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

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

[1]Universidade Federal de Uberlândia (UFU)

[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.

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

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

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

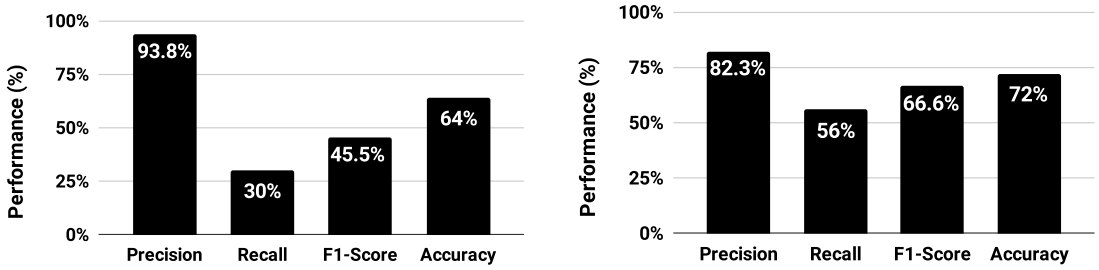
$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

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*.