

A Zero-Shot Aspect-Based Sentiment Analysis of Public Perception Toward AI Chatbots

Naufal Andila Fauzan

naufal21002@mail.unpad.ac.id

Digital Business Program, Faculty of Economic and Business, Universitas Padjadjaran, Jawa Barat, Indonesia

ABSTRACT

The rapid development of AI chatbots has sparked public discussion on social media regarding their performance, ethical implications, and related concerns. While past studies primarily focused on individual chatbot model using traditional sentiment analysis, this study implements a novel application of Zero-Shot Aspect-Based Sentiment Analysis (ABSA) on 17,562 tweets mentioning AI chatbots such as ChatGPT, Bard (now Gemini), and DeepSeek, utilizing an efficient sentiment extraction method without supervised training. Six aspects were analyzed to understand the sentiment pattern and the results show the discussion was dominated by negative sentiment, with Bard receiving the most positive sentiment, potentially shaped by brand trust and user familiarity. On the other hand, DeepSeek and ChatGPT attracted more criticism, especially related to performance and bias aspects. This study offers data-driven suggestions for developers, including improving response accuracy to shape user trust, reducing biased output, and developing real-time discourse analysis. Future work should incorporate multiple platforms to avoid bias, analyze more AI chatbot models, and include temporal sentiment for broader insights.

Keywords: Aspect-Based Sentiment Analysis (ABSA); Zero-Shot Learning; AI chatbots; Public Perception; Social Media Mining

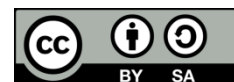
Article Info

Received : 11-02-2025

This is an open-access article under the [CC BY-SA](#) license.

Revised : 21-04-2025

Accepted : 30-06-2025



Correspondence Author:

Naufal Andila Fauzan
Digital Business Program,
Universitas Padjadjaran,
Sumedang, Jawa Barat, 40132.
Email: naufal21002@mail.unpad.ac.id

1 INTRODUCTION

Artificial intelligence, also known as AI, has become an integral part of our everyday lives. Generative Artificial Intelligence (GenAI) is one of the example of AI implementation. GenAI can produce data in various forms, including images, text, and more, using sophisticated generative algorithms. The algorithm works by learning the structures and patterns of the training data and then generating new data with similar underlying characteristics [1]. The influence of generative AI on industries and the public was profound, including the generating creative content, assisting in models and simulations building, and as a guidance in scientific exploration [2]. One-way users can leverage generative AI is by using an AI chatbot.

ChatGPT is one of the most popular AI chatbots. As of January 2023, two months after its release, ChatGPT had gained approximately 100 million active monthly users. When compared to other platform, such as Facebook and YouTube, the user growth on ChatGPT was particularly notable [4].

Following ChatGPT's success, in February 2023 Google launch its own AI chatbot named Bard [5]. Bard is an AI chatbot that functions similarly to ChatGPT, which mimics user conversations [6]. The

mechanism under Bard is supported by a machine learning model called the Language Model for Dialogue Applications (LaMDA) [5].

Moreover, another AI chatbot, DeepSeek, was released in December 2024. This China-made model shook the AI landscape with the release of its V3 model because their performance has exceeded GPT's in several aspects such as in the advanced reasoning model [9] [11].

Since its launch, numerous studies have explored the public perception of AI chatbots. Past research highlights the public attitudes toward ChatGPT, focusing on tweets sentiment and the occupation of the user by implementing XLM-T, unveiling that most sentiments were in a neutral tone, and the common topics range from AI, ChatGPT, and marketing to cybersecurity [4].

In addition, a past study leveraged a dataset of 21,515 tweets about ChatGPT that were labeled using VADER and the analysis were performed using transformer-based BERT, achieving 93.37% accuracy, and found that the sentiment of the discussion about ChatGPT was mostly positive because the ability to engage with users through conversation [26].

Furthermore, another study focuses on a specific domain of ChatGPT's implementation. In the healthcare sector, a survey shows that 43% of its respondents trust ChatGPT's ability to explain their health concerns [27]. In the education sector, a review paper aimed to gain insight into the impact and effectiveness of utilizing chatbots to understand better how to leverage AI chatbots effectively. The review highlights concern regarding the rise of AI in the educational sector, including the ethics, reliability, and accuracy of the information generated by AI chatbots [28].

While many research has examined the public perception of specific AI chatbots, a notable gap remains in understanding the public perception beyond a single AI chatbot [4]. Moreover, previous studies used only the traditional sentiment analysis approach to examine public perception [4], [14]. Sentiment analysis is defined as the process of recognizing and categorizing opinions expressed in text into specific sentiment categories, such as positive, neutral, or negative [18]. In the present day, with the growing need to identify sentiments at a more detailed level, traditional sentiment analysis methods do not have the capability to do such a task, as they are only capable of sentence and document-level analysis. It is required to implement a more advanced type of sentiment analysis, such as Aspect-Based Sentiment Analysis [21].

This study aims to fill the gap by examining public perception of sentiment toward multiple AI models on a broader level, rather than focusing on a single chatbot. Additionally, to provide a more detailed and in-depth understanding of public perception, this research employs Aspect-Based Sentiment Analysis (ABSA) to overcome the limitations of traditional sentiment analysis. This approach enables more detailed insight into how different AI models are perceived in a specific aspect, providing deeper insight.

To perform ABSA, the training dataset must be domain-specific. This is a challenge as the model that was trained in one domain cannot be applied to another [22]. ABSA also required a benchmark dataset, which is still limited, and for the dataset to be labeled with a manual annotation. Nevertheless, this has become a constraint to many because of how costly it is [23]. Zero-shot Learning (ZSL) comes to solve this issue as during the classification, it allows the process without the need for its instances to be labeled initially [24].

Many past studies have implemented ZSL to work with their ABSA models. For instance, a study implemented ZSL on datasets of online reviews to gain a deeper insight into aspects of the tourist experience and to understand the feedback, which provides valuable information for business decision-making [25]. The study uses the RoBERTa model as it has improvements in its training process, which enhances the overall model's performance. Moreover, a survey reveals that ZSL has been applied to many domains, including NLP. In the NLP field, ZSL has been used for language translation, understanding verbal speech, categorization of semantic utterances, extracting entities from the web, retrieving documents across documents, and identifying correlations between entities [24].

Despite the advancements in AI chatbots by many companies, flaws and errors remain inevitable. For example, AI chatbots from various companies, such as ChatGPT and Bard, often produce output that is not fact-based. This led to the coined term "AI hallucination," which refers to a product that implements GenAI, such as an AI chatbot, producing outputs that are completely nonsensical and untrue [12]. Biases in the dataset used to train the model could also potentially create problems, as they may produce outputs that discriminate against specific individuals and groups [13]. If these failures are not addressed, it could lead to devastating consequences that impact both individuals and society [12].

Therefore, this study aims to conduct an Aspect-Based Sentiment Analysis (ABSA) of ChatGPT, Bard, and DeepSeek by leveraging public discussions data on X (formerly Twitter). This analysis aims to offer a comprehensive view of the most discussed aspect and its sentiment polarity of AI chatbots while providing recommendation for chatbot development.

2 METHODOLOGY

This section covers the explanation of the methodology used for aspect-based sentiment analysis in this study. The whole process is structured into four main stages: data source, data cleaning and pre-processing, aspect extraction, and sentiment detection. A visual representation of the methodological workflow is illustrated in Figure 1.

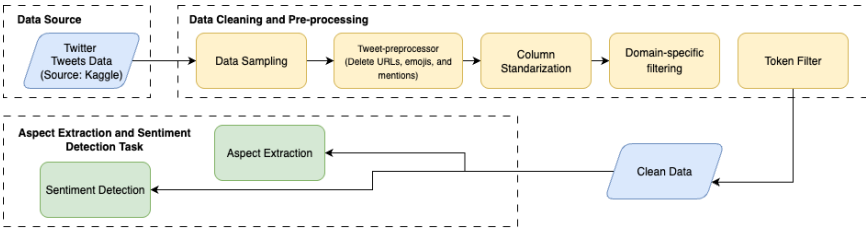


Figure 1. Overview of the Methodological Workflow

2.1 Data Source

The growing popularity of AI chatbots has led to an increase in public discussion about them. X (formerly Twitter) is one of the most popular and active platforms that hosts user-generated opinions and real-time user reactions on various topics and has been utilized by many researchers to perform such tasks [4], [18], [29], [30]. Therefore, tweets became an insightful source to understand people's opinions and the sentiment polarity behind it.

To understand the public sentiment towards AI chatbots, this research leverages three publicly available datasets sourced from Kaggle [31], [32]. Each of these datasets contained tweets that mention one of the three selected AI chatbots, including ChatGPT, Bard (now known as Gemini), and DeepSeek. The number of tweets of each dataset were visualized in Table 1.

Table 1. Summary of Tweet Dataset for Each AI Model

AI Model	Numbers of Tweets	Time Period
ChatGPT	151,000	February-March 2024
Bard	45,000	January-April 2024
DeBERTa	360,000	January 2025

The selection of the three datasets was designed to support three key research objectives. First, to understand the broad industry trends, we examine insights from the overall public sentiment and the most common aspect surrounding the conversations about AI chatbots. To achieve this, the three datasets were combined into a single entity, allowing for the analysis of broad public perception. Second, as we aim to compare the sentiment between three chatbots, we tries to enable a fair model comparison to ensure that no bias from data imbalanced take effect. A balanced dataset was assembled by sampling an equal number of tweets per chatbot allowing for a controlled analysis of sentiment across the three chatbots. Lastly, to deep dive into specific model, each dataset was analyzed individually to gain insight into chatbot-specific aspects that reflect user concerns. This approach enabled a clearer understanding of how each model is perceived without any influence of other chatbots.

2.2 Data Cleaning and Pre-processing

Given the large size of the collected dataset and considering the limitations of Google Colab as the computing environment used in this research, we first creating new column containing value of the AI model name to labels the tweet on each dataset, they all combined and randomly sampled into total of 20,000 tweets.

Moreover, since this study employs a transformer-based embedding model, such as DeBERTa and DistilBERT, data pre-processing is not necessary, as it can process raw data. This happens because all parts of the input text could hold a significant value for the sentiment and the aspect. However, in this research, tweets were employed as the input, and oftentimes, they also include noise and unnecessary characters; therefore, after sampling, the cleaning step was conducted. The tweet pre-processor library was used to clean the tweets by removing URLs, emojis, and mentions of another user's username. Then, the unnecessary columns in the dataset (including engagement metrics, author attributes, and language tags) were removed, and the remaining columns were renamed to ensure a smooth combining process.

To ensure that the dataset only contains tweets that are relevant to the main topic, this study conducts a domain-specific filtering check. This process manually checks a sample of tweets from the dataset and annotates the off-topic terms, then uses this reference to filter tweets. We found that some tweets contain terms such as "Crypto", "Ethereum", and "Bard protocol", which may be because the AI models we selected share

the same names as types of cryptocurrencies. The result of the filtering process helps to reduce noise that could distort the aspect and sentiment analysis. Following the cleaning and the pre-processing step, the dataset was reduced to 17,585 tweets that met the criteria.

However, as the transformer model has a maximum limit of 512 tokens per tweet, the dataset underwent a filtering process based on token length using AutoTokenizer from Hugging Face. This process tokenizes and calculates the total number of tokens per tweet, then tweets that contain more than 512 tokens are removed from the dataset to avoid errors during the analysis process. A small portion (0.1%) of the dataset was removed, which might include longer tweets that might contain richer context and more detailed opinions. Future work may consider a truncation or chunking approach to ensure that this insight is still included in the final dataset. This step reduced the dataset to a total of 17,562 tweets that only contain tweets with fewer than 512 tokens.

The cleaned dataset then used as for 3 different main objectives, first, to understand public sentiment, we use the overall dataset containing a total of 17,562 records, as it was previously unbalanced, to accurately reflect the public. For comparative analysis purposes, the data was stratified into 5,000 tweets for each of the AI models, resulting in a balanced dataset of 15,000 tweets. Lastly, for the model-specific deep dive, we took 1,000 tweets for each AI model. This approach was taken to ensure that the computational process feasible and more efficient, while maintaining the representation of each chatbot model.

2.3 Aspect Extraction and Sentiment Detection Task

To implement the Aspect-Based Sentiment Analysis on data in the form of user tweets regarding the topics of three AI chatbots, Zero-shot text classification was leveraged. Zero-shot methodology enables aspect and sentiment categorization in a more efficient way and saves time and cost compared to the traditional supervised learning pipelines, as it does not require any manual annotation [24].

In this research, aspect extraction is implemented as a multi-label classification problem. This approach assumes that a single tweet may contain multiple aspects. To address this, the cross-encoder/nli-deberta-v3-base model is leveraged. The selection of the model was based on past research and empirical evaluation.

For instance, a past research study conducted a sentiment classification using a COVID-19 tweet dataset reveals that DeBERTa achieved a 54% accuracy and 38% F1 score. This achievement surpasses baseline models, such as BERT (50% accuracy, 35% F1), Decision Tree (47%, 41%), and K-Nearest Neighbor (38%, 34%) (See Table 2) [33]. While this result is not specific to the ABSA task, it still provides valuable information that suggests DeBERTa's superior performance compared to other models, especially the baseline. Nevertheless, such evaluation on the ABSA task is recommended in future work to have a further understanding of the effectiveness of the zero-shot approach.

Table 2. Models' Performance Comparison on the Sentiment Classification Task

Model	Accuracy	F1 Score
K-Nearest Neighbor	38%	34%
BERT	50%	35%
DeBERTa	54%	38%

DeBERTa accuracy was achieved as the model implemented novel techniques, such as Distangled Attention (DA) and Enhanced Mask Decoder (EMD). While the BERT model works by combining the information inside the text and its position into one vector, DeBERTa, using the DA technique, will represent each token into two different vectors, for both word information and its position. Thus, the correlation between words was not only based on the information inside the text but also on the word's relative position in the text.

Moreover, similar to BERT, DeBERTa is also trained on a Masked Language Modelling (MLM) that tries to figure out which tokens were masked. However, during the decoding phase, DeBERTa took a different approach to incorporating its absolute positional information, enabling the ability to identify which token is the subject and the object within a sentence [34].

Furthermore, to find out which aspect to deep dive into the topic of AI chatbots, we pick a total of eight aspects based on the literature review of past studies surrounding public perception about the topics of AI chatbots. These aspects include "Bias", "Ethics", "Hallucination", "Performance", "Pricing", and "Other". Table 3 highlights the selected predefined aspects, as these aspects have been frequently mentioned in existing past studies on user concerns regarding AI chatbot systems.

Table 3. Predefined Model and Its Related Past Studies

Model	Related Past Studies
Ethics	[13]
Bias	[13]
Performance	[36]
Hallucination	[13]
Pricing	[39]

The model was then tasked to extract the most relevant aspect and record each of its instances. Moreover, an additional aspect, “Other”, was used to categorize outputs that either did not fit the predefined labels or had a low confidence score below 0.30. These approaches were implemented to avoid over-interpreting uncertain predictions.

In the case of the sentiment classification task, the sentiment polarity was assessed using a different model than the one used for aspect extraction. Sentiment classification will process the text in a single-label classification setup and categorize the text input into sentiment labels, such as positive or negative.

In this study, we leveraged a transformer-based model called distilbert-base-uncased-finetuned-sst-2-english, which enables a binary sentiment classification setup (positive vs. negative). This model was chosen based on its performance in past benchmark studies. Table 4 highlights findings from past studies; when compared to binary and multinomial tasks, the DistilBERT model achieves the highest accuracy compared to another transformer model [40].

Table 4. Performance Comparison on Binary and Multinomial Tasks

Model	Accuracy
SVM	71.18%
LSTM	74.40%
DistilBERT	81.02%

DistilBERT is known as one of the BERT pre-trained models that is more efficient in terms of size, velocity, and computational cost compared to its base model. This breakthrough is achieved by Knowledge Distillation (KD), a technique where the smaller model is trained to replicate the behavior of the base model. This results in the model being 40% more compact in size and 60% faster, while still retaining 97% of BERT’s performance [41].

To strengthen the justification for model selection, we conduct a manual annotation for 150 tweets with aspect and sentiment labels. These manually-annotated labels were then compared with the prediction from cross-encoder/nli-deberta-v3-base for aspect classification and distilbert-base-uncased-finetuned-sst-2-english for the sentiment task. The evaluation was conducted using various metrics, including precision, recall, and F1-score.

Table 5. Summary of Sentiment Classification Performance (DistilBERT vs Manual)

Sentiment Class	Precision	Recall	F1 Score
Negative	0.72	0.82	0.76
Positive	0.61	0.47	0.53
Accuracy Score			0.69

The result shows, as visualized in Table 5, that the sentiment classifier achieved an accuracy of 69%, with strong performance on the negative sentiment, reaching an F1 score of 0.72, while the positive sentiment shows an F1 score of 0.61. This suggests that the model is more sensitive to a negative term that appeared on the tweet. Positive tone on social media often gets mixed with sarcastic nuances, which makes it harder to capture when applying a binary sentiment classifier.

Table 6. Summary of Aspect Extraction tPerformance (DeBERTa vs Manual)

Sentiment Class	Precision	Recall	F1 Score
Bias	0.10	0.50	0.17
Ethics	0.60	0.60	0.60
Hallucination	0.00	0.00	0.00
Other	0.66	0.42	0.52
Performance	0.33	0.25	0.29
Pricing	0.15	0.33	0.21
Accuracy Score			0.38

On the other hand, as shown in the Table 6, the aspect extraction process achieved an overall 38% accuracy with the highest performance by ethic aspects, reaching 0.60 F1-Score while showing a balance between precision and recall. Meanwhile, the “Bias” category showed a recall of 0.50 but low on precision by only 0.1. This suggests that the model might be leaning towards overpredicting this category and resulting in many false positives. In contrast, hallucination was not detected at all with a 0 F1-score. This happened because the cues may be too subtle for the model to be able to detect them.

Overall, this evaluation provides empirical evidence that supports the application of the Zero-shot ABSA approach while acknowledging its boundaries. The result validates the calibration of the model and offers insight into which aspects are more effectively captured using a zero-shot setup and which may require extensive domain-specific fine-tuning in future work.

Moreover, as we handle large datasets, it is required to put the processing efficiency into factors to consider. To ensure these factors were fulfilled, this study uses 1000 rows per batch to leverage batch processing during the aspect extraction and sentiment detection to ensure the overall process works effectively and efficiently.

3 RESULTS AND DISCUSSION

This section presents the key findings and analysis from the sentiment classification and aspect extraction of tweets discussing AI chatbots. This finding offers a deep understanding of public perception towards AI chatbots and provides valuable insights into the aspect that truly matters based on users’ opinions and experiences.

3.1 General Public Sentiment on AI Chatbots

To explore the overall market trend and public opinions of AI chatbots, we utilized a dataset that combines three AI models. The dataset was unbalanced and meant to visualize the public discussion about AI models.

From the data that contained a total of 17,562 tweets, we observed that the sentiment was leaning towards negative sentiment across these three chatbot models. There are 11,993 tweets that were identified as negative sentiment, and the remaining 5569 tweets were identified as positive opinions. Figure 2 visualizes the representation of this sentiment distribution. This suggests that the overall public opinion on X towards the AI chatbot discussion was dominated by negative tweets and leaves room for the AI model developer for future development. Nevertheless, to understand which aspect becomes the biggest concern of the public, the aspect extraction was conducted.

As shown in Figure 3, the aspect extraction shows that the “Performance” aspect, with a total of 5451 tweets, was dominating, followed by the “Other” aspect with a total of 5439 tweets, and “Bias” with a total of 3776 tweets. This suggests that the public was concerned more with the performance and bias of an AI model and other aspects that were not included in the predefined aspect.

For the developer of the AI models, this section highlights the need to optimize performance, including reducing hallucination, increasing response accuracy, and making sure that the output remains fair and neutral. Moreover, as the “Other” aspect shows a high number of appearances, it shows that the public discussion also included broader topics outside the predefined topic. This shows that the evaluation of AI models needs to adopt a broader perspective based on public perception, and not only performance-based. Moreover, the “Other” aspect will be deep-dived in section 4.4.

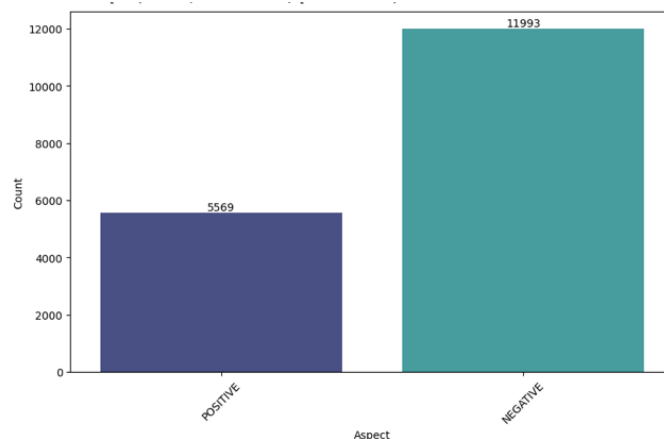


Figure 2. Sentiment Distribution of General Public Sentiment on AI Chatbots

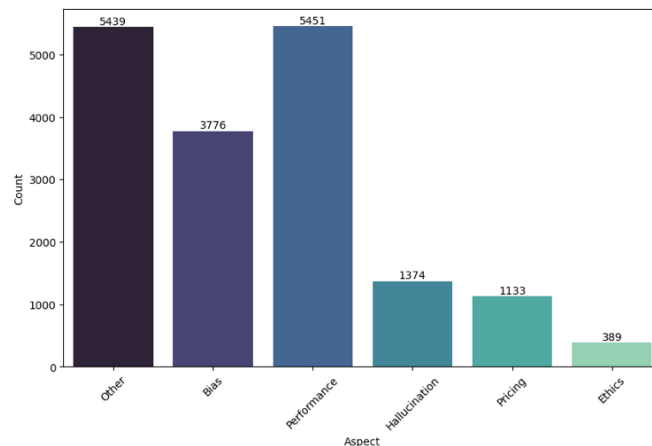


Figure 3. Aspect Distribution of General Public Sentiment on AI Chatbot

Comparative Analysis Between AI Chatbots

The fair comparison between the three chatbots was done by using a dataset that contains an equal number of tweets for each model. This approach ensures that the difference in the frequency of the aspect extracted, or sentiment, is not caused by the imbalances in the dataset.

The sentiment distribution across the AI models, as shown in Figure 4, reveals that all models were dominated by negative sentiment, which aligned with the findings in the previous section. It was also revealed that Bard received 562 negative tweets and 438 positive tweets (a ratio of 1.3:1), ChatGPT received 675 negative and 325 positive tweets (2:1), and DeepSeek received 722 negative and 278 positive tweets (2.6:1).

This comparative analysis revealed that, while every model was dominated by negative sentiment, the ratio shows that Bard is slightly balanced in negative-to-positive ratio, which may reflect the influence of the “Halo effect” as Bard was developed by Google (elaborated in Section 4.3). In contrast, DeepSeek received a higher negative-to-positive ratio, which might be influenced by skeptical commentary about DeepSeek’s performance or a lack of brand familiarity.

These results suggest the need for the developer to understand how user expectations and brand reputation affect the public sentiment towards their product. Addressing these issues requires more than technical improvements but also user engagement to build customer trust. Deeper sentiment analysis and aspect extraction of each AI chatbot’s model will be unveiled in the following sections.

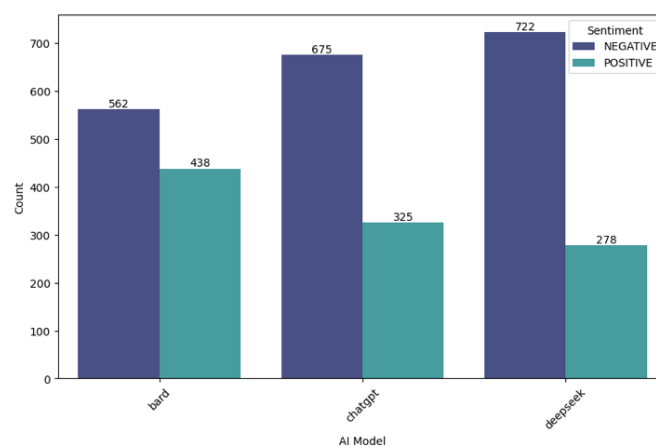


Figure 4. Sentiment Distribution Across AI Chatbots

3.2 Model-Specific Deep Dives

As the public visualization and comparative analysis have been done in the previous section, digging deeper into each model individually gave the ability to explore valuable contextual nuance. Aspect-based extraction was constructed to understand what people are deeply concerned about in each specific chatbot and to see how the sentiment around the conversation of each AI chatbot model.

3.2.1 ChatGPT

As visualized in Figure 5, the sentiment detection reveals that, from 1000 tweets, ChatGPT sentiment polarity was predominantly negative, with a total of 675 tweets, and the remaining 325 were positive.

A Zero-Shot Aspect-Based Sentiment Analysis...(Naufal Andila Fauzan)

Moreover, the aspect extraction process, as shown in Figure 6, reveals that the user was dominated by concerns on the aspects of “Other”, “Performance”, and “Bias”.

The “Other” aspect was predominant with general commentary that does not relate to the predefined aspect, for instance, a user tweeted, *“In addition to JPMorgan, other organizations have also blocked access to #ChatGPT. Verizon barred the #chatbot from its corporate systems, saying it could lose ownership of customer information or source code that its employees typed into ChatGPT,”* which reveals how some companies’ responses to ChatGPT in the workplace and the potential security risk associated.

However, if we investigate the ratio between the negative and positive tweets, it seems that the “Bias” performance has the highest ratio of negative to positive tweets. This suggests that when discussing bias-related topics in the context of ChatGPT, users were particularly critical. For example, a user tweeted, *“ChatGPT lists Musk, and Trump as ‘controversial’ noted personalities.”* This tweet shows that ChatGPT may be biased ideologically or politically in its responses by labeling specific figures as “controversial”. This user reaction indicates that they perceive the model’s responses as not heavily biased and not neutral, which might be influenced by the system behind the model, which might be constructed by a subjective filter or biased training data.

These findings suggest that OpenAI, as the ChatGPT developer, should consider putting attention towards fairness, transparency, and user trust. To address this issue, it is needed for them to analyze the model’s outcome, provide the user more features to explain how the chatbot constructs its responses, and improve the user feedback feature to allow the user to note on potential biased outputs. Moreover, as the “Other” aspect is dominating the public discussion about ChatGPT, it often discusses ChatGPT beyond the predefined aspect, which is mostly about technical metrics like societal impact and workplace responses. This highlights that to satisfy the user needs, improvement must also happen through effective communication and building user trust, not just focusing on model performance.

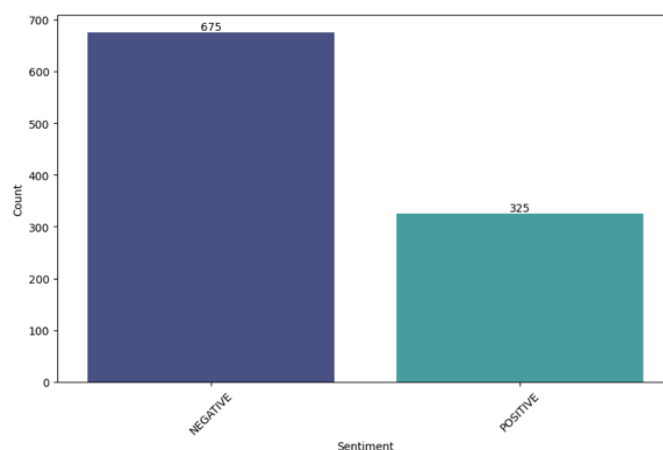


Figure 5. Sentiment Distribution of ChatGPT-related Tweets

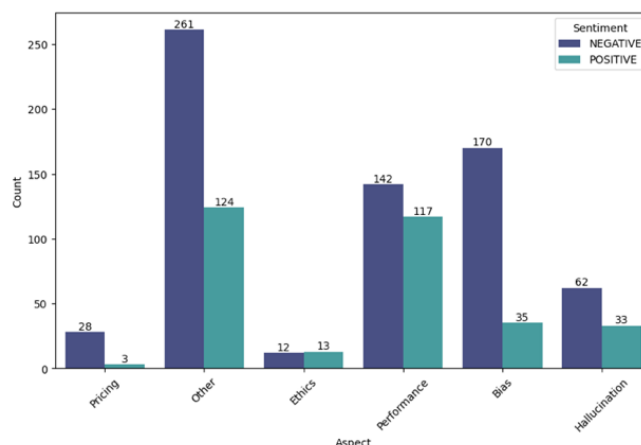


Figure 6. Aspect Distribution and Its Sentiment Distribution of ChatGPT-related Tweets

3.2.2 Bard

Released in 2023, Bard has gained attention due to the model being developed by one of the most influential companies, Google. Bard tweets sentiment distribution, visualized in Figure 7, shows that it was

dominated by negative sentiment, with a total of 562 tweets, and the remaining 438 were positive tweets. Moreover, a deeper look into the sentiment of the aspect-extracted, as visualized in Figure 8, reveals that most users were deeply concerned about three key aspects: “Other”, “Performance”, and “Bias”.

The “Other” aspect was filled with Bard-related tweets that do not belong to predefined categories of the aspect. For instance, one user tweeted, “*OpenAI: #ChatGPT is the future! Google: hold my #Bard*” reflecting a comedic or meme-based commentary about Bard’s releases. Another user tweeted, “*#Bard is a stupid name,*” expressing their opinion about their distaste for the name choice. These tweets highlight that Bard’s public discourse does not belong only to predefined aspects (which more falls into the technical side of the models) but also includes reactions to a wide range of topics that are still relevant to user sentiment.

Interestingly, the “Performance” aspect of Bard-related tweets was dominated by positive sentiment. This suggest that users are relatively optimistic about its capabilities, for example, a user expressed their excitement saying, “*Bard is here! Google's inevitable answer to ChatGPT has been released. Excited to compare the two side by side*”, and “*Google introduces #Bard, commits to staying bold with innovation and responsible in their approach to #AI.*”

This positive sentiment may be influenced by several factors outside its actual performance, including brand trust and ecosystem integration. Google’s track record in the tech world may create a halo effect where users link the positive qualities of a product based on the brand. One tweet says that “*...Google showing results from their new AI Bard without users having to click a single link...*” highlighting the excitement of integrating Bard on their search engine. Another asked, “*Anyone else looking forward to seeing how Google integrates Bard into search?*” This shows the anticipation of the new user experience.

These observations suggest that Bard’s sentiment tweets are not shaped only by its capabilities or model output, but also by brand trust and familiarity. For the Bard developer team, these results were a reminder to keep in mind the importance of not only improving technical performance but also strategically utilizing the brand trust that Google has while keeping up with user expectations and communicating how the model is positioned amidst their existing products.

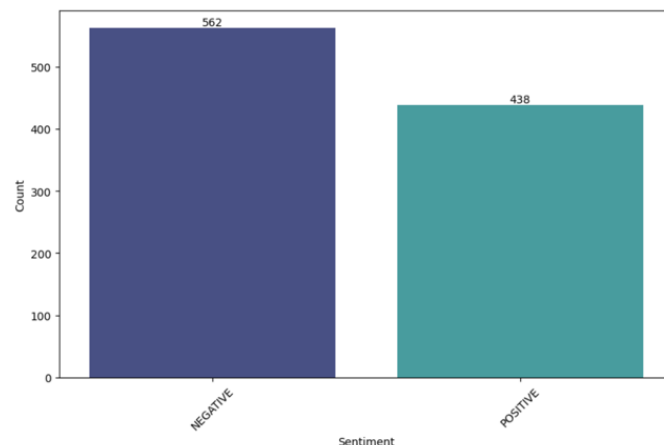


Figure 7.Sentiment Distribution of Bard-related Tweets

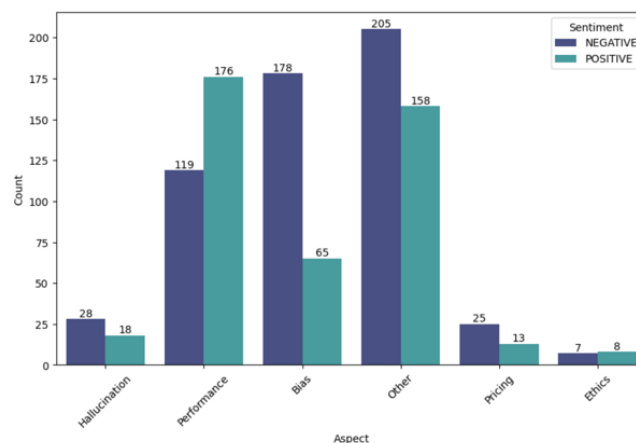


Figure 8. Aspect Distribution and Its Sentiment Distribution of Bard-related Tweets

3.2.3 DeepSeek

As a relatively new model released in late 2024, DeepSeek appears to receive more criticism and concern than praise from the public. As Figure 9 shows, from the total of 1000 tweets, it seems that the majority of DeepSeek tweets consisted of negative sentiment, with 722 negative tweets and only 278 positive ones. Moreover, the detailed breakdown of sentiment based on the predefined aspect is illustrated in Figure 10, highlighting that aspects that the public discusses are “Performance”, “Bias”, and “Other”.

Performance being the most aspect and its sentiment being majority negative suggested that the public discussion about DeepSeek performance attract more criticism than appraisal, for instance, a user throw critique by saying *“also ChatGPT is leagues better, the app was overall laggy and annoying to use compared to ChatGPT”* which a direct performance criticism towards DeepSeek when compared to ChatGPT. Another also compared both models, saying *“Slower, worse at maths, worse at reasoning, worse in complex tasks, simply cannot compete with the usefulness of ChatGPT. I used ChatGPT to summarise gigabytes of videos and dozens of documents into a table to print.”* This reaction may come from the fact that users had high expectations regarding DeepSeek as a new model, which is why it was compared to ChatGPT. As a result, this suggests that user sentiment is not only shaped by its performance, but also relative benchmarking against existing models in the market.

Interestingly, the “Bias” aspect also attracts significant criticism with a high negative-to-positive ratio of 4:1. This suggests that the user was overwhelmingly sensitive when discussing potential model content filtering and its ideological stances. For instance, a user tweets, *“Just look at the answers, this Chinese AI Chatbox is giving! On being asked about the Tianmen Square Massacre, #DeepSeek didn't reply. When asked, about South China Sea, it said it belongs to China.”* This finding shows concerns about content filtering as the model’s output appeared to be aligned with the Chinese government’s information policies, as it reflects China’s official stances on sensitive matters.

This insight shows that the public sentiment may be driven by the performance expectation of the user and distrust in the model output caused by bias. For DeepSeek’s developers, this highlights the need to improve the model’s responsiveness and transparency, highlighting the cross-cultural neutrality since the model was released globally. To address the user criticism, it might be needed for the developers to do benchmarking on existing AI model performance, refine the training data to reduce geopolitical bias and sensitive issues, and provide a direct disclaimer for restricted content to ensure improvement on user experience. This approach was meant so that DeepSeek could meet the standard set by existing AI models as current market leaders.

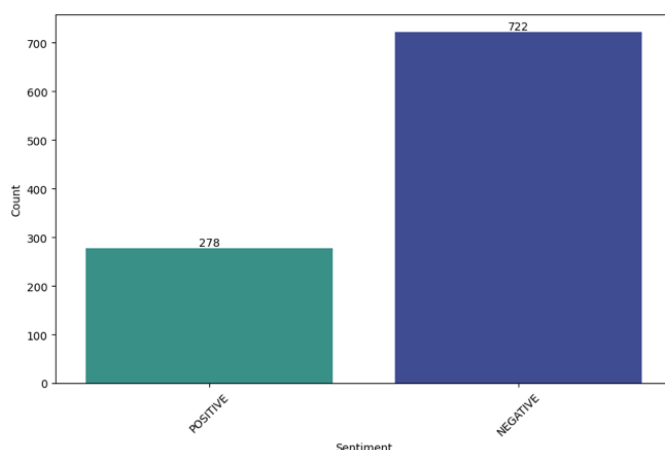


Figure 9. Sentiment Distribution of DeepSeek-related Tweets

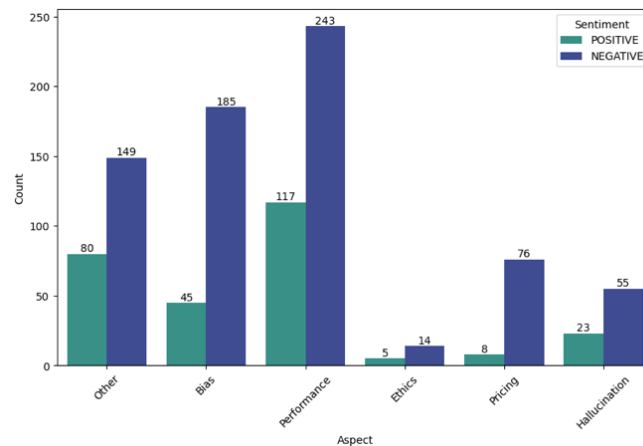


Figure 10. Aspect Distribution and Its Sentiment Distribution of DeepSeek-related Tweets

3.3 “Other” Aspect Sub-topic Exploration

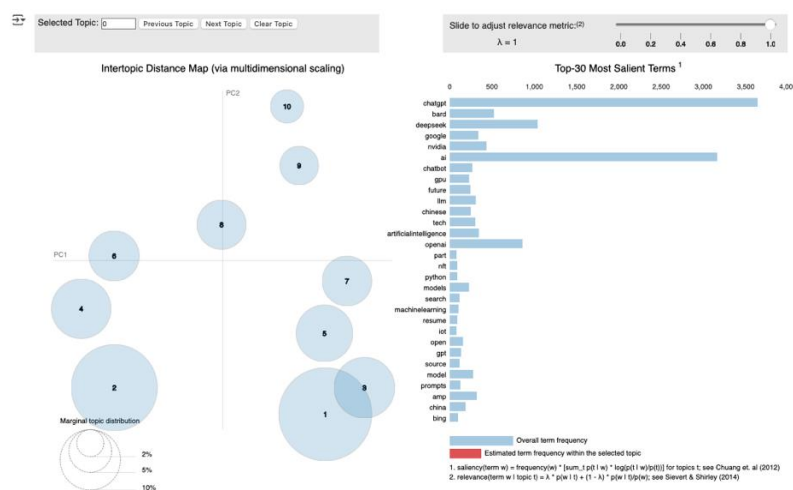


Figure 11. Result of Latent Dirichlet Allocation (LDA) on “Other” Aspect Tweets

Since the dataset that was used to understand the aspect that public discussion mainly talked about reveals that the “Others” category was the most dominant aspect, we explore the sub-topic to dig further by implementing LDA. This was done by first determining the number of topics appropriate by calculating the coherence score for values of k ranging from 2 to 10. The result shows that the highest score (0.455) was at $k = 10$, which was then selected as the number of topics for the LDA models.

Based on the LDA inter-topic distance map (see Figure 11), it reveals that most of the topics were well-separated, which indicates a meaningful. ChatGPT, Bard, DeepSeek, Nvidia, AI, Chatbot, and GPU were observed as the most salient terms included. This finding shows that, even though the tweets that belong in the “Others” category were related to the AI, it focuses on discussion outside the scope of the predefined model, which is still generally relevant to AI discourse. These tweets typically contain general commentary and discussion on AI and the related tools.

For example, a user tweeted, “Whether you love or hate generative #AI like the viral success of #ChatGPT, the technology is officially out in the wild, and Pandora’s Box has been opened.” This tweet directly addresses the popularity of AI by bringing ChatGPT as an example, but does not mention any term that relates to the predefined topic. Another user tweeted, “Our race should be using AI for agriculture and bettering our lives. Self-driving car is far-fetched. A PhD student should make a model that can oversee the growth of tomatoes and call out pest. Yes, to create that doesn’t need 28398GB GPU” which mentions AI and discusses its broad potential, but does not directly relate to the AI Models we are focusing on. This also reveals a limitation on the dataset, as tweets that did not correlate with AI models were included in the scraping process just because they mentioned AI in the sentences.

The LDA findings reveal some limitations in our dataset. First, as we observed that there are still off-topic tweets that are still included in the dataset, this might be caused by the tweets simply mentioning “AI” and using AI-related terms, which might confuse the tweets scrapper even when it did not discuss targeted AI

models. Because of that, future research might want to undergo an even more extensive data cleaning process or fix the X scraping criteria so that the data is only specific to AI models that we tried to understand. Second, this LDA reveals many subtopics inside the “Other” aspect that could potentially become the additional predefined aspects for future studies.

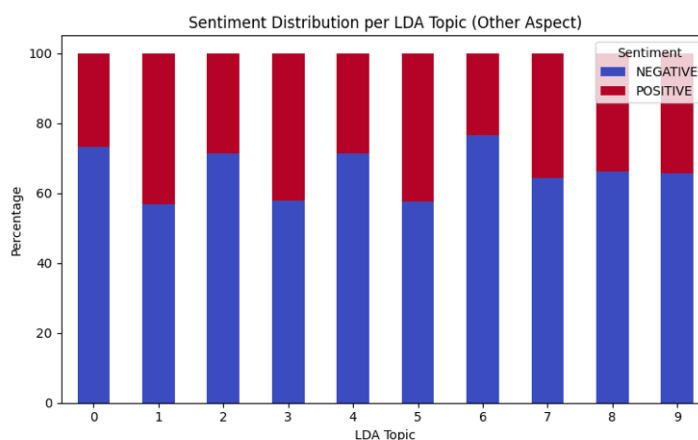


Figure 12. Sentiment Distribution per-LDA Topic

Moreover, to understand the tweets that belong to the “Other” aspect more deeply, we analyzed the sentiment within each LDA-derived topic. It revealed that negative sentiment dominates all 10 topics (see Figure 12), with Topic 6 showing the negative tweets dominating the sentiment distribution. This topic includes terms like “Nvidia”, “DeepSeek”, “AI”, “Market”, “China”, “Tech”, “Just”, “Chinese”, “Chips”, and “Like”. This suggests the biggest negative might come from the general commentary of the AI product. For example, a user tweeted “*Chinese AI app DeepSeek has overtaken ChatGPT as the top-rated free product on the Apple App Store in the US, UK, and China. Henry looks at what this might mean for the Western tech sector.*” shows that the user tweeted it react negatively on phenomena where DeepSeek dethrone ChatGPT as the top-rated app on the App Store.

Meanwhile, topic 5 was leaning towards more positive sentiment domination. This topic includes terms such as “AI”, “Chatgpt”, “Like”, “Data”, “GPU”, “DeepSeek”, “Human”, “LLM”, “Use”, and “power”. This shows that most tweets that belong to this topic may reflect public expectations about AI. For instance, a user tweeted “*For the people who thinks AI will replace Humans, yes it Will (I think) with more computing power and data it can replace humans (at least in theories) I Am Excited For it.*” unveils the positive expectation of the user on the future prospect of AI.

For developers, these finding shows that performing monitoring on the subtopics outside the main aspects, which are usually technical-related aspects, may offer deeper insights into user concerns that were not captured by the standard criteria. For instance, AI’s geopolitical context or the potential to implement AI in the common area may indirectly enable the adoption of chatbots and shape user trust. It’s recommended for the developer to build real-time public discourse analysis by implementing it on their feedback pipelines to enhance more proactive responses.

4 CONCLUSION

This study explored the public perception towards AI chatbots, including ChatGPT, Bard (now Gemini), and DeepSeek tweets, based on 17,562 tweets by implementing a Zero-Shot Aspect-Based Sentiment Analysis approach. The result shows that the overall public discussion of all models was dominated by negative sentiment tweets with a negative to positive ratio of 2:1. The aspect extraction highlights the most discussed aspects, which include “Other”, “Performance”, and “Bias”. A large number of tweets that belong to the “Other” category suggest that there is a huge number of concerns that cover broader aspects beyond the predefined technical aspects.

Bard became the most favorable model as it attained the smallest negative-to-positive ratio (1.3:1). This might be influenced by brand trust and its integration into other Google products. Meanwhile, both ChatGPT and DeepSeek seemed to attract more negative tweets, with sentiment ratios of 2:1 and 2.6:1, respectively. Notably, ChatGPT attracted criticism around the bias aspect, while DeepSeek’s negative sentiment was around its performance and bias, especially its content filtering on sensitive political matters.

These negative tweets suggest that user trust is built not only by the quality of the model responses, but also by how it handles addressing controversial issues.

The analysis of the “Other” aspect was done by implementing LDA, resulting in the discovery of hidden subtopics, for instance, geopolitical influence, hardware-related topics, and societal impact that were mostly classified as negative sentiment tweets. These topics were not included in the main predefined aspect, emphasizing the need to expand the chatbot evaluation beyond technical aspects.

These study findings suggest some actionable recommendations that the chatbot developer could implement. First, this study reveals that performance and bias appeared in a large number of tweets and both perceived as negative. It is needed to optimize the response accuracy and reducing the effect of bias on its outcome. It suggested that the developers implement a de-biasing feature and a direct disclaimer when the responses are potentially biased.

Second, as DeepSeek’s negative tweets were directed to its geopolitical-related responses, the developers need to put an urgency to ensure that the responses are neutral and transparent across a broad context since the models were released globally. Moreover, Bard’s positive sentiment shows that the public discourse was not only influenced by the model’s accuracy on its output, but also by user trust and experience. Developers can enhance user trust by prioritizing transparency in the model’s outcome and consider integrating with a system that feels familiar with the user. Lastly, as this study has demonstrated the application of Zero-shot ABSA, developers could implement it by developing a real-time sentiment monitoring that enables proactive tracking to support improvement based on real-time user concerns.

Nevertheless, this study faces several limitations. First, this study was only utilizing the tweets dataset and X users were known as more tech-savvy users and were demographically young. As a result, the findings may not generalize to broader populations. Second, while this study has already applied domain-based filtering in the pre-processing step, unfortunately, off-topic tweets are still observed in the dataset. Additionally, as the dataset we utilized only provides a brief visualization of public opinion during the data collection period, our study faces the temporal sentiment that could evolve over time.

Therefore, future studies address these limitations by implementing cross-validation by applying data from another platform or internet forum, such as Reddit, Quora, or other platforms to reduce platform-specific bias. In addition, a deeper exploration into different kinds of AI chatbot models and analyzing larger datasets could offer a valuable insight for future research. Improving preprocessing step by implementing more extensive off-topic filter would help to improve clarity and more targeted criteria for dataset. Furthermore, tracking the temporal evolution of sentiment would give a deeper insight on how the sentiment develop over time. Finally, as the “Other” aspect were observed having high frequency of uncategorized aspects, future work needs to expand the aspect definition to better capture more concern.

In conclusion, it is important to enhance user trust and experience by ensuring that the AI chatbot development and public sentiment are aligned. Large-scale ABSA provides a valuable tool to understand user feedback and can serve as strategic guidance for the continuous development of AI chatbot models.

CONFLICT OF INTEREST STATEMENT


The authors state no conflict of interest.

REFERENCES

- [1] S. Singh Sengar, · Affan, B. Hasan, S. Kumar, F. Carroll, and A. Bin Hasan, “Multimedia Tools and Applications Generative artificial intelligence: a systematic review and applications,” *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-20016-1.
- [2] B. Ramdurai and P. Adhithya, “The Impact, Advancements and Applications of Generative AI,” *International Journal of Computer Science and Engineering*, vol. 10, no. 6, 2023, doi: 10.14445/23488387/ijcse-v10i6p101.
- [3] P. Welsby and B. M. Y. Cheung, “ChatGPT,” *Postgrad Med J*, vol. 99, no. 1176, pp. 1047–1048, Oct. 2023, doi: 10.1093/postmj/qgad056.
- [4] R. Koonchanok, Y. Pan, and H. Jang, “Public attitudes toward chatgpt on twitter: sentiments, topics, and occupations,” *Soc Netw Anal Min*, vol. 14, no. 1, p. 106, 2024, doi: 10.1007/s13278-024-01260-7.
- [5] S. K. Singh, S. Kumar, and P. S. Mehra, “ChatGPT & Google Bard AI: A Review,” *2023 International Conference on IoT, Communication and Automation Technology, ICICAT 2023*, 2023, doi: 10.1109/ICICAT57735.2023.10263706.
- [6] D. Milmo, “Google trials its own AI chatbot Bard after success of ChatGPT.” Accessed: Feb. 18, 2025. [Online]. Available: <https://www.theguardian.com/technology/2023/feb/06/google-releases-its-own-ai-chatbot-bard-after-success-of-chatgpt>
- [7] S. Sutner, “ChatGPT bursts into Microsoft Bing as Google Bard rises.” Accessed: Feb. 18, 2025. [Online]. Available: <https://www.techtarget.com/searchenterpriseai/news/365530303/ChatGPT-bursts-into-Microsoft-Bing-as-Google-Bard-rises>
- [8] Daniel Newman, “Exploring The Ins And Outs Of The Generative AI Boom.” Accessed: Feb. 17, 2025. [Online]. Available: <https://www.forbes.com/sites/danielnewman/2023/03/14/exploring-the-ins-and-outs-of-the-generative-ai-boom/>
- [9] A. Rowe, “What Is DeepSeek? China’s New AI Is Now Open-Source.” Accessed: Feb. 18, 2025. [Online]. Available: <https://tech.co/news/what-is-deepseek>
- [10] K.-S. Huang, “China’s AI Shock? What DeepSeek Disrupts (and Doesn’t) – The Diplomat.” Accessed: Feb. 18, 2025. [Online]. Available: <https://thediplomat.com/2025/01/chinas-ai-shock-what-deepseek-disrupts-and-doesnt/>
- [11] S. Throne, “Putting DeepSeek to the test: how its performance compares against other AI tools.” Accessed: Feb. 18, 2025. [Online]. Available: <https://theconversation.com/putting-deepseek-to-the-test-how-its-performance-compares-against-other-ai-tools-248368>

- [12] V. Barassi, "Toward a Theory of AI Errors: Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI," *Harv Data Sci Rev*, vol., no. Special Issue 5, Nov. 2024, doi: 10.1162/99608F92.AD8EBBD4.
- [13] V. D. Kirova, C. S. Ku, J. R. Laracy, and T. J. Marlowe, "The Ethics of Artificial Intelligence in the Era of Generative AI," *J Syst Cybern Inf*, vol. 21, no. 4, 2023, doi: 10.54808/jsci.21.04.42.
- [14] C. Graham and R. Stough, "Consumer perceptions of AI chatbots on Twitter (X) and Reddit: an analysis of social media sentiment and interactive marketing strategies," *Journal of Research in Interactive Marketing*, vol. ahead-of-print, no. ahead-of-print, Jan. 2025, doi: 10.1108/JRIM-05-2024-0237.
- [15] S. Geman and M. Johnson, "Probabilistic Grammars and their Applications," in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 2015. doi: 10.1016/B978-0-08-097086-8.42161-6.
- [16] T. Huynh-The, Q. V. Pham, X. Q. Pham, T. T. Nguyen, Z. Han, and D. S. Kim, "Artificial intelligence for the metaverse: A survey," *Eng Appl Artif Intell*, vol. 117, p. 105581, Jan. 2023, doi: 10.1016/J.ENGAPPAI.2022.105581.
- [17] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0015-2.
- [18] A. Tanaltay, A. S. Langroudi, R. Akhavan-Tabatabaei, and N. Kasap, "Can Social Media Predict Soccer Clubs' Stock Prices? The Case of Turkish Teams and Twitter," *Sage Open*, vol. 11, no. 2, 2021, doi: 10.1177/21582440211004153.
- [19] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
- [20] S. Bhattacharya, D. Sarkar, D. K. Kole, and P. Jana, "Chapter 9 - Recent trends in recommendation systems and sentiment analysis," in *Advanced Data Mining Tools and Methods for Social Computing*, S. De, S. Dey, S. Bhattacharyya, and S. Bhatia, Eds., Academic Press, 2022, pp. 163–175. doi: <https://doi.org/10.1016/B978-0-32-385708-6.00016-3>.
- [21] S. Yin, "The Current State and Challenges of Aspect-Based Sentiment Analysis," *Applied and Computational Engineering*, vol. 114, pp. 25–31, Dec. 2024, doi: 10.54254/2755-2721/2024.18197.
- [22] L. Shu, H. Xu, B. Liu, and J. Chen, "Zero-Shot Aspect-Based Sentiment Analysis," Feb. 2022.
- [23] D. Van Thin, H. Quoc Ngo, D. Ngoc Hao, and N. Luu-Thuy Nguyen, "Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models," *Journal of Information and Telecommunication*, vol. 7, no. 2, 2023, doi: 10.1080/24751839.2023.2173843.
- [24] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," 2019. doi: 10.1145/3293318.
- [25] I. Nawawi, K. F. Ilmawan, M. R. Maarif, and M. Syafrudin, "Exploring Tourist Experience through Online Reviews Using Aspect-Based Sentiment Analysis with Zero-Shot Learning for Hospitality Service Enhancement," *Information*, vol. 15, no. 8, 2024, doi: 10.3390/info15080499.
- [26] S. R. et al., "Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach," *Information*, vol. 14, no. 9, p. 474, Aug. 2023, doi: 10.3390/info14090474.
- [27] S. Hussain et al., "Investigating public perception on use of ChatGPT in initial consultations prior to healthcare provider consultations," *Annals of Medicine and Surgery*, Oct. 2024, doi: 10.1097/MS9.0000000000002697.
- [28] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," 2023. doi: 10.1186/s41239-023-00426-1.
- [29] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *J Comput Sci*, vol. 36, 2019, doi: 10.1016/j.jocs.2019.05.009.
- [30] S. K. Kumar et al., "Stock Price Prediction Using Optimal Network Based Twitter Sentiment Analysis," *Intelligent Automation and Soft Computing*, vol. 33, no. 2, 2022, doi: 10.32604/iasc.2022.024311.
- [31] BwandoWando, "□ Tweets and Reactions on DeepSeek □," 2025. [Online]. Available: <https://www.kaggle.com/datasets/bwandoWando/tweets-and-reaction-on-deepseek-models>
- [32] sina tavakoli, "AI based platforms," 2023. [Online]. Available: <https://www.kaggle.com/datasets/sinatavakoli/ai-based-platforms?select=bard.csv>
- [33] Z. Wang, Y. Pang, and Y. Lin, "Large Language Models Are Zero-Shot Text Classifiers," Dec. 2023, doi: 10.48550/arXiv.2312.01044.
- [34] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *ICLR 2021 - 9th International Conference on Learning Representations*, Jun. 2021. Accessed: Jun. 19, 2025. [Online]. Available: <https://arxiv.org/abs/2006.03654>
- [35] N. D. Cohen, M. Ho, D. McIntire, K. Smith, and K. A. Kho, "A comparative analysis of generative artificial intelligence responses from leading chatbots to questions about endometriosis," *AJOG Global Reports*, vol. 5, no. 1, p. 100405, Feb. 2025, doi: 10.1016/J.XAGR.2024.100405.
- [36] A. Williams, "Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, p. 52, Sep. 2024, doi: 10.1186/s41239-024-00485-y.
- [37] D. Lee, M. Brown, J. Hammond, and M. Zakowski, "Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard," *Int J Obstet Anesth*, vol. 61, p. 104317, Feb. 2025, doi: 10.1016/j.ijoa.2024.104317.
- [38] R. L. Fleurence et al., "Generative Artificial Intelligence for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations: An ISPOR Working Group Report," *Value in Health*, vol. 28, no. 2, pp. 175–183, Feb. 2025, doi: 10.1016/J.JVAL.2024.10.3846.
- [39] L. Li, Y. Zhang, J. Sun, and Z. Yang, "Generative Artificial Intelligence Application Pricing: Navigating Upstream Strategy, Consumer Behavior, and Market Competition," 2024. doi: 10.2139/ssrn.4956548.
- [40] M. Valizadeh, P. Ranjbar-Noiey, C. Caragea, and N. Parde, "Identifying Medical Self-Disclosure in Online Communities," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 4398–4408. doi: 10.18653/v1/2021.naacl-main.347.
- [41] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, Accessed: Jun. 19, 2025. [Online]. Available: <https://arxiv.org/pdf/1910.01108>

BIOGRAPHIES OF AUTHORS

Naufal Andila Fauzan  is a final-year Digital Business student at Universitas Padjadjaran with a strong foundation in business analytics, data science, and consulting. He has international academic experience as an awardee of the Indonesian International Student Mobility Award (IISMA) and completed a semester at the University of Padova, Italy. Naufal has held multiple impactful roles, including Category Strategy Intern at TikTok Bytedance and Digital Marketplace Enabler Intern at Telkom Indonesia International, where he developed data dashboards, chatbots, and conducted strategic research. His technical proficiencies span Python, SQL, Tableau, and machine learning frameworks. He has also contributed to academia as a teaching assistant in programming courses and is currently working on research in sentiment analysis and stock price prediction. Naufal is passionate about leveraging data to drive business decisions and innovation. He can be contacted at naufal21002@mail.unpad.ac.id