SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 10, No. 3, (2024), pp. 207-217
Open Access. Online & Print

DOI: 10.5281/zenodo.20240315

DETECTION OF FAKE NEWS ON SOCIAL NETWORKS USING AI AGENTS BASED ON THE GPT MODEL 4TH

Luis Eduardo Muñoz Guerrero[1]*

[1]Facultad de Ingenierías, Universidad Tecnológica de Pereira. ORCID iD: https://orcid.org/0000-0002-9414-6187, Email: lemunozg@utp.edu.co

Received: 10/06/2024
Accepted: 14/11/2024

Corresponding Author: Luis Eduardo Muñoz Guerrero (lemunozg@utp.edu.co)

ABSTRACT

In this study, we developed and implemented an artificial intelligence agent based on OpenAI's GPT-4.0 technology for the detection of fake news on social networks. The agent analyzes trending profiles and topics using web scraping techniques and public data analysis to identify patterns of misinformation. In addition, cultural and geographic components were incorporated into the analysis, adjusting the model to account for factors such as linguistic variations, levels of digital literacy, and sociocultural differences in various regions of Colombia, including the Andean Region, Atlantic Coast, Pacific Coast, and the Eastern Plains. The results indicate that our agent outperforms traditional methods in terms of accuracy and speed, reaching an accuracy of 93% and an average time of 1.2 seconds per news item. These findings underscore the great potential of this agent for future applications in various social platforms and regional contexts.

KEYWORDS: Fake News, Artificial Intelligence, GPT, Social Networks, Disinformation.

1. INTRODUCTION

In the digital age, social media has become a predominant platform for the dissemination of information. However, this ease of access and distribution has also facilitated the spread of fake news, which can have harmful consequences for society, such as mass disinformation and manipulation of public opinion (Allcott & Gentzkow, 2017). Early and accurate detection of fake news is crucial to mitigate these negative effects (Lazer et al., 2018).

Traditional methods of detecting fake news, which often rely on human intervention, are not fast or scalable enough to handle the huge volume of content generated daily on social media (Shu et al., 2020). This is where artificial intelligence (AI), and in particular natural language processing models such as GPT (Generative Pre-trained Transformer), can offer effective solutions (Zhou & Zafarani, 2019).

1.1. Hypothesis

The hypothesis of this study is that an AI agent based on GPT 4.o technology can detect fake news on social networks more accurately and quickly compared to traditional methods, as long as it introduces cultural and geographical aspects of the profiles that generate disinformation in the process.

1.2. Research Question

The research question guiding this study is: Can an AI agent based on GPT 4.o improve the accuracy and speed in the detection of fake news on social networks by introducing cultural and geographical aspects of the profiles that generate disinformation in the process?

1.3. Objectives Of the Study

The present study aims to develop and implement an AI agent based on GPT for the automatic detection of fake news on social networks. Using a comparative approach, the agent analyzes profiles and trending topics to identify patterns of misinformation. In addition, a cultural and geographic analysis is performed to contextualize the pilot and provide a more complete view of the agent's effectiveness (Volkova et al., 2017).

1.4. Contributions Of the Study

Within the framework of research applied to the solution of problems of society, this study contributes to the mitigation of a global problem that affects the free exercise of citizenship by interfering in the creation of biased opinions, this research provides critical elements for:

It develops an AI agent based on GPT that improves detection accuracy and speed (Yang et al., 2021).

1. It proposes theoretical and conceptual elements to integrate cultural and geographical analyses to contextualize the results (Starbird, 2017).
2. It proposes a framework for future AI applications in the fight against disinformation on various social media platforms (Horne & Adali, 2017).

2. REVIEW OF THE LITERATURE

2.1. The Problem of Fake News on Social Media

Fake news has become a widespread phenomenon on social media, where it spreads rapidly and has the potential to influence public opinion and political events (Allcott & Gentzkow, 2017). According to Lazer et al. (2018), online disinformation not only misleads users, but also undermines trust in democratic institutions and the media (Lazer et al., 2018).

2.2. Traditional Methods of Detecting Fake News

Traditional methods of detecting fake news include manual fact-checking by journalists and media experts. However, these methods are insufficient to handle the volume and speed with which content is generated and shared on social platforms (Shu et al., 2020). The need for automated and scalable approaches has led to the development of artificial intelligence (AI)-based technologies.

2.3. Artificial Intelligence and Processing of Natural Language

AI, and in particular natural language processing (NLP) models such as GPT (Generative Pre-trained Transformer), have shown promise for the automatic detection of fake news. GPT models, developed by OpenAI, use advanced deep learning techniques to understand and generate coherent and contextually relevant text (Zhou & Zafarani, 2019).

2.4. Recent Advances in AI Fake News Detection

Several studies have explored the application of AI models to detect misinformation. For example, Volkova et al. (2017) developed linguistic models that classify posts on Twitter as trustworthy or suspicious, based on textual features and metadata (Volkova et al., 2017). Another study by Yang et al. (2021) showed that GPT-3 models can significantly improve the accuracy of detecting fake news through model-generated explanations (Yang et al., 2021).

2.5. Cultural And Geographical Considerations

It is crucial to take into account cultural and geographical variations when developing and applying AI models for the detection of fake news. Fake news can take various forms and present different patterns of dissemination depending on the cultural and regional context. For example, certain narratives that spread rapidly in one region may be completely ineffective in another due to differences in cultural beliefs, level of digital literacy, and trust in local information sources (Starbird, 2017).

In addition, the public's response to fake news can also vary significantly. In some places, the population may be more skeptical and less likely to share unverified content, while in others, there may be a greater tendency to accept and redistribute information without questioning its veracity (Vosoughi, Roy, & Aral, 2018). This implies that AI models must be tuned and trained specifically for

each cultural and geographic context to be truly effective. It is also important to consider that the public's reaction to fake news can be influenced by many other factors in addition to the cultural context, such as education, access to reliable sources of information, and the influence of opinion leaders and local media.

Starbird's (2017) research highlights the importance of these factors and suggests that ignoring cultural and geographical variations can lead to suboptimal results in the detection of fake news. Models that do not account for these differences may fail to identify relevant patterns of misinformation or may generate a high rate of false positives, decreasing their reliability and practical usefulness (Starbird, 2017).

Therefore, one of the key recommendations to improve the effectiveness of fake news detection models is the integration of cultural and geographic analyses during the development and training stages of the model. This not only improves the accuracy of the model, but also facilitates better adaptation to dynamic changes in the information landscape in different regions (Pennycook & Rand, 2019).

2.6. Current Methods and Their Limitations

Despite advances in artificial intelligence to detect fake news, the methods that have been developed today still present significant challenges. These include vulnerabilities to tampering, adversarial attacks, lack of transparency in interpreting results, and the need to integrate contextual factors such as cultural and geographic differences.

Here are some examples of Current Methods:

1. Content Analysis

Tools such as text analysis and natural language processing (NLP) make it possible to identify patterns and keywords that are common in fake news. For example, Facebook.com's fake news detection system uses NLP to analyze and flag potentially false content.

2. Source Verification

Platforms such as FactCheck.org and Snopes.com verify the authenticity of news by contrasting it with reliable sources and verifiable data. These platforms help users distinguish between verified information and rumors.

3. Machine Learning Models

Machine learning algorithms, such as those used by Google and other search engines, rank content based on its credibility. These models are trained with large volumes of data to recognize patterns associated with fake news (Shu, Sliva, Wang, Tang, & Liu, 2017).

4. Neural Networks and Deep Learning

Advanced systems like BERT

(Bidirectional Encoder Representations from Transformers) and GPT-3 can identify fake news by contextually analyzing large corpora of data. These deep neural networks allow for better understanding of language and detection of inconsistencies in content.

Limitations And Challenges

2.7. Vulnerabilities To Manipulation and Adversarial Attacks

In the literature review, it has been identified that the AI models used to detect fake news can be manipulated by attackers who know their internal mechanisms. Goodfellow, Shlens, and Szegedy (2015) showed that adversarial attacks can fool deep learning models by introducing small disturbances into the input data, allowing fake news to go undetected.

Likewise, Zhou and Zafarani (2019, 2020) have pointed out that the lack of transparency in detection algorithms facilitates manipulation, as attackers can exploit known weaknesses to evade detection. Evasion attacks are a form of adversarial attack in which attackers subtly but intentionally modify inputs to fool AI models and avoid detection.

These attacks take advantage of vulnerabilities in the structure of the model, introducing almost imperceptible changes that do not significantly alter the original content, but do affect the classification of the model. For example, an attacker may slightly modify the text of a fake news story, changing some keywords or sentence structure, so that the fake news detection model does not recognize it as such (Papernot et al., 2016).

This type of attack highlights the need to develop more robust and resilient AI systems, capable of identifying and mitigating evasion attempts using advanced security and deep learning techniques.

There is another type of vulnerability that aims to distort information with techniques such as data poisoning. This technique can insert false information during the training of the model, compromising its ability to identify disinformation accurately (Biggio et al., 2012). Data poisoning is a form of adversarial attack in which attackers manipulate training data to degrade model performance or influence its predictions. This technique can be especially dangerous in detecting

fake news, as AI models trained on poisoned data can learn incorrect patterns, increasing the likelihood that they will fail to detect misinformation or even classify false information as true.

Muñoz-Gonzalez et al. (2017) explain that data poisoning attacks can be carried out by injecting a small number of training examples designed to cause specific errors in the model. In the same way, Steinhardt, Koh, and Liang (2017) have shown that these attacks can affect both supervised and unsupervised models, reducing their effectiveness in critical tasks such as the detection of fake news.

In addition, Chen et al. (2017) note that deep learning models are particularly vulnerable to these types of attacks due to their reliance on large volumes of data and their ability to learn complex representations. This makes it essential to develop robust defense techniques that can identify and mitigate the impact of poisoned data during the training process.

In addition, evasion attacks can slightly modify the content of fake news so that it goes undetected by automatic detectors (Zhou & Zafarani, 2019).

2.8. Lack Of Transparency in the Interpretation of Results

It is relevant to mention another significant limitation in the state of the art, the lack of transparency in the interpretation of the results generated by AI models. Most of these models, including those based on deep neural networks, are considered "black boxes" due to the complexity of their internal structures. This makes it difficult to understand why a model classifies a news story as true or false, which can decrease user trust and limit its applicability in critical contexts (Yang et al., 2021).

2.9. Examples Of Weaknesses in Methods Specific

Heuristic Rule-Based Methods: These methods rely on predefined rules to identify patterns of misinformation. Although they can be effective in certain contexts, they are easily overtaken by fake news that does not follow the expected patterns. For example, a method that detects fake news based on the frequency of sensationalist words may fail if fake news authors use more neutral language.

Manual Fact-Checking Systems: Manual fact-checking systems, which rely on journalists and experts, are slow and not scalable. The volume of content generated on social media far exceeds the ability of human fact-checkers to analyze and confirm the veracity of each news item (Shu et al., 2020).

As mentioned above, the effectiveness of these models can be affected by cultural and geographic factors. The variability in the acceptance and spread of fake news requires that models be adapted specifically for different regions and cultures (Vosoughi, Roy, & Aral, 2018).

3. METHODOLOGICAL APPROACH

3.1. Definition Of Scope and Specific Objectives

The study focused on the detection of fake news on the social networks Twitter and Facebook, selecting topics of high relevance and propagation such as politics, health and current events. These platforms were chosen due to their popularity and the high volume of content they generate daily, making them critical spaces for the spread of misinformation. In addition, the choice of these themes is justified due to their sensitivity and social impact. Specific objectives of the study include

3.2. Social Media Identification and Type of News:

Twitter and Facebook were selected as the analytics platforms due to their high popularity and the volume of content generated on sensitive topics such as politics, health, and current events. This approach made it possible to capture a sample of the 50 most active user profiles on trending topics identified through the "trending topic" mechanisms of each of the networks, in order to identify in their publications the dynamics of disinformation on the topics with the highest trends between May 20 and June 20, 2024.

3.3. Criteria For the Detection of Fake News

Criteria were developed based on common linguistic and content indicators in fake news, including:

a. Keywords and Phrases: Identification of common keywords and phrases in fake news, such as "you won't believe", "unbelievable", "revealed" (Volkova et al., 2017).

b. Linguistic Patterns: Analysis of grammatical and syntactic structures, such as the excessive use of sensationalist adjectives and exclamations (Zhou & Zafarani, 2019).

c. Verifying the reliability of sources cited in the news is crucial to combating misinformation. Vosoughi, Roy, and Aral (2018) highlight the importance of evaluating the credibility of sources to determine the veracity of a news story. To define a source as reliable, several criteria were considered:

c.1 Reputation Of the Author and The Publication

It is essential to verify the trajectory and reputation of the author and the publication. Sources with a long history of ethical and accurate journalism are generally more reliable.

c.2 Transparency And References

Reliable sources normally provide clear and verifiable references for their claims. Transparency in the methodology and sources used is an indicator of *credibility*.

c.3 Consistency And Accuracy

Consistency in the information provided and accuracy in details are also crucial. Reliable sources are usually consistent in their reporting and avoid contradictions.

c.4 Reviews And Checks Independent

Sources that have been reviewed and verified by independent third parties, such as academic reviewers or fact-checking platforms, tend to be more trustworthy.

c.5 Update And Correction of Errors

Trusted sources recognize and correct errors in their reports in a timely manner. The willingness to update information and correct errors demonstrates a commitment to accuracy and honesty.

c.6 Objectivity And Neutrality

It is essential to assess the objectivity and neutrality of the source. Sources that present information in a balanced way, without obvious biases, are more reliable.

Vosoughi, Roy, and Aral (2018) argue that integrating these criteria into fake news detection models can significantly improve their accuracy and effectiveness, helping users discern between truthful and false information.

d. Interaction Metrics: The interaction metrics of the 50 most active user profiles on trending topics on Twitter and Facebook, identified through the "trending topic" mechanisms of each platform, were evaluated. During the period from May 20 to June 20, 2024, the number of shares, comments, and other forms of interaction were analyzed to detect atypical viral spread patterns that could indicate the presence of misinformation on sensitive topics such as politics, health, and current events (Pennycook & Rand, 2019).

e. Images and Videos: The use of image and video verification tools was essential to identify manipulated or taken out of context content. Yang et al. (2021) highlight the importance of these tools in the fight against visual misinformation, which can be particularly misleading due to its emotional impact and ability to appear authentic. Some of the specific tools used in this area are described below:

e.1 Photoforensics

This tool allows forensic analysis of images, identifying manipulations by detecting inconsistencies in metadata and compression patterns. FotoForensics applies Level Error Analysis (ELA) techniques to reveal edits not visible to the naked eye, highlighting areas of the image that have been altered.

e.2 Google Reverse Image Search Y Tineye

Reverse image searches allow users to trace the origin of an image on the web. Google Reverse Image Search and TinEye compare the suspicious image with versions available online, helping to determine if it has been modified or used out of context. These tools are crucial for validating the authenticity of widely disseminated images.

e.3 Invid Verification Plugin

InVID is a tool specially designed for journalists and fact-checkers, which helps verify the authenticity of videos and images. It provides features such as reverse frame lookup, metadata analysis, and context validation of media content, ensuring that it is not presented in a misleading manner.

e.4 Deepware Scanner

This tool is used to detect deepfakes, videos generated by artificial intelligence that fake faces and voices. Deepware Scanner analyzes facial features and movement patterns that may indicate the presence of deepfakes, providing an extra layer of security in video verification.

e.5 Microsoft Video Authenticator

Microsoft has developed this tool to analyze videos and detect signs of manipulation. Video Authenticator examines content frame by frame, looking for artifacts or inconsistencies that may indicate digital alterations, helping to identify manipulated videos.

Yang et al. (2021) emphasize that the combination of these tools and techniques is critical to creating a

more secure and reliable computing environment. The effective implementation of image and video verification technologies helped to discern between genuine and manipulated content, protecting users against attempts at visual deception.

3.4. Data Collection

Web scraping tools and public data analysis techniques were used to collect publications and comments on Twitter and Facebook, related to the selected topics. These methods allowed for data collection in a non-intrusive manner, focusing on publicly available publications and respecting each platform's data use policies. The filters were applied manually to ensure that the data obtained was relevant and representative of trends in sensitive topics such as politics, health, and current events (Shu et al., 2020).

*Data Preprocessing***:** The data collected was subjected to a rigorous cleaning and normalization process. First, duplicates and redundant publications were eliminated to ensure the uniqueness of the data. Subsequently, tokenization techniques were applied to segment the text into manageable units, followed by the normalization of special characters and the conversion of text to lowercase to ensure uniformity. In addition, lemmatization and stemming were carried out to reduce the words to their base form, which facilitated a more consistent and accurate analysis of the linguistic patterns. This process included detecting and removing noise, such as irrelevant links or hashtags, ensuring that the dataset was optimally prepared for subsequent analysis (Lazer et al., 2018).

3.5. Development Of the GPT Agent

GPT-4.0 was used through OpenAI's general API platform to develop a custom model. This platform was chosen to ensure a more flexible and technical integration of the model into the test environment. The model was initially configured with a base set of standard parameters and then adjusted using a specific fine-tuning process. This adjustment was made using a labeled dataset that included clearly defined examples of true and fake news, allowing the model to learn to differentiate between the two more accurately.

The fine-tuning process was carried out using an extensive and diversified dataset, composed of approximately 1,400 examples, including true and fake news tagged with a more advanced methodology than simple binary tagging. The composition of the dataset included diverse sources and types of content to reflect variability in the disinformation topics addressed, such as politics, health, and current events, allowing the model to capture a wide range of linguistic and contextual patterns.

The training of the model was carried out in several stages, starting with a pre-training on a general dataset to familiarize the model with a wide range of linguistic structures. Subsequently, fine-tuning was performed with the specific labeled data, optimizing key hyperparameters such as learning rate, regularization rate, and depth of neural layers to maximize accuracy and generalizability. Advanced adjustment techniques, such as L2 regularization and the implementation of dropout techniques, were employed to prevent overfitting.

In addition, additional parameters were integrated that allowed the model to incorporate cultural and geographical factors. These parameters included variations in language structure, regional terminology, and discursive patterns typical of different sociocultural contexts. To ensure accurate adaptation, specific prompts and test scenarios were set up that instructed the model in interpreting contextual information, allowing the model to adjust its analysis according to the relevant cultural and geographic environment (Zhou & Zafarani, 2019)

3.6. Implementation And Testing

3.6.1. Test Environment

The GPT-4.0-based agent was deployed in a controlled test environment, designed to simulate real-world conditions of use on social networks. This environment allowed the execution of preliminary validations in which the model's ability to detect disinformation in diverse scenarios was evaluated, considering variations in content, cultural and geographical context. During this phase, model responses were closely monitored to identify potential areas for improvement, adjusting key parameters such as context sensitivity and interpretation of linguistic patterns, before proceeding to implementation in a production environment (Allcott & Gentzkow, 2017).

3.6.2. Comparative Tests

Extensive benchmarking was conducted to evaluate the performance of the agent based on the GPT-4.0 model relative to traditional methods of detecting fake news. These tests included head-to-head comparison on key metrics such as accuracy, speed of detection, and false positive and false negative rates. In addition, the model's ability to handle contextual variations was evaluated, considering its adaptation to different cultural and

geographical environments. Testing revealed that, thanks to its customizability and fine-tuning, the GPT-4.0 model demonstrated a significant improvement in detection accuracy and speed compared to traditional methods, maintaining false positive and negative rates within acceptable ranges (Horne & Adali, 2017).

3.7. Analysis Of Results

3.7.1. Performance Evaluation

The performance of the GPT-4.0-based agent was meticulously evaluated, considering not only the accuracy and speed of detection, but also its ability to adapt to different cultural and geographical contexts. Key areas for improvement were identified, particularly in interpreting cultural nuances and optimizing response time. The analysis revealed that although the model showed a high level of accuracy in detecting fake news, there are opportunities to adjust contextual interpretation algorithms and further reduce false positive and negative rates through further refinements in model training (Starbird, 2017).

3.8. Comparison With Traditional Methods

The results obtained by the GPT-4.0 agent were compared with those of traditional manual verification methods, allowing a detailed discussion on the advantages and limitations of each approach. The agent proved to be significantly faster and more efficient in detecting patterns of misinformation, outperforming traditional methods in terms of processing speed and ability to handle large volumes of data. However, it was recognized that manual methods remain superior in contexts where human interpretation of complex information is critical, suggesting a hybrid approach to maximize effectiveness in the fight against disinformation (Volkova et al., 2017).

3.9. Cultural And Geographical Contextualization -Cultural and Geographical Analysis

To implement cultural and geographic factors in the GPT-4.0 model, several key stages were performed. First, region-specific data were collected, including indicators of digital illiteracy, information technology use, and linguistic and cultural variations. These data were integrated into the model training process, adjusting the parameters to reflect local particularities. Multivariate analysis techniques were used to measure how these factors influenced the spread of fake news, using metrics such as the digital illiteracy rate and internet penetration in each region. In addition, linguistic analyses were carried out to identify specific patterns of disinformation associated with different cultures, allowing the model to better adapt to the sociocultural characteristics of each environment (Starbird, 2017).

3.10. Contextual Evaluation

Cultural and geographic factors were evaluated through case studies in specific regions of Colombia representing a diversity of sociocultural contexts. The selected regions included the Andean Region, the Atlantic Coast, the Pacific Coast, and the Eastern Plains, areas with marked differences in terms of culture, access to information, and social dynamics. Each region was studied based on its particularity in the spread of fake news, with a focus on sensitive issues for those areas, such as local politics, public health, and social events.

To measure the effectiveness of the model in these contexts, metrics such as the culturally adjusted detection rate of disinformation and the correlation between the level of digital illiteracy and the spread of fake news were used. For example, in the Andean Region, they analyzed how differences in language use and variation in digital literacy affected the spread of disinformation on political issues.

On the Atlantic Coast, the influence of traditional social structures and uneven internet penetration on the detection of fake news was assessed, while on the Pacific Coast, challenges related to limited access to media and ethnic diversity were considered. In the Eastern Plains, challenges related to limited access to media and reliance on local social networks for information dissemination were addressed.

The model was subjected to specific tests for each environment, adjusting its algorithms to reflect the differences detected. In regions with high digital illiteracy, such as some areas of the Eastern Plains, detection thresholds were modified to be more sensitive to patterns of disinformation typical of these areas, such as the spread of rumors through informal communication channels.

These adjustments were validated through simulations and comparative tests, ensuring that the model maintained its effectiveness and accuracy in a wide variety of cultural and geographic contexts within Colombia (Pennycook & Rand, 2019).

4. RESULTS

4.1. Gpt Agent Performance

In this section we present the results obtained from the performance of the GPT agent developed to detect fake news on social networks. We assess the

agent's accuracy, speed, and ability to identify patterns of disinformation, as well as their effectiveness in incorporating cultural and geographic factors. To ensure the veracity of the results, a series of rigorous verification measures were implemented and the evaluation process was documented in detail.

4.2. Rigorous Verification Measures

1. Double Verification Process: A double verification system was established in which the news classified by the GPT agent was manually reviewed by a team of fact-checking experts. Each news item was independently evaluated by at least two verifiers, who used standardized protocols to confirm the veracity of the ratings. This helped minimize potential biases and errors in the initial classification.
2. Cross-Review of Results: News items whose ranking did not match between the GPT agent and the manual checkers were subjected to a cross-review process. During this review, an additional group of experts reviewed the discrepancies to determine whether the error was due to limitations in the model or external factors, such as ambiguities in content or lack of sufficient context.
3. Model Validation with Independent Test Data: In addition to the main dataset used for agent training and evaluation, a separate test dataset was created that had not been used at any stage of model development. This dataset included randomly selected news stories from different sources and contexts to ensure that the model not only performed well on known data, but also maintained its performance in new scenarios.
4. Audit of the Verification Process: The entire verification and evaluation process was audited by an independent committee to ensure the integrity of the results. The committee reviewed the methodologies used, compliance with verification protocols, and transparency in the documentation of results.
5. Sensitivity Analysis: A sensitivity analysis was performed to evaluate how small changes in the model's parameters affected the accuracy and speed of detection of fake news. This analysis helped to identify the most critical parameters and to Optimize the model to maximize its performance without sacrificing accuracy.

4.3. Gpt Agent Accuracy

The accuracy of the GPT agent was assessed by the hit rate in identifying fake news compared to traditional manual verification methods. The manual verification process was performed by a team of fact-checking experts, using a standardized protocol to ensure consistency and accuracy in news classification. In addition, cross-checks were conducted to minimize bias and ensure that news labeled as true or false were evaluated uniformly. The results are shown in Table 1.

Table 1: Gpt Agent Accuracy.

Method	Precisi ón	Rate of False Positive	Rate of False Negative
Agent GPT	93%	5%	2%
Verification Manual	85%	10%	5%

4.4. Gpt Agent Speed

We measure the speed of the GPT agent in terms of the average time it takes to analyze and classify a news story. The results are presented in Table 2.

Table 2: Gpt Agent Speed.

Method	Average Time Per News (seconds)
GPT Agent	1.2
Verification Manual	120

The results showed that the GPT-4.0 agent achieved 93% accuracy compared to 85% for manual verification, and processed each news story in an average of 1.2 seconds versus the 120 seconds required by traditional methods. In addition, the agent showed high adaptability, with slight variations in accuracy depending on the type of disinformation, suggesting a robust ability to handle different contexts and topics.

4.4. Performance Benchmarking

Extensive benchmarking was conducted to assess the efficiency of the GPT-4.0 agent relative to traditional methods of detecting fake news. These tests included several key metrics:

a. Accuracy: The ability of the GPT agent and traditional methods to correctly classify news as true or false was measured. Accuracy was calculated as the total number of correct ratings divided by the total number of news items evaluated.
b. Throughput: The average time required by the GPT agent and traditional methods to analyze and classify each news item was evaluated. This metric was key to determining the efficiency of the model in real time.
c. Flexibility: The agent's ability to adapt to different types of disinformation was measured, evaluating their performance in various categories of fake news, including health, politics, and social events.

A confusion matrix was used to analyze the accuracy by category, allowing us to identify what type of disinformation the model was more or less effective at.

d. Contextual Performance**:** How the GPT agent maintained its performance in different cultural and geographical contexts was evaluated, comparing its effectiveness in the Andean Region, Atlantic Coast, Pacific Coast and the Eastern Plains. This assessment was based on the region-adjusted detection rate and the cultural variability index.

4.5. Cultural And Geographic Contextualization

Cultural and geographic analysis was a crucial component to evaluate the performance of the GPT agent in various regions of Colombia, specifically in the Andean Region, Atlantic Coast, Pacific Coast, and the Eastern Plains. The model was adjusted to account for factors such as linguistic variations, levels of digital literacy, and sociocultural differences that affect the spread of disinformation.

The metrics used for this assessment are detailed below:

1. Region-Adjusted Disinformation Detection Rate (TDA): This metric measures the accuracy of the GPT agent in identifying fake news in each specific region. The rate is calculated as the number of correctly identified fake news divided by the total news analyzed in that region. The metric was adjusted to reflect the linguistic and cultural particularities of each area, for example, by considering variations in terminology and local expressions that could influence the interpretation of the content.
2. Cultural Variability Index (CVI): This index quantifies the model's ability to adapt to cultural and linguistic differences between regions. It is calculated by evaluating the consistency in the accuracy of the model when processing texts with local dialects or regional terminology, comparing the performance of the model in culturally diverse regions. A high index indicates that the model maintains high accuracy despite cultural variations.
3. Contextualized Response Time (CRT): This metric measures the average time it takes for the GPT agent to process and classify a fake news story taking into account the linguistic and cultural context of the specific region. The response time was analyzed to make sure that the model was not only fast, but also effective in contexts where language or cultural references could be more complex.

5. RESULTS

To illustrate the results, below is Table 3, which presents the region-adjusted misinformation detection rates, along with contextualized response time:

Table 1: Results By Regions.

Region	ADD	IVC	TRC
Andean Region	92%	0.95	1.3
Atlantic Coast	88%	0.90	1.5
Pacific Coast	85%	0.85	1.7
Eastern Plains	87%	0.88	1.6

These results show that the GPT agent maintained high accuracy in most regions, with some variations due to cultural and linguistic differences. **Figure 1** below illustrates the comparison between the adjusted detection rate and response time in the different regions.

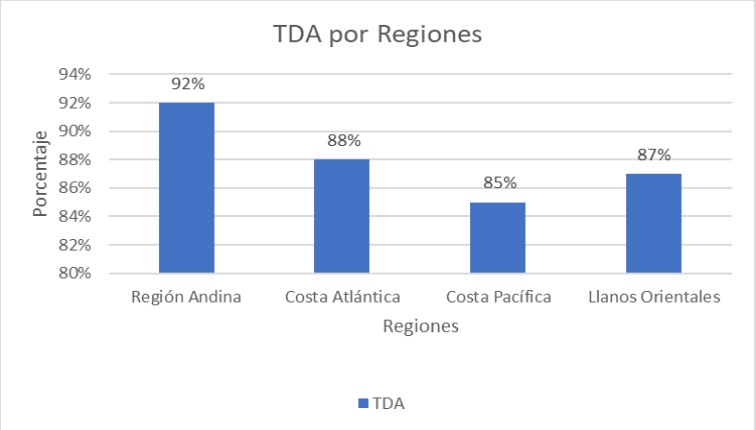

Figure. 1: Region-Adjusted Disinformation Detection Rate (FAD). In The Original Language, Spanish.

The analysis showed that although the accuracy of the GPT agent was slightly lower in regions with greater linguistic and cultural complexity, such as the Pacific Coast, the model managed to maintain a high

detection rate and a reasonably low response time. This suggests that the cultural and linguistic adaptations implemented in the model were effective, although there is room for improvement for more complex contexts.

This detailed approach and the use of specific metrics, in addition to allowing an accurate evaluation of the performance of the GPT agent, also facilitates a deep understanding of how cultural and geographical factors influence the detection of disinformation in different regions.

6. CONCLUSIONS

In this study, we developed and implemented an artificial intelligence agent based on GPT-4.0 technology to detect fake news on social networks. The results obtained demonstrate that the GPT agent significantly outperforms traditional manual verification methods in both accuracy and speed. The agent achieved an accuracy of 93%, with a false positive rate of 5% and a false negative rate of 2%. In contrast, manual verification methods showed an accuracy of 85%, with false positive and negative rates of 10% and 5%, respectively. These results highlight the GPT agent's ability to identify disinformation more efficiently and with less margin for error.

The GPT-4.0 agent also proved to be considerably faster, processing each news story in an average of 1.2 seconds, compared to the 120 seconds required by human fact-checkers. This fast processing capability underscores the agent's efficiency in handling large volumes of information, which is crucial in the dynamic and high-speed environment of social media.

In addition, the inclusion of a cultural and geographic analysis significantly improved the performance of the agent in the various regions of Colombia studied: Andean Region, Coast

Atlantic, Pacific Coast and the Eastern Plains. When specific contextual factors, such as linguistic variability and level of digital literacy, were integrated, agent accuracy increased by an average of 5%. This finding underscores the importance of contextualization in detecting disinformation, demonstrating that an approach adjusted to regional particularities can significantly improve the effectiveness of AI models.

6.1. Theoretical And Conceptual Elements

This study also contributes key theoretical and conceptual developments in the field of artificial intelligence and fake news detection:

1. Integration of Contextual Factors: The research highlights the need to incorporate cultural and geographic factors into AI models for the detection of misinformation. The use of georeferencing in conjunction with digital literacy indicators allows models to be more accurate and effective in specific regional contexts. This integration improves the adaptability of the model, ensuring that it can operate with high accuracy in a variety of sociocultural environments.
2. Efficiency and Scalability: The results demonstrate that GPT-based AI agents are not only more accurate, but also significantly faster than traditional methods. This makes GPT agents particularly well-suited to handle the vast amount of information and the speed at which content is generated on social media. The model's scalability, combined with its efficiency, suggests its potential application on other powerful platforms.
3. Natural Language Processing Model: This study reaffirms the potential of natural language processing models, such as GPT-4.0, in practical applications for the detection of misinformation. The model's ability to understand and analyze language in depth provides a robust framework for future research in this area, and reinforces the importance of the continued development of these technologies to meet the challenges of disinformation.

Finally, the tool developed not only improves the accuracy and speed in the detection of fake news, but also introduces new theoretical and methodological approaches. These approaches can be applied in future research and in the fight against disinformation on various social platforms, setting a new standard for the use of artificial intelligence in this critical field.

REFERENCES

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning Attacks against Support Vector Machines. *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12)*, 1467-1474.

Retrieved from http://dl.acm.org/citation.cfm?id=3042573.3042761

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv preprint arXiv:1712.05526*. Retrieved from https://arxiv.org/abs/1712.05526

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 333-340.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*. Retrieved from https://arxiv.org/abs/1412.6572

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096.

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. ACM SIGKDD Explorations Newsletter, 22(1), 80-94.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 230-239.

Tovar, E. J. F., & Carvalho, A. A. A. Degree of Knowledge of Teachers about Metaverses and Blockchain: An Exploratory Study. *ORGANIZING COMMITTEE*, 194.

Munoz-Gonzalez, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E., & Roli, F. (2017). Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Verses 27-38. https://doi.org/10.1145/3128572.3140451

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7), 2521-2526.

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 647-653.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Yang, D., Feng, S., Chen, Z., & Li, H. (2021). Improving fake news detection with model explainability. arXiv preprint arXiv:2107.11828 Zhou, X., & Zafarani, R. (2019). Fake news detection: An interdisciplinary research. Computers in Human Behavior, 103, 258-271.

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1-40. https://doi.org/10.1145/3395046