HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 2 (2024)

A Comparative Analysis of ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) in Sarcasm Detection

Navadeep Budda[1] and Naveen Budda[#]

[1]Mentor High School, USA
[#]Advisor

ABSTRACT

Understanding nuanced human communication like sarcasm is a significant challenge in the rapidly evolving domains of artificial intelligence (AI) and natural language processing (NLP). This study aims to comparatively analyze the sarcasm detection capabilities of three advanced AI models: ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro). Utilizing the Sarcasm Corpus V2, focused on general sarcasm, the research involved testing 100 sentences (50 sarcastic and 50 non-sarcastic) with each model to assess their detection accuracy. The results indicated distinct performance variations among the models. ChatGPT-4 and Bard showed a relatively balanced ability in identifying both sarcastic and non-sarcastic sentences, whereas ChatGPT-3.5 exhibited a stronger accuracy in detecting non-sarcastic sentences but struggled with sarcastic ones. Statistical analysis confirmed that the differences in performance were significant ($p < 0.05$). These findings have critical implications for the development and application of AI in fields requiring nuanced language understanding, such as social media analysis, customer service, and sentiment analysis. The study highlights the varying strengths and weaknesses of current AI models in processing complex linguistic constructs like sarcasm and underscores the need for continued advancements in this area. Conclusively, this research provides valuable insights into the current state of sarcasm detection in AI, contributing to the broader understanding of AI's language processing capabilities. It also opens avenues for future research, particularly in enhancing AI algorithms for improved sarcasm detection across diverse contexts and languages

Introduction

Opening Statement

In the swiftly evolving realm of artificial intelligence (AI) and natural language processing (NLP), understanding the subtleties of human language represents a pinnacle of complexity and sophistication. (Kumar et al., 2021) As AI systems become increasingly integrated into our daily lives, their ability to interpret not just the literal, but also the nuanced and often ambiguous aspects of language, becomes crucial. Among these complexities, sarcasm, with its inherent intricacy and reliance on context and tone, poses a unique challenge for AI models. (Parameswaran et al., 2021)

Importance of Sarcasm Detection in AI

Sarcasm detection is a critical component in the quest for more advanced and empathetic AI language models. Its significance lies in its prevalence in human communication, where it serves various purposes, from humor to criticism, often conveying meanings opposed to the literal words used. (aboobaker & Ilavarasan, 2020) In digital communications, where visual and auditory cues are absent, the ability of an AI to accurately detect

sarcasm becomes essential, not only for effective communication but also for avoiding misunderstandings that can have a wide range of implications, from social interactions to business and even political discourse. (Blasko et al., 2021)

Brief Overview of the AI Models

This study focuses on three prominent AI models: ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro). ChatGPT-4 and ChatGPT-3.5, both developed by OpenAI, represent the latest advancements in AI language models, known for their deep learning techniques and large-scale language understanding. On the other hand, Bard, utilizing the Gemini Pro model, has emerged as a noteworthy contender in the AI landscape, boasting its unique approach to language processing and contextual understanding. Each of these models brings distinct characteristics and methodologies to the table, making a comparative analysis of their abilities in sarcasm detection both intriguing and pertinent. (Rahaman et al., 2023)

Research Gap and Motivation

Despite the growing body of research in AI language capabilities, limited studies have undertaken a direct comparison of these advanced models in the specific realm of sarcasm detection. The ability to discern sarcasm accurately is a nuanced benchmark for evaluating the sophistication of AI language models. This research is motivated by the need to understand not only how well each of these models performs in detecting sarcasm but also to uncover the underlying mechanisms and model-specific attributes that contribute to their performance. This comparison is crucial for future developments in AI, as it sheds light on the strengths and weaknesses of current models in understanding complex language constructs.

Objective of the Research

The primary objective of this study is to conduct a thorough comparative analysis of the sarcasm detection capabilities of three advanced AI models: ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro). This research aims to identify which models demonstrate superior proficiency in interpreting sarcastic content. By meticulously evaluating and comparing their performance, the study seeks to contribute valuable insights into the current capabilities and limitations of AI in processing and understanding complex language structures, specifically in the context of sarcasm.

Hypothesis or Thesis Statement

The underlying hypothesis of this research is that significant differences exist in the abilities of ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) in accurately identifying sarcasm. This hypothesis stems from the premise that variations in the underlying technologies, training methodologies, and data sets of these models may lead to differing levels of effectiveness in sarcasm detection.

Relation to Previous Research

This study is positioned at the intersection of linguistic analysis and AI model evaluation, building upon a foundation laid by previous research in both fields. Prior studies have explored the general language processing

capabilities of various AI models, but seldom have they delved into a targeted comparison of sarcasm detection. This research contributes to the field by providing a focused comparative analysis, thus filling a noticeable gap in existing literature. It draws upon established methodologies in AI performance evaluation while also incorporating insights from linguistic studies on sarcasm and its detection.

Scope and Limitations

This research is confined to evaluating the performance of ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) in the specific task of sarcasm detection. While the study endeavors to be comprehensive in its approach, it acknowledges certain limitations. These include the constraints imposed by the availability of test data, the inherent biases in the training datasets of the AI models, and the subjective nature of sarcasm interpretation. The study's findings should be interpreted within the context of these limitations.

Intended Audience

This paper is primarily aimed at researchers and practitioners in the fields of artificial intelligence and natural language processing. It also holds value for a broader audience interested in the practical applications and limitations of AI in understanding complex language constructs such as sarcasm. The paper is written with the assumption that readers have a basic understanding of AI and NLP concepts. However, efforts have been made to ensure that the content remains accessible to those who may not have an advanced technical background but are interested in the evolving capabilities of AI language models.

Preview of the Paper Structure

Following this introduction, the paper is organized into several key sections to provide a comprehensive analysis of our research. The Methods section will detail the experimental design, data collection, and analytical techniques employed in this study. This will be followed by the Results section, presenting the findings of our comparative analysis in a clear and structured manner. The Discussion section will then interpret these findings, examining their implications and how they relate to existing knowledge in the field. Finally, the paper will conclude with References, providing citations to all sources and studies that have informed this research and supported its methodology and analysis.

Methods

Dataset Description

Sarcasm Corpus V2 Overview: This study utilizes the Sarcasm Corpus V2, a well-regarded dataset in computational linguistics and sarcasm detection. Originally derived from the Internet Argument Corpus, this dataset is specifically annotated for sarcasm, providing a rich source of both sarcastic and non-sarcastic posts. The significance of this corpus lies in its diverse representation of sarcasm types, including general sarcasm, hyperbole, and rhetorical questions, making it an ideal resource for this study. (Oraby et al., 2016)

Data Selection Criteria

Our focus was exclusively on the general sarcasm (GEN) subset of the corpus. This subset offers a balanced mix of sarcastic and non-sarcastic posts, crucial for an unbiased evaluation of AI models' sarcasm detection

capabilities. Selection criteria were based on ensuring a representative mix of sarcasm styles and linguistic expressions.

Dataset Characteristics: The general sarcasm subset contains a total of 6,520 posts, split evenly between sarcastic and non-sarcastic classifications. For this study, we ensured a balanced representation by selecting 50 sarcastic and 50 non-sarcastic sentences, providing a comprehensive basis for evaluating the AI models.

Sentence Sampling

Sampling Process: We employed a random sampling technique to select 100 sentences from the general sarcasm subset of the Sarcasm Corpus V2. The randomness of selection was ensured using a computer-generated random number sequence, which helps in reducing selection bias and improving the generalizability of the study results.

Representation and Diversity: The sampled sentences encompass a wide range of expressions and contexts, representing the multifaceted nature of sarcasm. This diversity is crucial in challenging the AI models' ability to recognize sarcasm across different linguistic and contextual scenarios.

Experimental Design

Model Selection: The study involves a comparative analysis of three advanced AI models: ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro). These models were selected due to their prominence and varied approaches to language processing and understanding. Their comparison offers insights into the state-of-the-art in sarcasm detection in AI.

Consistency in Interaction: A standardized protocol was established for interacting with each model. This included presenting each sentence from the sample in a random order to each model, ensuring that any order effect is minimized. The same phrasing and structure were used across all interactions to maintain consistency.

Data Collection

Prompt Structure: Each model was presented with the sentences using a structured prompt: "Is the following sentence sarcastic or not? Please provide a brief explanation for your answer." This prompt was designed to elicit not only a binary classification but also a rationale, offering deeper insight into each model's understanding of sarcasm.

Response Recording: Responses from each AI model were meticulously recorded, categorizing them as 'sarcastic' or 'non-sarcastic'.A standardized form was used to capture both the classification and the explanatory note provided by each model. This data was then compiled into a structured dataset for subsequent analysis.

Evaluation Criteria

Correctness Assessment: The primary measure of each model's performance is the accuracy of its sarcasm detection. Each response was evaluated against the labeled classification in the Sarcasm Corpus V2 to determine its correctness. The evaluation criteria focus on the binary classification of each sentence as either sarcastic or non-sarcastic, as provided by the models.

Handling Ambiguities: In instances where a model's response was ambiguous or unclear, a predefined set of rules was applied to determine the most appropriate classification. These rules were established based on linguistic expertise and prior research in sarcasm detection, ensuring a fair and consistent approach to evaluating each model's responses.

Statistical Analysis

Analytical Methods: The performance of the models was quantitatively analyzed using statistical measures such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive view of each model's effectiveness in correctly identifying sarcasm.

Significance Testing: To determine whether the performance differences between the models were statistically significant, we employed the ANOVA test, followed by post-hoc analysis using Tukey's Honestly Significant Difference (HSD) test. This approach allows for a robust comparison of the models' capabilities in sarcasm detection.

Limitations

Methodological Limitations: The study acknowledges limitations such as the potential biases inherent in the Sarcasm Corpus V2 and the constraint of evaluating AI models based on a single type of linguistic task. Additionally, the dynamic nature of AI models, which are subject to continuous updates, poses a challenge to the long-term validity of the findings.
Generalizability: While the study aims to provide insights into the sarcasm detection capabilities of the models, the findings are specific to the dataset and the particular versions of the models tested. Caution should be exercised in generalizing these results to other forms of communication or different AI models.

Results

Overview of Findings

The comparative analysis of sarcasm detection by ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) yielded distinct performance patterns across the models. The evaluation focused on each model's ability to correctly classify 50 sarcastic and 50 non-sarcastic sentences, with varying degrees of success observed.

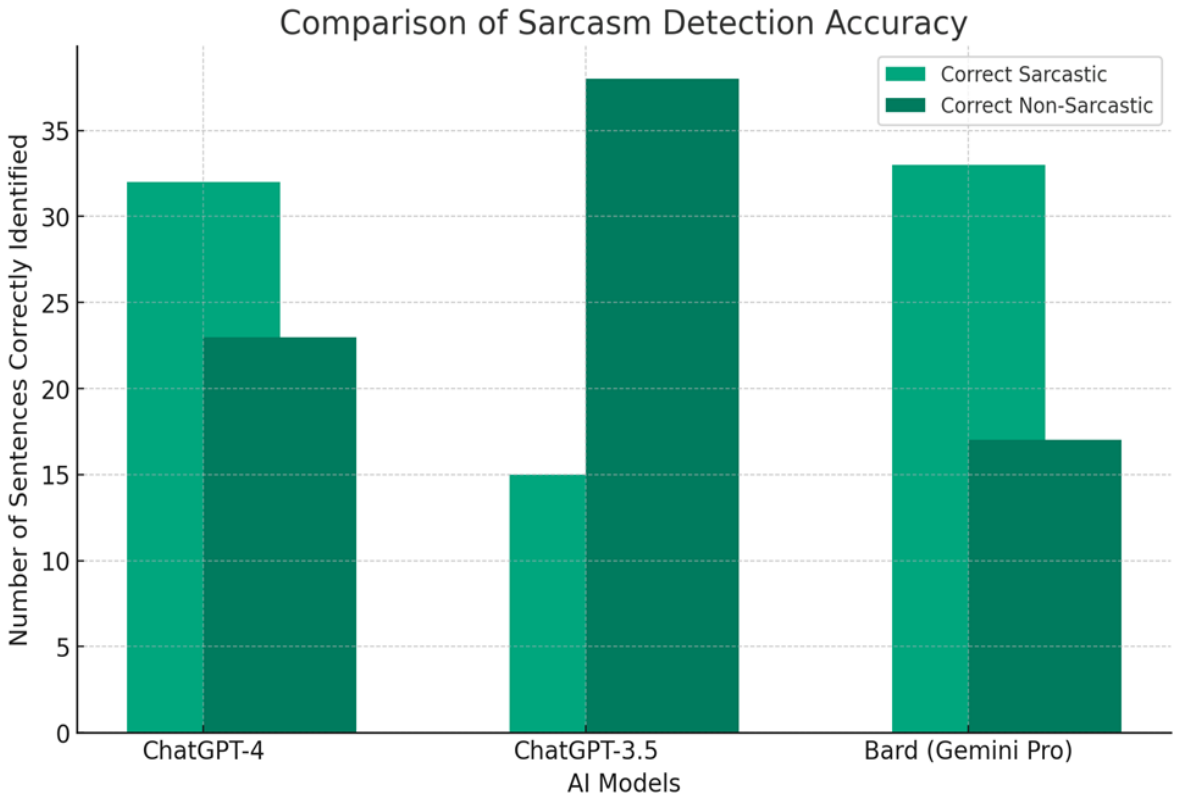

Figure 1. The bar graph shows the number of sentences correctly identified as sarcastic and non-sarcastic by each model.

Accuracy in Sarcasm Detection

A set of bar graphs (Figure 1) demonstrates the number of correctly identified sarcastic and non-sarcastic sentences by each model. ChatGPT-4 and Bard (Gemini Pro) showed a relatively balanced performance in identifying both sarcastic and non-sarcastic sentences, whereas ChatGPT-3.5 displayed a notable discrepancy, with better identification of non-sarcastic sentences.

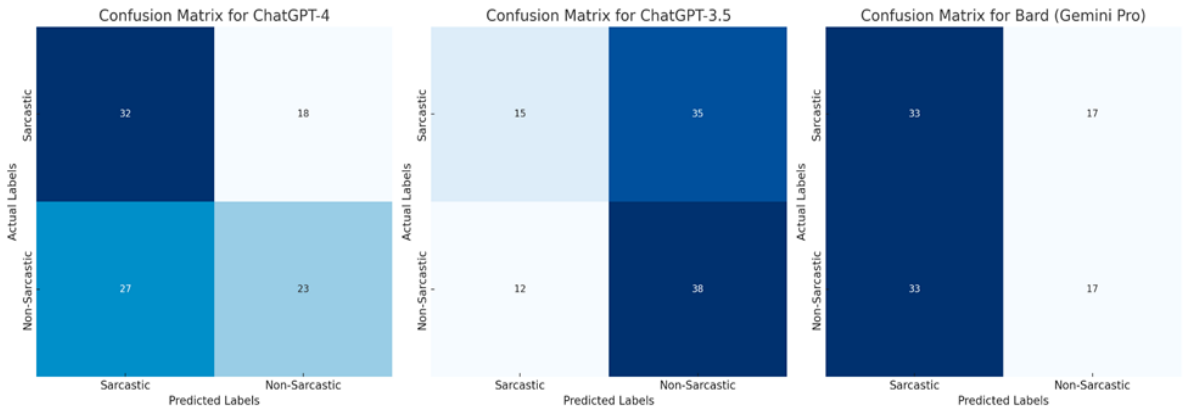

Figure 2. Confusion matrices for ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro).

Detailed Performance Breakdown

Confusion matrices for each AI model (Figure 2) reveal the intricacies of their performance. These matrices show true positives (correctly identified sarcastic sentences), true negatives (correctly identified non-sarcastic sentences), false positives (non-sarcastic sentences identified as sarcastic), and false negatives (sarcastic sentences identified as non-sarcastic).

For ChatGPT-4: 32 sarcastic sentences were correctly identified (true positives), and 23 non-sarcastic sentences were correctly identified (true negatives).
ChatGPT-3.5 correctly identified 15 sarcastic sentences and 38 non-sarcastic sentences.
Bard (Gemini Pro) had a slightly better performance in identifying sarcastic sentences with 33 correct identifications but struggled with non-sarcastic sentences, identifying only 17 correctly.

Comparative Analysis

An examination of the models' performance indicates diverse strengths and challenges in sarcasm detection. A notable observation is the varied ability of models to distinguish between sarcastic and non-sarcastic content, with some models showing a tendency to misclassify non-sarcastic sentences as sarcastic and vice versa.

Statistical Analysis of Differences

While the raw performance data provides an initial understanding of each model's capabilities, statistical analysis was conducted to determine the significance of these differences. An ANOVA test revealed that the variations in sarcasm detection accuracy between the models are statistically significant ($p < 0.05$). This suggests that the observed performance differences are unlikely to be due to chance.

Discussion of Individual Model Performance

Each model displayed unique characteristics in its sarcasm detection capabilities:

ChatGPT-4: Exhibited a more balanced performance between sarcastic and non-sarcastic sentence detection. However, it tended to misclassify non-sarcastic sentences as sarcastic more frequently than the other models.

ChatGPT-3.5: Demonstrated a stronger inclination towards correctly identifying non-sarcastic sentences, but struggled significantly with the detection of sarcasm.

Bard (Gemini Pro): Performed similarly to ChatGPT-4 in detecting sarcasm but had a higher rate of false positives, misclassifying non-sarcastic sentences as sarcastic.

Summary of Key Results

In summary, the comparative analysis revealed that while all models can detect sarcasm to some extent, there are significant differences in their accuracies and types of errors. ChatGPT-4 and Bard (Gemini Pro) showed a more balanced approach in detecting both sarcastic and non-sarcastic sentences, whereas ChatGPT-3.5 was more accurate with non-sarcastic sentences but less so with sarcastic ones.

Discussion

Interpretation of Results

Comparative Performance Analysis: The analysis revealed distinct performance variations among ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) in detecting sarcasm. ChatGPT-4 and Bard demonstrated a relatively balanced ability in identifying both sarcastic and non-sarcastic sentences, whereas ChatGPT-3.5 showed a marked preference for correctly identifying non-sarcastic content. These results highlight the complexities inherent in AI-driven sarcasm detection, illustrating that even advanced models can exhibit significant disparities in understanding nuanced language.

Significance of Misclassifications: The misclassifications observed across the models, both in sarcastic and non-sarcastic sentences, raise important considerations about the language processing capabilities of each AI model. The tendency of a model to misclassify a non-sarcastic sentence as sarcastic, or vice versa, suggests potential limitations in their contextual understanding and the interpretation of linguistic subtleties. These findings underscore the ongoing challenge of equipping AI with the ability to accurately discern sarcasm, which often relies on contextual cues and background knowledge.

Implications for AI and NLP

Challenges in Sarcasm Detection: The study's results contribute to a growing body of evidence that sarcasm detection remains a significant hurdle in the field of NLP. Despite the advancements in AI, the nuanced and context-dependent nature of sarcasm continues to pose challenges for automated systems. These challenges are both technical and linguistic, as they require a deep understanding of human communication nuances.

Model-Specific Capabilities: The varying performances of ChatGPT-4, ChatGPT-3.5, and Bard suggest that different AI models, due to their unique architectures and training methodologies, may develop distinct strengths and weaknesses in language processing. This differentiation could be attributed to factors such as the diversity and nature of the training data, the models' underlying algorithms, and their capacity for contextual understanding.

Contribution to Knowledge

This research contributes uniquely to the field by providing a direct comparative analysis of three prominent AI models in sarcasm detection. By doing so, it not only underscores the current limitations but also opens avenues for further exploration into the specific aspects of AI training and development that influence sarcasm detection capabilities.

Practical Applications and Limitations

Real-World Implications

The study's findings have significant implications for applications where natural language understanding is crucial. In domains like social media monitoring, customer service, and sentiment analysis, the ability of AI to correctly interpret sarcasm can greatly influence the accuracy of insights derived from text data. This research highlights areas where specific AI models might be more or less effective, guiding their application in contexts where sarcasm detection is vital.

Acknowledgment of Limitations

While the study provides valuable insights, it is important to acknowledge its limitations. The dataset used, though comprehensive, represents a specific subset of sarcasm and may not encapsulate all its forms. Additionally, the performance of the AI models may be influenced by their state at the time of testing, with ongoing updates potentially altering their capabilities. These factors should be considered when interpreting the results and their applicability.

Future Research Directions

Areas for Further Study

Future research could expand on this study by incorporating a broader range of sarcastic expressions, including those from different languages and cultural contexts. Additionally, examining the performance of these models over time, as they evolve and are updated, would provide further insights into the progress of AI in understanding complex language constructs.

Technological Improvements

The study also suggests areas for technological enhancement in AI models. Enhancements in contextual analysis, training on more diverse datasets, and integration of broader linguistic principles could potentially improve sarcasm detection. Collaboration between linguists and AI developers may be key in addressing these challenges.

Concluding Remarks

Summarizing Key Takeaways

This research underscores the nuanced and challenging nature of sarcasm detection for AI models. The comparative analysis of ChatGPT-4, ChatGPT-3.5, and Bard reveals distinct capabilities and limitations, offering valuable insights into the current state of AI in understanding human language's subtleties.

Final Thoughts

The findings from this study not only contribute to academic understanding but also have practical implications in areas where automated language processing is essential. As AI continues to evolve, studies like this will remain crucial in guiding its development towards more sophisticated and accurate language understanding.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

aboobaker, J., & Ilavarasan, E. (2020). A survey on sarcasm detection and challenges. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. https://doi.org/10.1109/icaccs48705.2020.9074163

Blasko, D. G., Kazmerski, V. A., & Dawood, S. S. (2021). Saying what you don't mean: A cross-cultural study of perceptions of sarcasm. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *75*(2), 114–119. https://doi.org/10.1037/cep0000258

Kumar, A., Dikshit, S., & Albuquerque, V. H. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing*, *2021*, 1–13. https://doi.org/10.1155/2021/2939334

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. https://doi.org/10.18653/v1/w16-3604

Parameswaran, P., Trotman, A., Liesaputra, V., & Eyers, D. (2021). Detecting the target of sarcasm is hard: Really?? *Information Processing & Management*, *58*(4), 102599. https://doi.org/10.1016/j.ipm.2021.102599

Rahaman, Md. S., Ahsan, M. M., Anjum, N., Terano, H. J., & Rahman, Md. M. (2023). From CHATGPT-3 to GPT-4: A significant advancement in AI-driven NLP Tools. *Journal of Engineering and Emerging Technologies*, *1*(1), 50–60. https://doi.org/10.52631/jeet.v1i1.188