Comparative Analysis of Misinformation Detection: State-of-the-Art LLMs Are Not Readily Useable Fake News Classifiers

Anonymous ACL submission

Abstract

This study investigates the effectiveness of state-of-the-art Large Language Models (LLMs), including OpenAI's GPT-4o, GPT-4o-mini, and o1-mini, in identifying misinformation. We compare their performance against simpler text-based classifiers, namely a Naive Bayes classifier and a logistic regression model using text-derived features. Our evaluation utilizes two datasets: the ISOT Fake News dataset and the LIAR2 dataset, which respectively include political news articles and statements. Despite their advanced capabilities, our findings show that LLMs perform comparably to, or worse than, simple models like Naive Bayes or featured-based models in misinformation detection, especially with straightforward prompts. This research contributes to the ongoing debate on the role of LLMs in misinformation detection and highlights the need for alternative approaches in this domain.[1]

1 Introduction

Misinformation and fake news are regarded as a major challenge in our modern societies (Meel and Vishwakarma, 2020). Consequently, numerous researchers have tried to build increasingly complex computational models with the objective of accurately identifying misinformation without the need for human intervention. However, we propose that the layperson will select the most straightforward and convenient method for validating the content of a given news source. The omnipresence of Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022), in our daily lives may lead individuals to rely on these easily-accessible LLMs to verify the truth of information. Thus, we aimed to investigate whether state-of-the-art LLMs, specifically OpenAI's ChatGPT-4o , ChatGPT-4o-mini and o1-mini (OpenAI, 2024a,b,c), are capable of accurately classifying misinformation relative to simpler text-based classifiers.

In this paper, we explored the predictive capabilities of LLMs in a binary classification task. We compared the the performance of the LLMs against a simple Naive Bayes classifier based on word counts, as well as against a logistic regression classifier trained on hand-crafted features involving character-based patterns like punctuation and capitalization as well as word-level patterns like emotionality and lexical quality. We evaluated the models on two datasets: the ISOT Fake News (Ahmed et al., 2018) and the LIAR2 dataset (Xu and Kechadi, 2024), which is an extension of the widely used LIAR dataset (Wang, 2017). Despite both datasets containing solely English examples, the instances of the ISOT dataset consist of entire news articles, split into title and body, whereas the LIAR2 dataset consists of single statements spanning one or two sentences. We show that, when prompting LLMs to solve the task, we are not able to significantly outperform a simple feature classifier.

2 Related Work

The study of misinformation detection has been a prominent area of research in recent decades. With the advent of LLMs, researchers began utilizing these systems' language abilities for the detection of fake news. Previous studies of LLM applications in the context of misinformation detection may be broadly categorized into two approaches: simple prompting and the integration of LLMs into more sophisticated models.

Concerning the simple prompting approach, the predictive performance of LLMs is typically benchmarked against that of small language models (SLMs) or other existing models for the detection of fake news. Hu et al. (2024) analyzed the misinformation detection capabilities of GPT-3.5-turbo in zero- and few-shot conditions but assessed su-

[1]The code is made publicly available on GitHub: https://github.com/montmorency3/Fake-News-Detector

perior performance of a fine-tuned BERT model (Devlin et al., 2019), a widely used SLM. A comparative analysis was conducted by Caramancion (2023) between OpenAI's GPT models (-3.5 & -4), Google's Bard/LaMDA, and Microsoft's Bing AI based on their performance on news headlines, where GPT-4 not only outperformed GPT-3.5, but also the other two models. Pelrine et al. (2023) compared the performance of GPT-4 against several adaptations of BERT in a binary classification task on the LIAR dataset consisting of political statements (Wang, 2017), which saw GPT-4 outperforming all BERT-based models. This present study builds upon previous research by evaluating the latest OpenAI models, specifically GPT-4o, GPT-4o-mini, and o1-mini with simple prompting. Additionally, the performance was evaluated on political statements and complete news articles, consisting of title and body.

Other researchers investigated the use of LLMs for fake news detection just as one step in a more complex model. For example, some promising approaches involve leveraging LLMs to provide rationales for SLMs (Hu et al., 2024) or incorporating the LLM's classification as features to standard machine learning algorithms (Teo et al., 2024). However, due to the practical motivation behind this research, we leave the exploration of the effectiveness of OpenAI's GPT-4o and o1 series as model components open to future research.

3 Methods

3.1 Datasets

For the purpose of this study, the ISOT Fake News dataset (Ahmed et al., 2018) and the LIAR2 dataset (Xu and Kechadi, 2024) were investigated.

The ISOT dataset consists of political articles: truthful ones collected from Reuters.com as well as fake ones gathered from websites flagged as unreliable by Politifact. After cleaning and removing duplicates, a set of 38,205 unique articles each longer than 200 characters remained. Due to cost and API usage limitations, only 2.5 percent of the dataset (955 articles) were used as test set in the experiments.

The LIAR2 dataset consists of 22,962 political statements with an average length of 17.7 tokens per statement. Although statement truthfulness was categorized from 0 to 5, to aid comparisons to the ISOT dataset, we collapsed the six categories into 'true' (9,723 statements) and 'false' (13,239 statements).

3.2 Naive Bayes

As a baseline text-processing model, a Naive Bayes classifier was employed to benchmark the predictive power of the raw text data. For this task, the ISOT dataset was preprocessed by removing links, mentions, and potentially skewed indicators of article sources. The resulting vectorized word counts were fed into the sklearn Naive Bayes model to solve this binary classification task.

3.3 Simple Feature Extraction

Previous works have already addressed how text features can help improve the classification of documents (Kumar and Taylor, 2024). Specifically, one could aim to show whether certain features about written articles or statements could be used to correctly guide the classification of these text documents via a simple logistic regression model. To achieve this, several features were extracted under the premise of being representative of fake news articles at large.

To begin, character-based features were computed. The *punctuation* score was derived by counting the number of exclamation and interrogation points, with both types being more likely to be abused in fake articles. Similarly, the *tag* and *citation* scores respectively measured the number of tagged elements in an article (username handles and hashtags) as well as the number of quoted passages. Other metrics based on *capitalization* and *numeric* counts were computed by tallying the number of uppercase letters and digits in a text fragment. Although these features may not directly reflect the nature of fake against true articles, they are simple enough to extract that it would be wasteful to neglect them.

Next, word-based features were extracted. The text inputs were vectorized at a word level, with punctuation and symbol-likes stripped away, and the remaining tokens were lemmatized using the NLTK **WordNetLemmatizer**[2] function (Bird et al., 2009). From here, several measures of text complexity were derived using simple metrics. The first of these was a *uniqueness* score based on the proportion of unique words in the text instance. The assumption made here relies on the idea that real articles will be more carefully edited, and hence use

[2] https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=lemmatize

less redundant vocabulary. Furthermore, text quality was assessed with metrics for the proportion of *short*, *medium*, and *long* words present in the instance, where transition cutoffs were respectively 7 and 10 characters in length. This simple feature was used as a proxy measure for article quality, where articles with high counts of shorter words may come across as weaker from a lexical sense. Such length-based features were employed in a Tweet validity task using the recent Truth Seeker Dataset (Dadkhah et al., 2024).

In addition, a sentiment analysis approach was implemented to score the emotional tone of text instances. In general, there is a belief that false information is often embedded within inflammatory and reactionary language, whereas true articles tend to reflect emotional neutrality. As such, scoring the sentiment of articles seemed like an excellent feature that could assist the discrimination task at hand. Sentiment was scored under two distinctions: intensity and valence. *Intensity* represents the proportion of emotional words utilized in the text, and *valence* scores reflect the proportion of positively and negatively charged words in the vocabulary. The lexical database used for scoring was derived by Wilson et., al (2005) for the purpose of disambiguating word polarities in a context-dependent manner. Their lexical corpus consists of nearly 9,000 annotated examples scored using intensity (either strongly or weakly toned) and valence (either positive, negative, both, or neutral) (Wilson et al., 2005). This corpus was used to derive the sentiment scores within our target datasets.

Lastly, Named Entity Recognition (NER) was utilized to enhance the predictive power of the model. Named entities such as people, organizations, and locations were identified and were then categorized and counted. The idea here was to count the number of named entities that did not match a verified database of real-world entities to identify the presence of potentially fabricated or unreliable references, which would hence contribute to the classification of fake news articles (Tsai, 2023).

For the purpose of the ISOT dataset, features were extracted separately for the text body and the article title. Typically, one would expect fake articles to contain flashy and evocative titles, which would be reflected in saturated title scores. For the LIAR2 dataset, only the short statements were considered separately. To test the relevance of these features, the base logistic regression classifier[3] from the scikit-learn library was used (Pedregosa et al., 2011). Moreover, various permutations were employed to assess the individual contributions of each feature to the decision process.

3.4 Large Language Models

Given our aim to benchmark the predictive performance of state-of-the-art LLMs in misinformation detection, OpenAI's latest models, GPT-4o, GPT-4o-mini, and o1-mini (OpenAI, 2024a,b,c), were prompted to perform this binary classification task.

Models: Released in May 2024, GPT-4o tied, or even surpassed, results on a wide range of benchmarks. Despite the focus lying on multi-modal training, GPT-4o outperformed GPT-4 on most text-based benchmarks. In July 2024, OpenAI released the cost- and resource-efficient GPT-4o-mini, a small version of GPT-4o. Finally, in September 2024, OpenAI released the o1 series which focuses on complex reasoning. Due to the high cost of API usage, this work focuses on o1-mini, leaving the exploration of o1-preview and the final o1 model (when released) open for future research. Compared to GPT-4o, o1-mini has less world knowledge and performs worse in language-related tasks but higher tasks involving complex reasoning.

Model	Dataset	Invalid Answers
GPT-4o	ISOT	0.0178
	LIAR2	0.0
GPT-4o-mini	ISOT	0.0
	LIAR2	0.0
o1-mini	ISOT	0.0304
	LIAR2	0.1263

Table 1: Proportion of Invalid Responses per Model and Dataset

Prompting: To test the predictive capacity of the LLMs, we adapted the prompt used by Pelrine et al. (2023) which asks the LLM to evaluate the validity of a text input as either "0" (false) or "1" (true). Our modified version of the prompt was designed as a trade-off between modeling the behavior of a "typical" user and receiving unambiguous and readily analyzable results. For the ISOT news articles, we used the following prompt:

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

> "role": "user", "content": "I want you to rate the truthfulness of the following news article. Do not provide any explanations, just answer with 0 or 1, where 0 represents false and 1 represents true.\n\nTitle: TITLE \nText: TEXT"

For the LIAR2 statements, we used the prompt:

> "role": "user", "content": "I want you to rate the truthfulness of the following statement. Do not provide any explanations, just answer with 0 or 1, where 0 represents false and 1 represents true.\n\nStatement: STATEMENT"

A temperature value of 0.0 was used for GPT-4o and GPT-4o-mini in order to ensure reproducibility and in accordance with the findings of Pelrine et al. (2023) who identified this value as optimal. For o1-mini, the temperature parameter was not yet available, so the default value of 1.0 was utilized. Additionally, a max_completion_token value of 1024 (ISOT) or 512 (LIAR2) was used for o1-mini, which limits the amount of internal reasoning due to cost constraints. Moreover, some models produced invalid responses, which were excluded from subsequent analyses. The proportions of invalid responses per model and dataset are presented in Table 1.

4 Results

4.1 ISOT Fake News Dataset

The main results for the ISOT dataset are presented in Table 2. The baseline Naive Bayesand feature-based models exhibited the best performance, with a notable gap in accuracy being apparent relative to all LLMs. Even when specific feature subsets are removed, the feature classifier achieves competitive performance against the LLMs. Additionally, there were significant differences in accuracy between the different LLMs: while GPT-4o exhibited a performance that was nearly on par with the feature-based approach, GPT-4o-mini and particularly o1-mini demonstrated inferior performance.

4.2 LIAR2 Dataset

The main results for the LIAR2 dataset are presented in Table 3. The overall performance of all models was worse compared to the ISOT dataset. For statement data, GPT-4o demonstrated the highest performance, yet exhibited only a slight advantage over GPT-4o-mini and the Naive Bayes model.

Approach	Model	Accuracy
Naive Bayes	Word Counts	**0.9372**
Feature	Body + Title	0.9236
Feature	Body	0.8534
Feature	No NER	0.8524
Feature	No Sentiment	0.8272
Feature	No Character-Based	0.8262
LLM	GPT-4o	0.8582
LLM	GPT-4o-mini	0.7592
LLM	o1-mini	0.6685

Table 2: Misinformation Detection Accuracies in the ISOT Dataset

Approach	Model	Accuracy
Naive Bayes	Word Counts	0.6777
Feature	Statement	0.6130
LLM	GPT-4o	**0.6812**
LLM	GPT-4o-mini	0.6755
LLM	o1-mini	0.6466

Table 3: Misinformation Detection Accuracies in the LIAR2 Dataset

Then, o1-mini had a moderate decline in performance relative to GPT-4o-mini and Naive Bayes, and the feature-based approach achieved the lowest performance among the studied models.

5 Discussion

The objective of our study was to assess the predictive capacity of contemporary Large Language Models in the context of misinformation detection. To this end, we implemented a feature-based model and a word-count-based Naive Bayes model, and we compared their predictive performance to that of the target LLMs across two datasets.

For the ISOT dataset, the Naive Bayes model demonstrated the highest performance. This may be partially explainable by overfitting, as all the truthful examples of the ISOT dataset stem from the same website. However, the feature-based model involving title and body achieved a similar accuracy score while using a cleaner subset of training features. Interestingly, the LLMs performed notably worse, with significant differences between GPT-4o, GPT-4o-mini and o1-mini. Furthermore, a decline in performance was observed across all models when switching from the ISOT to the LIAR2 dataset, which is likely attributable to the increased task difficulty associated with classifying statements as opposed to complete news

Feature	Body Weights	Title Weights	Feature	Body Weights	Title Weights
Punctuation	-1.538	-0.933	Long	-0.823	-0.592
Tags	-0.886	-0.641	Length	0.487	-0.403
Citations	-0.139	0.439	Intensity	-1.727	-0.014
Capitalization	-0.182	–	Positive	0.510	0.403
Numeric	-0.276	-0.128	Negative	0.174	0.002
Uniqueness	-0.024	-0.250	Location	0.084	–
Short	-0.397	-2.103	Organization	0.034	–
Medium	0.748	-1.379	Person	0.051	–

Table 4: Feature Coefficients in Best Logistic Model on the ISOT Dataset

articles. The feature-based approach suffered the steepest decrease, likely due to scarcity of the extracted features present in the statements. Because of the unconvincing performance of the LLMs across both datasets, we advocate against prompting LLMs for the verification of news

Unsurprisingly, GPT-4o performed better than GPT-4o-mini on both datasets, which fits with the overall inferior performance of GPT-4o-mini across several benchmarks (OpenAI, 2024b). OpenAI o1-mini performed even worse on both datasets, probably due to its limited world knowledge and reduced language abilities. However, these properties might make o1-mini more suitable as part of a more sophisticated model than as a standalone misinformation detector. For example, Hu et al. (2024) explored using GPT-3.5 to provide rationales for SLMs, which may reflect an environment in which o1-mini's enhanced reasoning abilities could prove effective.

As indicated by the decreased performance of the feature-based model without using the title of news articles (see Table 2), the analysis of the title provides valuable information for detecting misinformation. A further in-depth analysis of the relative contributions of each feature is presented in Table 4. The character-based features and the features used as proxies for language quality were largely significant, indicating their importance for misinformation detection in news articles. Additionally, sentiment analysis proved to be an effective approach, as indicated by high coefficients of intensity and valence scores. Comparatively, the entity features established via NER did not offer much predictive support beyond the baseline text model. The aforementioned features offer valuable insights that could inform the development of robust LLM-based misinformation detection models. One potential approach is to explicitly prompt the LLM to focus on these features, which could enhance the model's ability to identify misinformation. Further investigation of this question is recommended for future research.

Conclusion

Our findings point out the limitations of Large Language Models (LLMs) in the task of misinformation detection. When using a simple prompting approach, LLMs are unable to reliably categorize the truthfulness of news articles and statements. The feature-based model excels for news articles, but falls short for single statements. Explicitly instructing LLMs to rely on such features could be a promising avenue for future research.

Limitations

We recognize the following limitations of our work:

- Our work focused on OpenAI's latest models, specifically GPT-4o, 4o-mini and o1-mini, without paying attention to previous models or other well-known LLMs, such as Claude, LLaMA or Gemini.

- We constructed a very simple, zero-shot prompt to resemble the LLM usage of a layperson and didn't investigate other prompting techniques, such as Chain-of-Thought prompting. As indicated by previous research (e.g. Hu et al., 2024), higher performance could be expected when using more sophisticated prompting approaches.

- The datasets we used focused on fake news detection in political statements and articles, limiting the generalizability of our results to other important domains, such as economy or health, as well as to non-English contexts.

Statement of Contributions

As a group, we collaborated extensively throughout the project. We met several times in the lead-up to writing the paper, both online and in person. These meetings allowed us to brainstorm ideas, divide responsibilities, provide feedback on each other's progress, and set deadlines. Alex focused on Named Entity Recognition and the Naive Bayes classifier, Domenico worked on the extraction of other text-based features, with a particular emphasis on sentiment analysis, as well as the implementation of the basic logistic classifier, and Lennart concentrated on the integration and implementation of the Large Language Models (LLMs) aspects of the project. Furthermore, Alex developed the initial preprocessing and Naive Bayes classifier scripts that were later on enhanced and optimized by Lennart, and Domenico set up online cluster access to run our computationally intensive pieces of code. Lastly, each of us reported on our individual findings, and we contributed to the report collectively while revising each others sections as a group.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *SECURITY AND PRIVACY*, 1(1):e9.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Kevin Matthe Caramancion. 2023. News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking.

Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A. Ghorbani. 2024. The largest social media ground-truth dataset for real/fake content: Truthseeker. *IEEE Transactions on Computational Social Systems*, 11(3):3376–3390.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Ajay Kumar and James W. Taylor. 2024. Feature importance in the age of explainable ai: Case study of detecting fake news misinformation via a multimodal framework. *European Journal of Operational Research*, 317(2):401–413.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024a. Gpt-4o.

OpenAI. 2024b. Gpt-4o-mini.

OpenAI. 2024c. Openai o1-mini.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4.

Ting Wei Teo, Hui Na Chua, Muhammed Basheer Jasser, and Richard T.K. Wong. 2024. Integrating large language models and machine learning for fake news detection. In *2024 20th IEEE International Colloquium on Signal Processing Its Applications (CSPA)*, pages 102–107.

Chih-Ming Tsai. 2023. Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis. *Electronics*, 12(17):3676.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Cheng Xu and M-Tahar Kechadi. 2024. An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006–88021.