Public Attitudes Toward ChatGPT During Early-Release Stage on Twitter: Different Perceptions in Health Versus Teaching Professionals

Creasa Chang, Hesong Zhao, Nora Li, Siyuan Ji, Xinlong Chen, Zimo Zhu

December 8, 2023

ABSTRACT

Our study investigates public perceptions of ChatGPT on Twitter during its initial release. Analyzing 39,054 English tweets posted within a month of ChatGPT release from Kaggle, we utilized sentiment analysis, topic modeling, and zero-shot classification to understand general and occupation-based attitudes toward ChatGPT. Our results indicate predominant positive and neutral sentiments in early-stage ChatGPT discussions, consistent across varied topics. Amongst "professionals," health and business sectors showed more positivity, whereas the cultural and teaching sectors exhibited skepticism, suggesting the influence of professional backgrounds on ChatGPT perceptions.

KEYWORDS: ChatGPT, Sentiment Analysis, Topic Modeling, Twitter, Natural Language Processing, Occupation

1 INTRODUCTION

Motivated by the rapid growth of ChatGPT, an advanced AI technology, our study aims to evaluate its public acceptance and uncover potential biases. The increasing prominence of Generative AI highlights the need for insights into societal impacts and ethical considerations for developers and policymakers (Dwivedi et al., 2023). Our research mainly analyzes Twitter sentiments towards ChatGPT—positive, negative, or neutral—delving into sentiment distributions across key discussion topics and variations based on users' occupations. This study is interesting since it leverages Twitter as a representative discussion platform to understand ChatGPT's social impact across different communities. We compared three sentiment analysis models, applied LDA for topic identification, and conducted occupation-based sentiment analysis. Our findings reveal that neutral sentiments predominate closely, followed by positive ones, and health professionals exhibit the most positive attitudes towards ChatGPT, providing crucial insights into public perceptions of this emergent technology.

2 RELATED WORK

In the previous work, Karanouh used VADER to analyze sentiment in ChatGPT-related headlines, finding 1,503 negative occurrences compared to 3,427 positives, indicating negative sentiments were approximately 56% less frequent than positive ones (Karanouh, 2023). Haque *et al.* focused on the ChatGPT early adaptors on Twitter and applied LDA, highlighting 'Entertainment and Creativity' as a highly positive topic, while 'Q&A Testing' and ‘Impact on Educational Aspects’ were less positive (Haque et al., 2022). Leiter *et al.* analyzed 300,000 ChatGPT tweets employing the XLM-Roberta model, noting a prevalence of neutral over positive sentiment (Leiter et al., 2023). Our research identifies a gap in existing studies by highlighting the top-performing model for sentiment analysis and exploring sentiments across various professions.

3 DATA

Our study leverages a Kaggle (Prata, 2022) dataset extracted from Twitter containing 39,052 English tweets from December 5th to December 25th, 2022, a period essential for capturing emergent public sentiment towards ChatGPT. The dataset's depth, with its 12 columns of user account information and demographic details, is the best fit for our sentiment analysis because it includes diverse user interactions during ChatGPT's early exposure. The 'Text' column's completeness, our primary analysis focus, ensures a dependable base for assessing public sentiments to ChatGPT. Despite missing and invalid entries in 'User Description,' the substantial 25,536 retained entries solidify the robustness of our occupational sentiment exploration. The lengths of 'Text' and 'User Description' display right-skewed distributions (Figures 1 and 2), predominantly 100-120 characters, highlight the succinct nature of

Twitter communications and are instrumental in revealing nuanced user sentiments and behaviors.

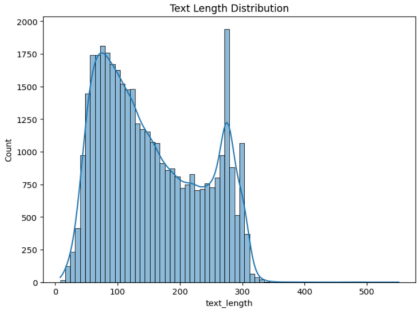


Figure 1: "Text" Length Distribution

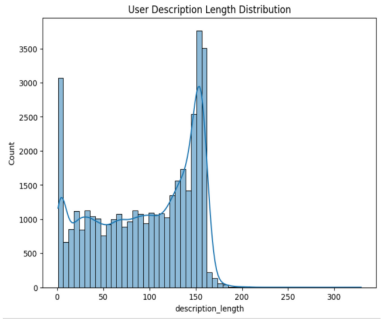


Figure 2: "User description" Length Distribution

4 APPROACH AND EXPERIMENTS

4.1 Comprehensive Sentiment Analysis of the Entire Tweet Dataset

Our sentiment analysis framework required thorough preprocessing for data quality, involving URL removal, tokenization, case normalization, stop word elimination, stemming, and spam detection to maintain data integrity. Based on established literature (Tan, Lee, & Lim, 2022), the sentiment was categorized into negative(-1), neutral(0), and positive(+1). Utilizing cleaned and structured data, we selected the Valence Aware Dictionary and sEntiment Reasoner (Vader), TextBlob, and the Robustly Optimized BERT Approach (RoBERTa) as our analytical tools, each noted for their capabilities within the sentiment analysis domain. Vader, with its compound scoring system, classifies sentiments as positive at scores above 0.05 and as negative below -0.05, attuned to the nuances of social media (Hutto & Gilbert, 2014). TextBlob employs polarity scores for sentiment definition, setting thresholds at > 0.05 for positive and < -0.05 for negative, thereby capturing a broad sentimental range within texts (Loria, 2018). RoBERTa is subjected to fine-tuning extensive tweet datasets for nuanced contextual sentiment analysis (Tan, Lee, & Lim, 2022). We employed Fleiss' Kappa (Landis & Koch, 1977), a quantitative metric for measuring inter-rater agreement, to validate the reliability of our sentiment assessments. In the validation phase, we engaged eight raters to label 100 tweet sentences across the three sentimental categories. This step was vital for validating unsupervised learning models and ensuring our methodologies align with human judgment. Through this structured approach, our sentiment analysis process was enhanced to be methodologically sound and robustly validated.

4.2 Comprehensive Topic Modeling and Classification Analysis of Tweet Corpus

This section adopts a multi-dimensional clustering approach to categorize user tweets into well-defined clusters. We utilized the Elbow Method, which graphically identifies the point of inflection, "the elbow," suggesting the optimal cluster count (Shi et al., 2021). To assess cluster integrity, we computed the Silhouette Score for a visual gauge of each point's cohesion within its cluster (Rousseeuw, 1987). Subsequently, we engaged Latent Dirichlet Allocation (LDA), a sophisticated generative probabilistic model for text corpora (Blei, Ng, & Jordan, 2003), to extract keywords from each cluster, clarifying their thematic essence. Post-clustering, we enlisted eight human raters (Section 4.1) for manual cluster labeling, affirming statistical and contextual relevance.

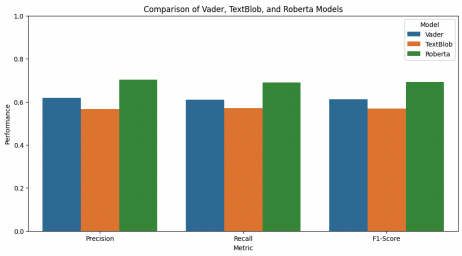
4.3 Professional Based Sentiment Analysis

Section 4.1's sentiment analysis reveals diverse sentimental responses. This section discusses the interplay between users' occupations and initial sentiments towards GPT. We employ zero-shot classification, suitable for unsupervised learning in classifying user descriptions (Yin, 2019). The International Standard Classification of Occupations (ISCO) by the International Labour Organization informs our occupational classification (International Labour Organization, n.d.). Comparing Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) performances, GPT emerges as superior for corpus annotations (Pawar & Makwana, 2022). BERT's limitations include a 512-token capacity (Feng et al., 2020) and fine-tuning inconsistencies (Zhou, 2021). Conversely, GPT-3 excels in varied NLP tasks, especially in zero-shot contexts, suitable for analyzing informal social media text (Zhang & Li, 2021). We applied OpenAI's API for zero-shot classification with ISCO to analyze and categorize 36,368 user descriptions, presented in a [JSON Output format]{"user_0": index}. After discarding invalid data, we classified professions and graphically illustrated the sentimental distribution among them, offering insights into the sentimental landscape across different occupations.

5 RESULTS

5.1 Comprehensive Sentiment Analysis of the Entire Tweet Dataset

In the text preprocessing, we successfully refined the 'Text' column by identifying and excluding 422 spam tweets, primarily marked by "MidJour" or "Giveaway." The inter-rater reliability, measured by Fleiss' Kappa, achieved a score of 0.52, categorized as a 'moderate agreement' strength (0.41-0.61) (Landis & Koch, 1977). The comparative evaluation of models indicated RoBERTa's supremacy (Figure 3), with its precision, recall, and F1-score significantly outperforming Vader and Textblob, substantiating the preeminence of BERT models, mainly credited to their bidirectional encoding framework (Ramalingam, 2022). When RoBERTa was applied to analyze the tweet dataset, the distribution of sentiment labels showed a predominance of neutral sentiment at 45%, followed by positive sentiment at 40.6% and negative sentiment at 14.4% (Figure 4). This trend supports the recent findings of Ogobuchi (Ogobuchi et al., 2023) for the primary occurrence of neutral sentiment in the landscape of social media sentiment analysis.



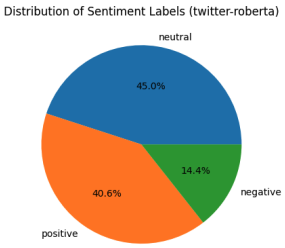
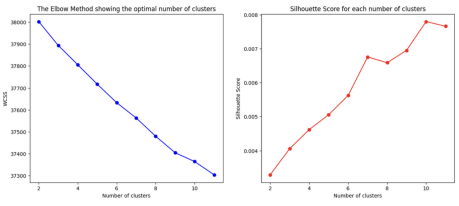
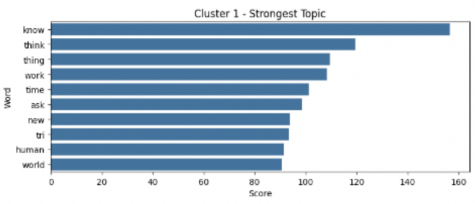


Figure 3: Comparison of Three Models' Performances **Figure 4: Sentiment Analysis Distribution in RoBERTa**

5.2 Comprehensive Topic Modeling and Classification Analysis of Tweet Corpus

In this analysis, we synthesized the Elbow Method and Silhouette Score to uncover the optimal number of clusters, ultimately selecting nine as the most coherent configuration. This decision, informed by the Elbow Method's initial indication (Figure 5), was corroborated by the Silhouette Score's affirmation of maximal cluster integrity at this count (Rousseeuw, 1987). Employing LDA, we extracted prominent thematic elements from each cluster; notably, our evaluators termed Cluster 1, with the most tweets number, as "ChatGPT's Cognitive Worldview" (Figure 6). Our sentiment distribution examination across the clusters discerned distinct sentimental trends: "Diverse Dimension" predominated in positive sentiments (76%), "Human-like Perception" in negative sentiments (21%), and "Artistic Authorship" featured a preponderance of neutral sentiments (54%) (Figure 7).





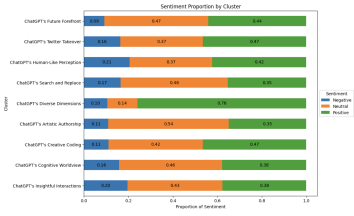


Figure 5: Elbow Method and Silhouette Score **Figure 6: Prominent Words in Cluster 1** **Figure 7: Each Topic's Sentiment Distribution**

5.3 Professional Based Sentiment Analysis

Upon exclusion of invalid data, our examination encompassed 25,536 entries. Within this refined dataset, 'professionals' constituted 17,501 entries, accounting for 68.53% (Figure 8). A trend of predominantly positive sentiments was observed across professions. Delving into the ISCO's 'professional' subcategories, 'health professionals' exhibited the highest positivity at 45% (Figure 9), potentially reflecting the beneficial perception of GPT models in healthcare contexts (Ray, 2023). Conversely, 'teaching professionals' registered the lowest positivity at 32% (Figure 9), which may signal the sector's unique instructional challenges and reservations towards GPT model adoption in educational settings (Huang, 2023). These disparate findings highlight the nuanced sentiment dynamics among various occupational groups on social media platforms.

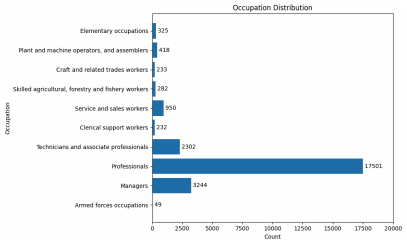
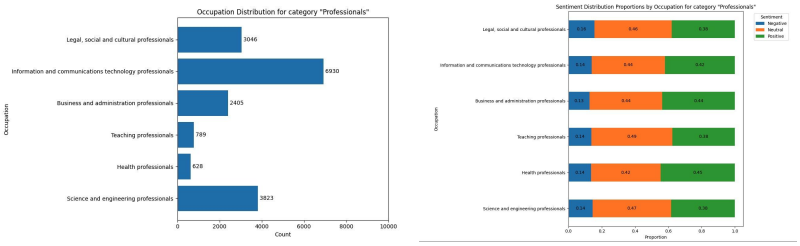


Figure 8: Tweet Counts by Occupation Distribution



Figures 9: Tweets Count and Sentiment Proportions in 'Professionals'

6 DISCUSSION

6.1 Limitations and Future Work

Our project faces three critical limitations. Firstly, the data lacks breadth in users' location and occupation details. The dataset, limited to English tweets from December 2022, may not accurately represent the broader population. Secondly, our occupation classification, employing the ISCO's 10-category system, fails to capture the full diversity of occupations. Lastly, the validation approach, mainly reliant on a small human group (eight individuals) and Extrinsic Evaluation Metrics, lacks depth and may benefit from additional, diverse methods (Cao et al., 2022).

To mitigate these issues, we propose expanding data collection over a longer period, encompassing multiple social media platforms and languages, and incorporating richer demographic information. A more refined sentiment scoring system, such as a five-point scale (-1, -0.5, 0, 0.5, 1), could offer greater precision. Additionally, breaking down broad occupation categories into specific roles, like distinguishing between doctors, nurses, and researchers within healthcare, might provide deeper insights. Finally, supplementing human labeling with sophisticated validation techniques like Inter-Annotator Agreement for Topic Modeling and Intrinsic Metrics for Topic Coherence could enhance the robustness of our model validation (Bruijn, 2020).

6.2 Insights and Contributions

Our experimental results reveal a nuanced reception of ChatGPT across various sectors. Significant findings include the dominance of professionals in the discourse and notable positivity in the Health and Business sectors, attributed to AI's practical applications in diagnostics and finance. However, the Cultural and Education sectors exhibited skepticism, mainly due to concerns about intellectual property and plagiarism risks associated with generative AI. This insight informs our recommendation for ChatGPT to engage with sector professionals in creating tailored strategies. Such collaboration can address ethical concerns, prevent AI misuse, and enhance user experiences, especially for those with neutral attitudes, thereby broadening ChatGPT's beneficial impact.

7 ETHICAL CONSIDERATIONS

In our study, we confront key ethical considerations, such as ensuring user consent and addressing privacy ambiguities in social media data analysis. This includes dealing with the reliability of user-provided details like location and occupation (Meserole et al., 2022). Methodologically, our text preprocessing risked overgeneralization, potentially omitting crucial context and affecting data validity. Using LDA for topic labeling presents methodological challenges, including subjective bias influenced by human interpretation. Additionally, employing ChatGPT for occupational classification highlighted socio-technical concerns, potentially propagating and perpetuating implicit representational biases inherent in its training data (Weidinger *et al.*, 2021).

8 CONCLUSION

Our study offers critical insights into ChatGPT's public perception on Twitter, highlighting generally “positive” and “neutral” sentiments across various sectors. Our topic modeling identified nine themes within the Tweets, including “Interactions and Perception,” “Artistic and Coding Production,” and “Future Prospects”. In addition, “professionals” dominated the discourse in our analysis, showing varied engagement levels. Particularly, the positive sentiment in Health and Business contrasts with skepticism in the Cultural and Education sectors, highlighting the diverse implications of AI. This research clarifies the diverse public perceptions of ChatGPT on Twitter, offering a scholarly perspective on its sector-specific impacts and sentiment.

REFERENCES

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Bruijn, L. de. (2020, July 18). *Inter-annotator agreement (IAA)*. Medium. https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3

Cao, Y. T., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., & Galstyan, A. (2022). *On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations*. ACL Anthology. https://aclanthology.org/2022.acl-short.62/

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Carter, L. (2023). *"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy*. International Journal of Information Management, 71(0268-4012), 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT Sentence Embedding. ArXiv. /abs/2007.01852

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). *"I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*. School of Computer Science, University of Adelaide, Australia. https://arxiv.org/pdf/2212.05856.pdf

Huang, Y. (2023). *Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching*. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1181712

Hutto, C., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

International Labour Organization. (n.d.). *Classification of occupation*. Retrieved from https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

Karanouh, M. (2023, August 3). *Mapping chatgpt in mainstream media to unravel jobs and Diversity Challenges: Early quantitative insights through sentiment analysis and word frequency analysis*. arXiv.org. https://arxiv.org/abs/2305.18340

Landis, J. R., & Koch, G. G. (1977). *The Measurement of Observer Agreement for Categorical Data*. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). *ChatGPT: A Meta-Analysis after 2.5 Months*. Natural Language Learning Group (NLLG), Faculty of Technology, Bielefeld University. https://arxiv.org/pdf/2302.13795.pdf

Loria, S. (2018). *textblob Documentation*. Release 0.15, 2(8), 269.

Meserole, C., Lee, N. T., Keller, J. B., & Darrell M. West, N. T. L. (2022, March 9). *How misinformation spreads on social media-and what to do about it*. Brookings. https://www.brookings.edu/articles/how-misinformation-spreads-on-social-media-and-what-to-do-about-it/

Ogobuchi, D. O., Udo, E. U., Rosa, R. L., Rodríguez, D. Z., & Kleinschmidt, J. H. (2023). *Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis*. Computers & Security, 135, 103476. https://doi.org/10.1016/j.cose.2023.103476

Pawar, C. S., & Makwana, A. (2022). *Comparison of BERT-Base and GPT-3 for Marathi text classification*. In Lecture Notes in Electrical Engineering (pp. 563–574). https://doi.org/10.1007/978-981-19-5037-7_40

Prata, Marília. (2022, December 20). *Chatgpt Tweets*. Kaggle. www.kaggle.com/code/mpwolke/chatgpt-tweets.

Ramalingam, S. (2022). *Sentiment analysis on Covid-19 vaccination reviews using BERT and comparative study with LSTM, VADER, and Text Blob models*. NORMA@NCI Library. Retrieved December 15, 2022, from https://norma.ncirl.ie/6639/

Ray, P. P. (2023). *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*. Internet of Things and Cyber-Physical Systems, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shi, C., Wei, B., Wei, S. et al. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Journal of Wireless Communications and Networking*, 2021(31). https://doi.org/10.1186/s13638-021-01910-w

Tan, K. L., Lee, C. P., & Lim, K. M. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access*, 10, 21517-21525. https://doi.org/10.1109/ACCESS.2022.3152828

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., . . . Gabriel, I. (2021). Ethical and social risks of harm from Language Models. ArXiv. /abs/2112.04359

Yin, W. (2019, August 31). Benchmarking Zero-shot text Classification: Datasets, evaluation and Entailment approach. *arXiv*. Retrieved from https://arxiv.org/abs/1909.00161

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, 1(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhou, Y. (2021, June 27). A closer look at how fine-tuning changes BERT. *arXiv*. Retrieved from https://arxiv.org/abs/2106.14282