EDITOR: Erik Cambria, cambria@ntu.edu.sg

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

Can Generative AI Models Extract Deeper Sentiments as Compared to Traditional Deep Learning Algorithms?

Mohammad Anas, Anam Saiyeda, *Jamia Hamdard University, 110062, New Delhi, India*

Shahab Saquib Sohail, *VIT Bhopal University, 466114, Madhya Pradesh, India*

Erik Cambria, *Nanyang Technological University, 639798, Singapore*

Amir Hussain, *Edinburgh Napier University, EH11 4BN, Edinburgh, UK*

Abstract—Recent advances in the context of deep learning have led to the development of generative AI models, which have shown remarkable performance in complex language understanding tasks. This study proposes an evaluation of traditional deep learning algorithms and generative AI models for sentiment analysis. Experimental results show that RoBERTa outperforms all models, including ChatGPT and Bard, suggesting that generative AI models are not yet able to capture the nuances and subtleties of sentiment in text. We provide valuable insights into the strengths and weaknesses of different models for sentiment analysis and offer guidance for researchers and practitioners in selecting suitable models for their tasks.

Sentiment analysis stands as a pivotal task in natural language processing (NLP), crucial for comprehending and extracting polarity and emotions embedded in textual data. This ability to automatically decipher sentiments holds wide-ranging applications, from business and market analysis to social media monitoring and customer feedback assessment [1]. In an era marked by an unprecedented surge in textual data, sentiment analysis emerges as an invaluable tool for organizations and researchers aiming to extract insights from the vast sea of unstructured text.

Recent strides in NLP, notably the advent of generative AI models like OpenAI's ChatGPT and Google's Bard, have ushered in a revolution in sentiment analysis. These models, grounded in transformer architecture and enriched through extensive pre-training on diverse text data, exhibit exceptional prowess in handling intricate language understanding tasks [2].

ChatGPT is a member of the Generative Pretrained Transformer (GPT) family that epitomizes a pinnacle achievement in conversational AI. Fueled by deep neural networks and trained on a diverse array of internet text, ChatGPT showcases human-like language understanding and generation capabilities [3]. Widely adopted for tasks requiring contextual understanding, such as sentiment analysis and natural language inference, it excels in grasping nuances in language and context [4]. Furthermore, in [5], authors have explored its taxonomy, applications, and assess its trustworthiness. The work advocates its capability and potential in several real-life scenarios.

On the other hand, Bard, with its bidirectional architecture and autoregressive training approach, establishes itself as a robust performer in various NLP benchmarks. Its prowess in contextual comprehension and capturing long-range dependencies positions it as a formidable contender in sentiment analysis. The rise of these generative AI models raises intriguing questions about their efficacy in sentiment analysis compared to traditional methods.

Digital Object Identifier 10.1109/MIS.2024.XXXXXXX

This paper delves into these questions, exploring unique insights, challenges, and the models' capacity to capture the intricate subtleties of sentiment expressed in text. These models find applications in diverse fields, such as Data Science, where they aid scientists in automating workflow components [6]. As the volume of textual data on social media expands, the focus shifts to interpreting online user reviews, aiming for a more thorough perspective on user-generated content and sentiment intensity.

The review of current literature suggests that important research investigating generative AI models performance for sentiment analysis of product reviews is missing. To this end, we have comprehensively evaluated the performance of two generative AI models, namely Bard and ChatGPT, for sentiment analysis. In addition, we compare how these perform as compared to traditional deep learning algorithms, namely SenticNet [7], a long short-term memory (LSTM) network [8], and RoBERTa [9].

RELATED WORK

In the realm of ChatGPT applications across diverse fields, a comprehensive investigation by [10] assessed its performance in approximately 25 varied analytical NLP tasks, employing human evaluation criteria. These tasks spanned domains such as emotion recognition, toxicity detection, sentiment analysis, and stance detection. Turning to healthcare researchers' perceptions of ChatGPT [11], a study harnessed deep learning methodologies, utilizing pretrained BERT models for sentiment analysis and topic modeling. The analysis focused on social media posts authored by healthcare researchers, providing valuable insights into their emotional sentiments regarding ChatGPT. Another evaluation of ChatGPT [12] delved into its performance across three affective computing challenges: Big Five personality prediction, sentiment analysis, and suicide tendency detection. This assessment compared ChatGPT with RoBERTa, a Bag-of-Words (BoW) baseline, and pretrained word embeddings (word2Vec), offering a comprehensive analysis.

In the realm of assessment item production [13], an investigation aimed to ascertain whether ChatGPT and Bard could generate coherent assessment items. This involved a reliability analysis contrasting the performance of ChatGPT and Bard tools with that of proficient and trained human evaluators, focusing on the complexity of writing prompts. A comparative assessment of chatbot technologies [14], including ChatGPT, Bard, and Microsoft Bing, combined quantitative scrutiny of ChatGPT's performance metrics with qualitative analysis of user feedback. The findings revealed noteworthy distinctions, with ChatGPT displaying superior accuracy and relevance, Bard excelling in response time, and Microsoft Bing standing out for remarkable user satisfaction and engagement.

In the context of myopia care [15], a study presented a benchmarking analysis of generative AI models, specifically ChatGPT-3.5, ChatGPT-4.0, and Bard, focusing on their ability to provide precise responses to frequently encountered myopia care-related inquiries. Notably, ChatGPT-4.0 exhibited a higher degree of accuracy compared to its counterparts. Sentiment analysis of tweets related to ChatGPT [16] employed a sophisticated BERT model. The model, built on a deep transformer architecture, comprised three dense layers of neural networks to process and analyze sentiment embedded in tweets.

For sarcasm analysis [17], a comprehensive examination incorporated traditional deep learning techniques like LSTM and BERT. The objective was to discern and categorize the level of contextual information or response required for effective sarcasm detection. Evaluating product opinions from online retailers and merchants [18], an approach employing the Skip-gram architecture incorporated LSTM to discern intricate patterns within textual data. This enhanced the feature extraction process for semantic and contextual information of words. In evaluating the research trend of generative AI models [19], comparing ChatGPT, GPT4, and Bard publication trends with early COVID-19 research also revealed insightful patterns.

Our study embarks on an in-depth exploration of sentiment analysis using product reviews as the primary dataset. We conduct a comparative analysis of sentiment analysis performance across various models, including ChatGPT and Bard. Our research uniquely focused on comparing the accuracy of ChatGPT and Bard specifically in the context of sentiment analysis, filling a gap in the literature that has received limited attention. The study sheds light on the nuanced strengths and weaknesses of generative AI models in capturing sentiment nuances within product reviews.

PROPOSED EVALUATION

Dataset

The selected dataset comprises an extensive collection of consumer reviews, encompassing over 67,000 evaluations of various Amazon products (combined all three datasets available under Consumer Reviews of Amazon Products on Kaggle), including popular items such as the Kindle, Fire TV Stick, and more.

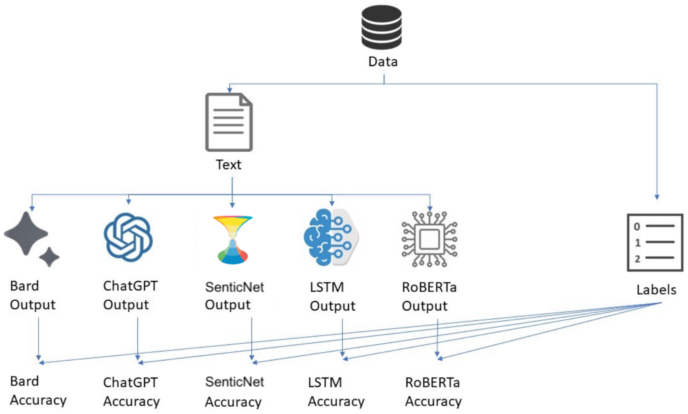

FIGURE 1. Architecture for the comprehensive evaluation of foundational models and generative AI models

This dataset was sourced from Datafiniti's Product Database, a reputable repository of product-related information[1]. Each entry in the dataset provides comprehensive insights into consumer sentiment and product characteristics. This diverse dataset offers a rich resource for our research, enabling us to conduct a thorough investigation into the performance of various sentiment analysis models across a wide range of consumer reviews. The inclusion of product-specific details ensures a nuanced evaluation of sentiment and facilitates the identification of trends and patterns within the dataset.

Methodology

The proposed mechanism for the evaluation is represented in Fig. 1. To facilitate a comprehensive evaluation of the selected models, we began by splitting our dataset into two subsets. Specifically, we reserved 500 product reviews for testing and comparison purposes, while the remaining reviews were allocated for training purposes. This separation allowed us to ensure an unbiased evaluation of our chosen methods.

Generative AI Models

For our investigation, we selected OpenAI's ChatGPT and Google's Bard. We initiated the evaluation process by passing the reserved testing reviews to each model. We tasked both models with the responsibility of rating these reviews on a scale of 1 to 5, reflecting customer satisfaction levels. This approach allowed us to gauge their respective capabilities in assessing customer sentiment based on the textual content of the reviews.

SenticNet

SenticNet is a commonsense-based neurosymbolic framework that belongs to a family of knowledge bases built by means of deep learning [20]. In particular, it employs unsupervised and reproducible subsymbolic techniques such as auto-regressive language models and kernel methods to build trustworthy symbolic representations that extract polarity from text in a fully explainable manner. The resource is available both as a knowledge base and in the form of a API suite[2] for various sentiment analysis tasks.

LSTM

LSTM is a recurrent neural network that excels at capturing long-term dependencies, making it ideal for sequence prediction tasks. We utilized the Keras preprocessing Tokenizer to convert the product reviews into sequences suitable for input into the LSTM model. Further preprocessing steps, including padding sequences, were applied to ensure compatibility with the model architecture. The reviews were then passed through the LSTM model for comprehensive analysis.

RoBERTa

RoBERTa is a large language model based on the transformer architecure. To begin, we employed the pretrained RoBERTa Tokenizer in conjunction with RoBERTaForSequenceClassification[3]. This allowed us to transform the product reviews and subsequently feed them into the model for training and evaluation.

[1] https://kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products

[2] https://sentic.net/api

[3] https://huggingface.co/transformers/v2.9.1/model_doc/roberta.html

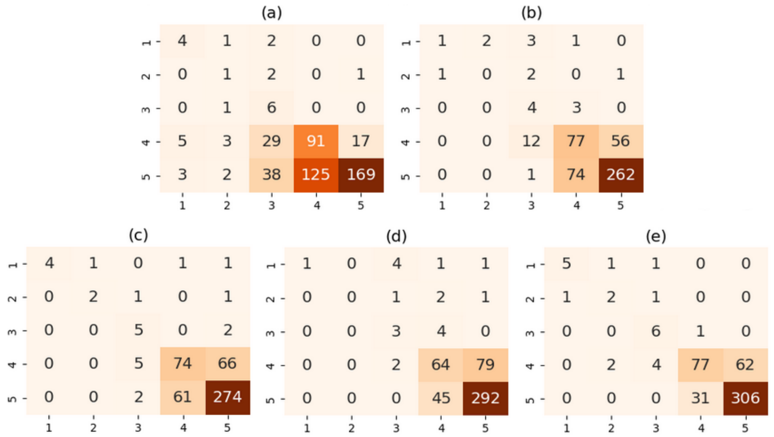

FIGURE 2. Confusion matrix for a) Bard, b) ChatGPT, c) SenticNet, d) LSTM, and e) RoBERTa

TABLE 1. Performance metric for all the models

Metrics	Bard	ChatGPT	SenticNet	LSTM	RoBERTa
accuracy	0.54	0.69	0.72	0.72	0.79
precision	0.74	0.70	0.73	0.71	0.78
recall	0.54	0.69	0.72	0.72	0.79
F1-score	0.59	0.36	0.72	0.71	0.78

DISCUSSION

The results of our sentiment analysis experiments offer valuable insights into the performance of various models in classifying sentiments within product reviews. Among the models examined, RoBERTa stands out with its exceptional accuracy, precision, recall, and F1-score, as shown in Table 1. This underscores the effectiveness of cutting-edge transformer-based architectures in capturing intricate sentiment nuances. ChatGPT and Bard also demonstrated competitive results, with strong accuracy (0.68) and (0.54), affirming their potential in the sentiment analysis domain. Their balanced performance is well-suited to applications requiring both precision and recall.

Meanwhile, SenticNet achieved respectable results, especially considering that its strength is not on accuracy but rather explainability of sentiment results, e.g., in terms of emotion-cause pair extraction. LSTM delivered good results too, aligning with the state-of-the-art models and highlighting the utility of deep learning techniques in this context. ChatGPT and Bard, both prominent contenders in our sentiment analysis study, demonstrated commendable performance in Fig. 3. However, upon closer examination, distinct characteristics emerged that shed light on their suitability for different applications. ChatGPT exhibited impressive adaptability in restructuring responses to meet specific requirements and addressing a wide range of queries.

Its capacity to handle multiple reviews simultaneously, providing separate ratings for each, adds to its versatility. In contrast, Bard displayed a more specific approach in its responses, offering precision in its analysis. Nevertheless, Bard encountered challenges when dealing with multiple reviews concurrently, which impacted its overall efficiency. Notably, ChatGPT exhibited a higher accuracy rate of 0.70, reflecting its proficiency in sentiment analysis. In contrast, Bard achieved an accuracy rate of 0.54, indicating its potential but also room for improvement. These distinctions highlight the nuanced capabilities of each model, and their choice should align with the specific requirements of the sentiment analysis task at hand.

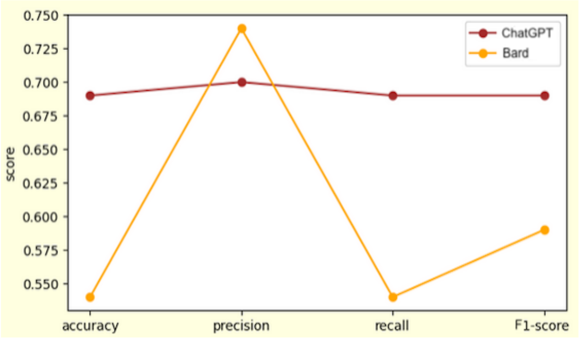

FIGURE 3. Performance evaluation of ChatGPT and Bard

CONCLUSION

This study presents an evaluation of generative AI models and traditional deep learning techniques for sentiment analysis. In particular, we compared Bard, ChatGPT, SenticNet, LSTM, and RoBERTa in terms of accuracy, precision, recall, and F1-score. The results demonstrate that RoBERTa, a transformer-based large language model, outperforms all models on all metrics, achieving an accuracy of 0.79 and an F1-score of 0.78. This indicates the superiority of large language models in capturing the nuances and subtleties of sentiment expressed in text. RoBERTa's robust performance aligns with the growing prominence of transformer architectures in the field of NLP.

When iteratively trained on huge amounts of data, however, large language models may be prone to the so-called 'short blanket dilemma': while trying to improve their accuracy on some tasks, they may inadvertently become worse for tasks which they previously excelled at. While generative AI models are generally considered better than traditional deep learning algorithms, we shall not forget that they are fine-tuned for generative tasks like dialogue generation or text completion. Hence, they may not perform as well as traditional deep learning algorithms trained for a specific downstream task like sentiment analysis. Nonetheless, ChatGPT's adaptability and ability to handle multiple reviews simultaneously make it suitable for affective computing applications requiring versatility, while Bard's precision in analysis make it suitable for applications requiring in-depth sentiment analysis. In the future, the explainability of these models will play an increasingly important role. To this end, we also provided a holistic analysis of the model behaviors through confusion matrices, which reveal not only the predictive prowess but also the interpretability of each model.

Overall, this study provided valuable insights into the comparative performance of traditional deep learning algorithms and generative AI models for sentiment analysis, aiding researchers and practitioners in selecting suitable frameworks for their tasks. This research suggests multiple compelling avenues for advancing sentiment analysis. Firstly, there's a need to explore how different domains and datasets influence model performance. The study focused on product reviews, and extending this analysis to diverse text types like social media posts, news articles, or personal blogs could reveal how models handle varying levels of complexity, formality, and subjectivity inherent in different language styles.

Secondly, a deeper dive into the intricacies of sentiment, especially concerning sarcasm, irony, and humor, is warranted. Developing models with heightened sensitivity to these nuanced emotional cues could substantially refine sentiment analysis. Additionally, investigating the impact of domain-specific knowledge on sentiment analysis outcomes holds promise. Integrating domain-specific lexicons and contextual information into models might enhance their capacity to discern sentiment within specific domains like finance or healthcare. Thirdly, there's potential in synergizing different sentiment analysis techniques. Hybrid models that combine traditional deep learning approaches with generative AI models could offer a more comprehensive and accurate understanding of sentiment.

Lastly, addressing the ethical implications and biases in sentiment analysis is imperative. Future research should focus on developing methods to mitigate bias, ensuring responsible and ethical use of models in real-world applications. Embracing explainable AI methodologies and incorporating user feedback mechanisms will contribute to model interpretability and adaptability, aligning sentiment analysis research with the evolving landscape of linguistic expression.

REFERENCES

1. L. Oneto et al., "Statistical Learning Theory and ELM for Big Social Data Analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, 2016, pp. 45–55.
2. E. Cambria et al., "Seven Pillars for the Future of Artificial Intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, 2023, pp. 62–69.
3. M. Amin, E. Cambria, and B. Schuller, "Can ChatGPT's Responses Boost Traditional Natural Language Processing?" *IEEE Intelligent Systems*, vol. 38, no. 5, 2023, pp. 5–11.
4. B. Gupta et al., "ChatGPT: A Brief Narrative Review," *Cogent Business & Management*, vol. 10, no. 3, 2023, p. 2275851.
5. S. Sohail et al., "Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, 2023, p. 101675.
6. H. Hassani and E. Silva, "The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces are Revolutionizing the Field," *Big data and cognitive computing*, vol. 7, no. 2, 2023, p. 62.

7. E. Cambria et al., "SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis," *LREC*, 2022, pp. 3829–3839.
8. S. Hochreiter; and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, 1997, pp. 1735–1780.
9. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
10. J. Kocoń et al., "ChatGPT: Jack of All Trades, Master of None," *Information Fusion*, vol. 99, 2023, p. 101861.
11. S. Praveen and V. VVajrobol, "Understanding the Perceptions of Healthcare Researchers Regarding ChatGPT: A Study Based on Bidirectional Encoder Representation from Transformers (BERT) Sentiment Analysis and Topic Modeling," *Annals of Biomedical Engineering*, 2023.
12. M. Amin, E. Cambria, and B. Schuller, "Will Affective Computing Emerge from Foundation Models and General AI? A first evaluation on ChatGPT," *IEEE Intelligent Systems*, vol. 38, no. 2, 2023, pp. 15–23.
13. A. Khademi, "Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance," *arXiv preprint arXiv:2304.05372*, 2023.
14. S. Bhardwaz and J. Kumar, "An Extensive Comparative Analysis of Chatbot Technologies-ChatGPT, Google Bard and Microsoft Bing," *ICAAIC*, 2023, pp. 673–679.
15. Z. Lim et al., "Benchmarking Large Language Models' Performances for Myopia Care: A Comparative Analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard," *EBioMedicine*, vol. 95, no. 104770, 2023.
16. M. Mujahid et al., "Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach," *Information*, vol. 14, no. 9, 2023, p. 474.
17. A. Kalaivani and D. Thenmozhi, "Sarcasm Identification and Detection in Conversion Context using BERT," *Second Workshop on Figurative Language Processing*, 2020, pp. 72–76.
18. Y. Wu, M. Al-Daweri, and V. Kumar, "Emotional Analysis Based on LSTM-CNN Hybrid Neural Network Model," *International Conference On Software Engineering and Computer Systems*, 2023.
19. J. Shobana and M. Murali, "An Efficient Sentiment Analysis Methodology based on Long Short-Term Memory Networks," *Complex & Intelligent Systems*, vol. 7, no. 5, 2021, pp. 2485–2501.
20. F. Xing, F. Pallucchini, and E. Cambria, "Cognitive-Inspired Domain Adaptation of Sentiment Lexicons," *Information Processing and Management*, vol. 56, no. 3, 2019, pp. 554–564.

Mohammad Anas is currently a final-year Bachelor of Technology (BTech) student at Jamia Hamdard University, India. His research interests include deep learning, large language models, and natural language processing. Contact him at mohammadanas@jamiahamdard.ac.in.

Anam Saiyeda is currently an assistant professor at Jamia Hamdard University, India. Her research interests lie in cloud computing, smart education, and genetic algorithms. Contact her at anamsaiyeda@jamiahamdard.ac.in.

Shahab Saquib Sohail is currently an assistant professor at VIT Bhopal University, India. His research interests include computational social science, recommender systems, and social network analysis. Contact him at shahabsaquibsohail@vitbhopal.ac.in.

Erik Cambria is currently a professor at Nanyang Technological University, Singapore. He earned his PhD through a joint programme between the University of Stirling and MIT Media Lab. His research focuses on neurosymbolic AI for explainable sentiment analysis in domains like social media monitoring, financial forecasting, and AI for social good. Contact him at cambria@ntu.edu.sg.

Amir Hussain is currently a professor at Edinburgh Napier University, UK. He obtained his PhD in Electronic and Electrical Engineering from the University of Strathclyde in 1996. His research interests include brain-inspired machine learning and cognitive technology for modeling and control of complex systems. Contact him at a.hussain@napier.ac.uk.