International Journal of Computer Techniques – Volume 12 Issue 5,
September - October - 2025

# ChatGPT and NLP as Tools for Unstructured Data Analysis

[1]Rakesh Rohan Budige

[1]Department of Computer Science, University of Illinois Springfield, IL, USA

Corresponding Author: Rakesh Rohan Budige, Email: rakeshrohanbudige@outlook.com

Abstract

Unstructured data that makes up almost 80–90% of the world's digitized information poses great challenges in extraction, interpretation, and decision-making because of the lack of a pre-defined format and the complexity of semantics. Natural Language Processing (NLP) has been a pivotal discipline to overcome the challenges and has itself maturated from rule-driven systems to transformer centered architectures that can comprehend the context of the language. Of these developments, ChatGPT is a formidable generative model that integrates deep learning with linguistic reasoning in order to comprehend and synthesize unstructured data at a very large scale. The present paper theorizes the role of ChatGPT and NLP in the analysis of unstructured data and outlines the main applications in business intelligence, healthcare, engineering, and cyber-security spaces. Advantages are in terms of scalability, contextual precision, and flexibility with the shortcomings ranging from hallucination to bias and domain adaptation challenges. The paper concludes the discussion by charting the future of infusing ChatGPT with explainable AI, knowledge graphs, and multimodal systems. All in all, the present paper theorizes the role of NLP and ChatGPT in terms of deriving actionable knowledge from unstructured data.

Keywords: ChatGPT, Natural Language Processing (NLP), Unstructured Data Analysis, Transformer Models, Artificial Intelligence Applications

## 1. Introduction

The information era has witnessed unprecedented growth in the creation of unstructured information in the form of free-form text, social media content, emails, transcriptions of audio, images, and technical reports. Structured sets of data, in that they fit into table-like structures and are well-formatted to be consumed by typical analytical software, stand in striking contrast with unstructured information that is necessarily complex, context-rich, and semantically rich. Extraction of insight from this type of information is central to making decisions in applications ranging from business and medicine to engineering and public policy. Conventional machine learning models almost always disappoint in the capability of capturing the vagueness, variability, and finesse that is present in natural language [1].

Natural Language Processing (NLP) has been the go-to enabler of resolving these challenges by offering tokenization techniques, entity recognition, sentiment analysis, summarization, and

contextual embeddings. The most recent enhancements in transformer architecture have entrenched NLP more in the management of humongous unstructured data with unprecedented precision hitherto unseen. Riding this trajectory of development is the innovation of the ChatGPT that unifies deep learning and human-like linguistic reasoning for the production of coherent and context-conscient output [2].

Notably, burgeoning study supports the practical viability of similar models. Kadam and Pitkar (2025) showed that demand forecasting precision is increased and the efficiency of decision making and supplier communication in supply chain management is increased using ChatGPT while it efficiently processes unstructured data like supplier emails and market reports and customer inquiries [3] . Their results illustrate the feature of the generative and interpretative power of ChatGPT going beyond structured analytics and serving as a multipurpose tool for the processing of unstructured information. Expanding from this basis, the current paper investigates the theoretical basis of ChatGPT and NLP in unstructured data analysis in various disciplines with a focus on advantages and challenges and future prospects.

## 2. Understanding Unstructured Data

Unstructured data is that information that does not conform to a pre-defined model or schema and cannot be neatly organized into typical rows and columns. It comprises items such as social messages, comments from customers, emails, audio or video files with transcriptions, and articles from researcher reports. Structured data in the form of relational databases can be directly executed and processed in comparison to unstructured data that is often characterized by its irregularity, vagueness, and reliance upon context. It is a prized mine of information and a big headache to companies that seek to extract it in a systematic way.

The strength of unstructured data lies in the ability it has to record fine expressions of human behavior and dense contextual relationships that cannot be encoded in structured forms. Customer comments, say, may contain sentiment, intent, or dissatisfaction that cannot be taken into account in quantitative sales data alone. Similarly, engineering reports or clinical notes often contain long narratives that are required to take precise decisions. Unstructured information is thus a valuable supplement to structured data that brings a richer understanding of phenomena in the natural world.

But processing unstructured information at a large scale is fraught with a range of difficulties. Textual data typically encompasses polysemy, sarcasm, or specialized domain terms that are difficult to interpret. Multimedia or audio information raises complexities further through the necessity of multimodal fusion. Further, the sheer amounts of unstructured information daily produced, estimated at being in zettabytes, requires higher-level tools that can deal with information at an efficient level without losing contextual semantics. These considerations

therefore make a case for advanced methods of the NLP and the like kind of generative models of the chatGPT type that deal with variability, indeterminacy, and big-data streams [4].

3. NLP Foundations for Unstructured Data Analysis

Natural Language Processing (NLP) is an art of computer science that is focused on enabling machines to understand, analyze, and generate human language. It is grounded in the core tasks that make up unstructured data analysis, including tokenization, part-of-speech marking, named entity recognition (NER), sentiment analysis, and topic modeling. These processes make it possible to transform raw textual inputs into structured formats that computer models can intelligibly process and understand. Through the bridging of the communication-human and processing-machine gaps, NLP avails the possibility of organizations deriving actionable information from the torrential amount of textual information that pervades the world today.

Historically, NLP developed from rule-based systems that depended on handcrafted dictionaries and grammars to statistical models that incorporated probabilistic reasoning. Deep learning created a paradigm shift with the capability of learning language patterns from very large sets of data instead of pre-specified rules. Transformer-based architectures like BERT and GPT are the newest development in terms of using the power of self-attention mechanisms to comprehend long-range dependencies in a language. These models offer contextual embeddings that not only comprehend the individual word but also comprehend the meaning of the word in the context of the text around it and achieve a very high level of accuracy in summarization, machine translation tasks, and question-answering tasks [5].

Despite these advances, a variety of shortcomings remains. NLP models inherently underperform with domain-Specific Language and require fine-tuning to support specialized sectors like medicine or engineering. NLP models are also expensive in terms of computing, requiring high power and storage capacity. Additional challenges of bias and lack of interpretability and data protection also restrain wider adoption. But the trajectory of NLP development unveils the transformative ability of NLP in the interpretation of unstructured data, particularly in conjunction with generative systems of AI like ChatGPT that shift from analysis to reasoning and synthesis [6].

4. ChatGPT as a Tool for Unstructured Data

ChatGPT is an OpenAI generative pre-trained transformer that is arguably the most advanced tool of unstructured data analysis. Built upon massive text corpora and fine-tuned to a contextual level of comprehension, it employs the transformer architecture's self-attention in order to glean relationships between words and phrases at long ranges of text. Unlike NLP models before it that were only task-specifically trained, the more generalized ChatGPT is in a position to summarize, translate, analyze sentiment, and chat interactively with very little additional input. Through the

virtue of that flexibility, it is particularly well-suited to the kind of unstructured data analysis where nuance and context are in high demand and that is where a great deal of unstructured data resides.

One of the typical strengths of ChatGPT is the capability of not only parsing but also of synthesizing information and generating well-structured and context-aware output. It is capable of summarizing lengthy technical reports, summarizing research articles in salient insights, or clustering customers' comments into thematic sets. Its conversational interface is also a convenience facilitator that allows the users to question unstructured datasets in natural-language format and retrieve insightful interpretation back from it. When compared with the typical NLP pipelines involving numerous models per different task, the ChatGPT encompasses the entire suite of these operations under a consistent and multipurpose framework that reduces complexity and upholds efficiency.

Theoretical applications span industries. In a commercial enterprise, ChatGPT could read hundreds of thousands of customer reviews to identify germinating product trends. In medicine, it could read physician notes to support diagnosis or summarize clinical trial results. In manufacturing and engineering, ChatGPT can read design documents, maintenance records, and failure reports to identify patterns and support decision-making. Such diversity speaks to the power of a scalable unstructured data analysis tool that underlines structured analytics with depth of analysis, contextual reasoning, and dynamic real-time adaptability.

5. Applications Across Domains

Transformative capability of ChatGPT and NLP in unstructured data processing exists in many industries where existing approaches are either incapable of providing nuance or cannot achieve scalable business outcomes. By virtue of the power of contextual understanding, they create fresh opportunities for companies to find insights, automate tasks and facilitate better decision making.

Business Intelligence: One of the most immediate applications is in market and customer sentiment analysis. ChatGPT can sift through thousands of comments online, tweets, or questionnaire responses to discover patterns in the behavior of consumers. More than sentiment analysis, it is in a position to cluster answers into valuable sets, with product quality, say, or with promptness of service or prices, so that firms get to concentrate on where they are in the best position to innovate and improve. It unlocks market trends insight that is otherwise trapped in unstructured sets of data [7].

Healthcare: In medicine, unstructured data is comprised of clinical notes, articles of research, and patient interactions. ChatGPT is able to summarize long medical literature, extract valuable results for practitioners, and even extract risk factors from reports submitted by physicians.

These capabilities promise to hasten evidence-based decision-making, reduce administrative burden, and support early disease detection in conjunction with structured health records [8].

Engineering and Manufacturing: These disciplines produce vast amounts of unstructured information in the form of technical manuals, design spec sheets, maintenance records, and failure reports. Such documents can be analyzed using ChatGPT to spot patterns of recurring issues, anticipate future failures, and simplify design validation. For example, analyzing long lists of incident records might uncover patterns of component wear or manufacturing workflow inefficiencies that engineers can correct proactively instead of reactively. Further, previous studies of digital transformation of supply chain quality management identified that the combination of AI, IoT, and big data analytics greatly enhances the level of transparency, reliability, and quality control of engineering processes. Expanding on the logic from the previous studies, ChatGPT adds more value to the utility of the digital tools through the provision of advanced natural language that is capable of summarizing technical narratives, underpinning predictive strategies of maintenance, and simplifying information sharing across multidisciplinary teams [9].

Education and Research: ChatGPT can support researchers in rapidly collating literature across disciplines, providing concise extracts, and identifying thematic intersections across studies. In education, it could find applications as a brilliant grader, intelligent system of feedback, and personalized learning track system through natural-language interactions with student work [10].

Cybersecurity and Risk Management: Unstructured data from incident reports, threat advisories, or even online forums often hold early indicators of security vulnerabilities. By analyzing this information using ChatGPT, you can reveal potential risks, support forward-looking mitigation steps, and enhance protection against disruptions [11].

These diverse examples illustrate why NLP and ChatGPT extend far beyond theoretical constructs and come forward as do-it-all instruments for extracting actionable insight from the sheer ocean of unstructured data generated every day.

## 6. Benefits and Challenges

The pairing of ChatGPT and NLP in unstructured data analysis has a few prominent strengths. First, the tools are scalable and permit an organization to analyze millions of documents, papers, or customer reviews in a few seconds. The capability to analyze a whole lot of information with very limited human intervention translates into a great deal of time and money saved. Second, the models are contextually aware and move beyond the keyword level of searches to extract sentiment, intent, and fine-grained meaning in the document. The quality of results is enhanced and more informed decisions are taken. Third, the versatility of ChatGPT lends itself to being

used in a very wide range of industries—from business intelligence to engineering—without the need for rigorous retraining and therefore a cross-industry tool of great worth.

Even with these benefits come challenges. Perhaps the foremost of these is that of precision and reliability. Though chatGPT produces highly naturalizing responses, it may in turn produce "hallucinations," or responses that are correct-sounding but unfounded in fact. These are risks in areas of healthcare or engineering, where precision is paramount.

Another problem is the protection and security of data. Most unstructured data sets such as medical records or business emails hold confidential information. Installation of the AI software without sufficient protection risks exposure of the organizations to non-compliance and moral challenges.

Finally, issues of bias, interpretability, and domain adaptation restrain adoption. Models from large, generic datasets might misinterpret specialized vocabularies and transfer biases or produce inconclusive outcomes. Furthermore, their "black-box" nature makes users skeptical of results at a fundamental level. These factors require the implementation of a hybrid approach that combines the power of AI with human oversight in order to obtain the maximum benefit while limiting risk.

## 7. Future Directions

NLP in the future of unstructured data analysis lies in the combination of the strengths of big models and tailoring in the domain of application. Another promising direction is the combination of knowledge graphs with ChatGPT to improve grounding in the facts and reduce hallucinations. By anchoring the responses to verified databases, models will obtain more transparent and reliable outcomes in the higher-stakes applications of healthcare and engineering.

Another is domain-specialized fine-tuning of the type that is possible with ChatGPT for specialized vocabularies and situations. In medicine, for example, fine-tuning of clinical records can improve diagnostic support, and in engineering, design norms and technical manual fine-tuning can improve the capacity for definitely interpreting complex reports. Such specialized fine-tuning would significantly improve the repertoire of applications of the generative models in the workplace.

Further, the development of explainable AI (XAI) and multimodal systems will decide the future of the analysis of unstructured data. Explainability will allow conclusions to be understood and improve accountability and trust. Multimodal models that are capable of processing natural language but also images and video and audio will provide a more complete and richer analysis of the world data. These directions of development jointly imply a future where NLP and ChatGPT will not only be powerful but also trustworthy, specialized, and integrated into decision systems in a smooth manner.

## 8. Conclusion

Unstructured data has been one of the most valuable and under-leveraged assets of the modern digital era. Its variability, complexity, and sheer magnitude demand advanced tools that are capable of drawing information that is beyond the capability of conventional methods to access. NLP and transformer-based models in particular, like ChatGPT, register a revolutionary jump ahead by providing contextual understanding, creation of language and mass-scale information creation.

This paper has examined the contribution of ChatGPT and NLP to the analysis of unstructured data with a focus upon both promise and constraint. Following recent work like that of Kadam and Pitkar (2025), where ChatGPT had been shown to be successful in enhancing the forecasting and decision-making in supply chain management, it is well that these models transcend structured analytics into more wide-ranging, context-rich applications [3]. Areas of application range from business intelligence through healthcare and engineering and education and cybersecurity, highlighting the applications across industries of generative AI.

But the deployment of the same has to accompany sharp considerations of precision, confidentiality, and moral implications. Issues of hallucinations, biasing, and interpretability seek hybrid solutions that integrate the efficacy of AI and human oversight. In the years ahead, advances in knowledge graph integration, domain fine-tuning, explainability, and multimodal analysis hold great promise of enhancing reliability and expanding the horizon of unstructured data analysis. In the long term, ChatGPT and NLP cannot be a substitute for human knowledge but excellent synergists that offer companies the ability of extracting actionable knowledge from vast repositories of unstructured information.

## 9. References

1. Mohapatra DP, Thiruvoth FM, Tripathy S, et al (2023) Leveraging Large Language Models (LLM) for the Plastic Surgery Resident Training: Do They Have a Role? Indian Journal of Plastic Surgery 56:. https://doi.org/10.1055/s-0043-1772704

2. Zhang H, Shafiq MO (2024) Survey of transformers and towards ensemble learning using transformers for natural language processing. J Big Data 11:. https://doi.org/10.1186/s40537-023-00842-0

3. Kadam, A., & Pitkar, H. (2025). Optimizing Supply Chain Management with ChatGPT: An Analytical and Empirical Multi-Methodological Study. *Journal of Computer Science and Technology Studies*, *7*(1), 337-350. https://doi.org/10.32996/jcsts.2025.7.1.25

4. Yang H, Xiang K, Ge M, et al (2024) A Comprehensive Overview of Backdoor Attacks in Large Language Models Within Communication Networks. IEEE Netw 38:. https://doi.org/10.1109/MNET.2024.3367788

5. Ahmed SF, Alam MS Bin, Hassan M, et al (2023) Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif Intell Rev 56:. https://doi.org/10.1007/s10462-023-10466-8

6. Naseem U, Dunn AG, Khushi M, Kim J (2022) Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics 23:. https://doi.org/10.1186/s12859-022-04688-w

7. Fang Y, Li X, Thomas SW, Zhu X (2023) ChatGPT as Data Augmentation for Compositional Generalization: A Case Study in Open Intent Detection. In: FinNLP-Muffin 2023 - Joint Workshop of the 5th Financial Technology and Natural Language Processing and 2nd Multimodal AI For Financial Forecasting, in conjunction with IJCAI 2023 - Proceedings

8. P. Ganguly and R. Garine, "Deep Learning for 30-Day Hospitalization Prediction in Dialysis Patients," *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, 2024, pp. 792-796, doi: 10.1109/ICDICI62993.2024.10810852.

9. Akash Abaji Kadam, Tejaskumar Vaidya, & Subba rao katragadda. (2025). Digital Transformation of Supply Chain Quality Management: Integrating AI, IoT, Blockchain, and Big Data. *Journal of Economics, Finance and Accounting Studies* , *7*(3), 41-49. https://doi.org/10.32996/jefas.2025.7.3.5

10. Mahapatra S (2024) Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. Smart Learning Environments 11:. https://doi.org/10.1186/s40561-024-00295-9

11. Humphreys D, Koay A, Desmond D, Mealy E (2024) AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI and Ethics 4:. https://doi.org/10.1007/s43681-024-00443-4