and accuracy of the responses. Specifically, when requesting the API to identify if the provided text contained characteristics of fake news, a significant improvement in the results was observed. This suggests that directing the question in a more specific and focused manner on the characteristics of fake news enhances the model's effectiveness in detecting this type of content.

Our preliminary results showed that two different models of ChatGPT, named `text-davinci-003` and `text-davinci-002`, were effective in detecting fake news. Therefore, our findings suggest that GPT-3.5 has the potential for accurately classifying fake news when prompted with specific questions about its characteristics. Future research could explore how this model can be further optimized for detecting fake news on social media platforms and other online sources where misinformation is prevalent. Additionally, it would be interesting to investigate how other language models compare to GPT-3.5 in terms of their effectiveness in detecting fake news. Finally, we intend to consider a broader database, in addition to testing the system in different languages.

Referências

Aslam, N., Ullah Khan, I., Alotaibi, F. S., Aldaej, L. A., and Aldubaikil, A. K. (2021). Fake detect: A deep learning ensemble model for fake news detection. *Complexity*.

Baarir, N. F. and Djeffal, A. (2021). Fake news detection using machine learning. *2020 Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*.

Government, A. (2023). Is it true? Can COVID-19 vaccines alter my DNA? `https://www.health.gov.au/our-work/covid-19-vaccines/is-it-true/is-it-true-can-covid-19-vaccines-alter-my-dna`. Acessed: August 11, 2023.

Khivasara, Y., Khare, Y., and Bhadane, T. (2020). Fake news detection system using web-extension. *2020 IEEE Pune Section International Conference*.

OpenAI (2023). Models - OpenAI API. Disponível em: `https://platform.openai.com/docs/models/gpt-3-5`. Acessed: August 11, 2023.

Raza, S. and Ding, C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *Int. Jrnl. of Data Science and Analytics*.

Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., and de Figueiredo Nicolete, L. D. (2021). The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. *Journal of Public Health*.

Tribunal Superior Eleitoral (2022). Pílulas contra a desinformação: notícias falsas circulam 70% mais rápido do que as verdadeiras. Disponível em: `https://bit.ly/43DYRmy`. Acessed: August 11, 2023.

Özbay, F. and Alatas, B. (2019). A novel approach for detection of fake news on social media using metaheuristic optimization algorithms. *Elektronika ir Elektrotechnika*.