7th International Scientific Conference ITEMA 2023 – Selected Papers
DOI: https://doi.org/10.31410/ITEMA.S.P.2023.73

Application of Large Language Models for Text Mining: The Study of ChatGPT

Miloš Živadinović[1]

Received: December 21, 2023
Accepted: April 12, 2024
Published: May 28, 2024

Keywords:
Large language models;
ChatGPT;
Text mining

Abstract: *The appearance of Large Language Models (LLMs) has brought advancements in natural language processing (NLP), making it more available to everyone. This paper examines the application of LLMs in text mining, with a focus on ChatGPT by OpenAI. The author provides a brief overview of LLMs, highlighting their structure and training techniques, as well as parameter tuning. Utilizing ChatGPT as an example of an LLM, this paper identifies the model's capabilities and constraints in extracting insights from textual data. Based on the author's findings, they suggest several applications of LLMs for text mining that provide better text comprehension and set the tone for further research.*

1. INTRODUCTION

The concept of generative artificial intelligence has been expanding for the last two years, most notably with the public appearance of large language models. Generative artificial intelligence can be defined as the application of artificial intelligence and machine learning algorithms to generate new content (Goodfellow et al., 2016). The generative artificial intelligence market is expected to reach more than 667 billion USD value, which is a huge increase over 2022 market value of 29 billion USD (Generative AI Market Size, Share and Industry Trends [2030], n.d.).

Besides the ability to generate new content such as bodies of text (de Rosa & Papa, 2021), images (Rombach et al., 2022), or music (Wang et al., 2023), generative artificial intelligence can be used to transform existing content into more manageable forms. This can be useful if we need to digest large amounts of text or get a quick summary of the information at hand. These capabilities in generative artificial intelligence are commonly provided by large language models (LLM) (Radford & Narasimhan, 2018).

Due to the way how large language models work, they are suited to be a good candidate for the process of text extraction. ChatGPT (n.d.) is one of the more popular and most commonly used large language models due to its availability to a wide range of users. Our goal is to see if ChatGPT is a suitable option for the process of extracting data from text.

2. CONCEPT OF LARGE LANGUAGE MODELS

Large language models (LLM) represent a relatively new language model based on generative pre-trained transformers (Radford & Narasimhan, 2018). Generative pre-trained transformers represent the application of the transformer neural network (Vaswani et al., 2017) trained with unsupervised learning on large datasets. Training results yield probabilities between tokens

[1] Faculty of Organizational Sciences, Jove Ilića, 154, 11000, Belgrade, Serbia

inside the training dataset, usually words. The resulting model is specialized in supervised learning for specific tasks, such as language translation or text mining.

The first publicly available LLM is the GPT-1 model (Radford & Narasimhan, 2018), followed by the GPT-2 and GPT-3 models (Brown et al., 2020). The key difference between the models is in the amount of parameters used for training. The original GPT-1 model used 117 million parameters, compared to GPT-2 with 1.5 billion parameters and GPT-3 with 175 billion parameters.

An alternative to the GPT language model family is the BERT (Bidirectional Encoder Representations from Transformers) language model from Google (Devlin et al., 2019). The main difference when compared to the GPT family of large language models is in the way how next tokens are predicted in the sentence. BERT utilizes bi-directional context and word masking in order to perform predictions, while GPT is autoregressive and has a unidirectional path for determining context (left-to-right compared to BERT which performs context determination left-to-right and right-to-left).

The use of non-public datasets for training is a common characteristic of GPT and BERT language model families. The lack of public datasets or dataset descriptions keeps models hidden from general research contributions and potential improvements from third parties. The creation of the LLaMA (Touvron et al., 2023) family of models with the collaboration from Meta and Microsoft is a direct response to the problem of dataset availability.

Besides utilizing publicly available datasets, LLaMA utilizes several novel improvements to the transformer architecture, such as the application of the SwiGLU activation function (Shazeer, 2020) and rotary embeddings (Su et al., 2023). LLaMA, with its current iteration LLaMA 2 (Touvron et al., n.d.) provides significantly better performance than GPT and BERT language models.

Even though LLaMA and BERT provide better performance than the GPT language model family due to the improvements to the transformer architecture and larger amount of parameters used for training, the key blocker in public adoption is the accessibility of these models. One of the original goals of OpenAI, the company behind the GPT family of language models, was the benefit of AI for mankind (About, n.d.). ChatGPT (n.d.) is one of their key publicly available solutions that became available in November 2022. It is a front-end to the GPT-3 language model which would allow regular users to harness its capabilities for their day-to-day tasks and language model evaluation.

3. CURRENT WORK

Text mining is defined as a process with the goal of extracting meaningful information from text (L. Sumathy & Chidambaram, 2013). Text mining uses many different techniques, most notably natural language processing (NLP) and statistics to perform its goals. Our focus will be on the current state of large language models, both GPT and BERT, applied to the area of text mining.

One of the most common applications of text mining can be seen in the word prediction features of modern text editors. Utilizing text mining, the text editor can predict what would be the next word in the sentence. This is commonly used in programming environments (Allamanis & Sutton, 2013). This has evolved with the advent of large language models, now allowing tools to generate complete computer programs (Nijkamp et al., 2022) from user descriptions.

Another application for text mining is the improvement of text comprehension. Text comprehension is presented as the ability of the language model to answer questions, summarize text and hand and provide definitions of unknown concepts. One of these models is Galactica (Taylor et al., 2022), a large language model trained and focused on retrieving scientific data. This model also has the ability to reason given prompts and generate answers, allowing the end user to gain better insight into the information at hand. Another example regarding text mining with scientific data is the SciBERT (Beltagy et al., 2019) large language model which demonstrates improvements concerning the original BERT model.

Similarly, BioGPT (Luo et al., 2022) is a large language model belonging to the GPT family focused on textual mining and retrieval of biomedical data. The focus of BioGPT is end-to-end relation extraction, question answering, document classification and text generation. It is also worth mentioning that it outperforms the GPT-2_{medium} general purpose, large language model. There are similar implementations based on the BERT family of large language models, such as BioBERT (Lee et al., 2020).

4. CHATGPT AND ITS CAPABILITIES

As mentioned before, the concept behind the GPT family of large language models was introduced by Radford and Narasimhan (2018) and its improvements were introduced by Brown et al. (2020). The popularity of the GPT family of large language models has increased with the introduction of ChatGPT (n.d.) by OpenAI (n.d.), an originally GPT-3 large language model specialized for textual conversations. At the time of writing, there have been about 1.7 billion unique visits to the ChatGPT website (Similarweb, n.d.) which offers interaction with the model.

The popularity of ChatGPT comes from an intuitive user interface available via the web and natively on Android and iOS. The additional feature is the tuned GPT-3.5 and GPT-4 models made to represent human interaction, even though the training and model data are proprietary.

Giving additional instructions to the large language model is performed using prompts (Liu et al., 2023). Utilizing prompts we are able to customize the behavior of the large language model by setting baseline points and formatting for the returned answers. The process of utilizing prompts for large language model customization is called prompt engineering. This is an alternative for retraining the model whenever we wish to have a different context, allowing us to have a large model that we can engineer later on for our use cases.

Prompt engineering is one of the key features of using ChatGPT. Applying prompt engineering with ChatGPT allows the end user to get better results and in the required context, without retraining the model. With the application of prompt patterns (White et al., 2023) we are able to further increase the usability of ChatGPT.

Another key feature is the ability to use specialized plugins (Chat Plugins, n.d.), such as specialized search engines (Wolfram Plugin for ChatGPT, n.d.) and web browsing (Browsing, n.d.). This extends the original abilities of ChatGPT with new features that aren't a part of the original trained model, since they depend on outside services.

All of the features above are able to be combined, allowing the creation of complex questions and tasks for ChatGPT, utilizing multiple data sources and contexts.

5. TEXT MINING WITH CHATGPT

We have performed experiments on the following three text mining use cases with ChatGPT: sentiment analysis, document summarization and concept extraction in the domains of news articles, scientific papers and technical documentation.

Each experiment started with the following prompt: "Hello ChatGPT, I will need you to perform **text_mining_use_case** on text I will give you in the next prompt" where **text_mining_use_case** are sentiment analysis, document summarization or concept extraction. Prompt engineering was applied later in order to refine the returned results and provide better quality for the end user.

The following data was used for testing – a news article from Forbes (Tucker, n.d.), a scientific paper regarding the discovery of biological nerve conductivity (Hodgkin & Huxley, 1952) and a technical document outlining the extraction methods in Oracle database based data warehouses (Database Data Warehousing Guide, n.d.). Since the scientific text in question is large, we are focusing on the part "Refractory period".

Sentiment analysis can be defined as a method of extracting emotional sentiment from a body of text (L. Sumathy & Chidambaram, 2013). Sentiment analysis is used to convey the emotional meaning from bodies of text in order to gauge perceptiveness. Before testing, we read and analyzed the texts manually in order to determine the sentiments. The news article had positive sentiments, while the scientific text and technical document had neutral sentiments.

In the case of the news article, the sentiment is mostly positive and it has been presented descriptively. Prompt engineering with the following statement "Format the sentiment analysis by ranking" gives us a numerated list of sentiments identified and references inside the text. Performing advanced sentiment analysis with the prompt "As an analyst, perform advanced sentiment analysis" gives us a detailed breakdown of the discovered sentiments that are positive with further explanation of the referenced text.

Scientific text sentiment was discovered to be neutral, with an explanation about the process that determined its neutrality, compared to a newspaper article where sentiment has been referenced in the text. Similarly, the results were the same for the technical documentation, but with detailed sentiment ranking due to formatted subtitles in text. Asking additional questions regarding advanced sentiment analysis confirms the objectivity of these documents since sentiment is neutral.

Table 1. ChatGPT discovered sentiment comparison

Type of document	Expected main sentiment	Discovered main sentiment
News article	Positive	Positive
Scientific text	Neutral	Neutral
Technical document	Neutral	Neutral

Source: Own research

Document summarization is another branch of text mining that concerns the processing of text into more manageable parts that convey the same meaning (L. Sumathy & Chidambaram, 2013). This allows the reader to quickly go through large bodies of text and gain better knowledge of the text importance at hand. The key metric used for summarization was the final word count and the subjective measure of the authors about the validity of the summarization.

ChatGPT was able to perform document summarization of the news article by 48.79%, without further prompt engineering. After utilizing the prompt "Summarize the text even more" we were able to summarize it even more, bringing us to a total of 67.63% of summarization.

Due to the nature of scientific texts and technical documentation, document summarization gave us better results than news articles. ChatGPT managed to summarize the scientific text in question by 59.92% and with prompt engineering by 69.08%. The technical document was summarized by 58% and with prompt engineering by 80.21%.

Table 2. ChatGPT document summarization results

Type of document	First document summarization	Second document summarization	Author's subjective validity of summarization
News article	48.79%	67.63%	Valid
Scientific text	59.92%	69.08%	Valid
Technical document	58%	80.21%	Valid

Source: Own research

The process of concept extraction focuses on identifying and isolating key concepts from bodies of text in order to increase comprehension and organization of knowledge (L. Sumathy & Chidambaram, 2013).

Because of the way how ChatGPT is organized and the architecture of the GPT model, it is trivial for ChatGPT to perform concept extraction. Using the starting prompt "Hello ChatGPT, I will need you to perform concept extraction on text I will give you in the next prompt", we are able to summarize all three text's key concepts and ideas. Prompt engineering can help us with additional concept extraction (such as clarification of extracted concepts or their connections) in the next steps after the original prompt.

6. FURTHER RESEARCH

The provided use cases are a balance between the usability and practicality of ChatGPT in day-to-day work. With the complexity of large language models and text mining, there is room for further research in these areas.

One of the main directions would be deeper research into the applicability of ChatGPT as a solution to text mining. The performance of ChatGPT in text mining different kinds of text could give promising results, as well as the ability to set the context with prompt engineering. This can be approached with the applicability of prompt patterns (White et al., 2023) and unlearning (Yao et al., 2023) as a method to improve output.

Another important approach would be the capability of ChatGPT to perform advanced logical reasoning on mined bodies of text. This approach stems from the current research that assumes ChatGPT is a Chinese room and unable to perform advanced logical reasoning (Ling, 2023), but it is worth discovering the threshold of logic that can be applied to the text at hand.

Combining prompt engineering with the process of advanced logical reasoning, we should be able to build data mining pipelines (Wei et al., 2022) that can translate to specific data mining patterns. These emerging patterns can be used for text mining applications on different language models, such as LLaMA, focusing on a multi-model approach to text mining.

7. CONCLUSION

We have established that ChatGPT as a solution is a viable tool for text mining of three different kinds of text, most notably newspaper articles, scientific papers and technical documents. Testing was done for sentiment analysis, document summarization and concept extraction. All three use cases gave us positive results for the usage of ChatGPT. Starting prompt was "Hello ChatGPT, I will need you to perform **text_mining_use_case** on text I will give you in the next prompt" where **text_mining_use_case** are sentiment analysis, document summarization or concept extraction".

ChatGPT is able to perform sentiment analysis with results matching the expected values. With prompt engineering, detailed sentiment descriptions are returned by ChatGPT. Regarding document summarization, word count is lowered on average by 55.57% and by applying prompt engineering these results improve to 72.03%. No additional prompt engineering is required for concept extraction using ChatGPT.

References

About. (n.d.). Retrieved December 19, 2023, from https://openai.com/about

Allamanis, M., & Sutton, C. (2013). Mining source code repositories at massive scale using language modeling. 2013 10th Working Conference on Mining Software Repositories (MSR), 207–216. https://doi.org/10.1109/MSR.2013.6624029

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3613–3618. https://doi.org/10.18653/v1/D19-1371

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners (arXiv:2005.14165)*. arXiv. http://arxiv.org/abs/2005.14165

Browsing. (n.d.). Retrieved December 15, 2023, from https://openai.com/blog/chatgpt-plugins#browsing

ChatGPT. (n.d.). Retrieved July 18, 2023, from https://chat.openai.com

Chat Plugins. (n.d.). Retrieved December 15, 2023, from https://platform.openai.com/docs/plugins/introduction/chat-plugins-beta

Database Data Warehousing Guide. (n.d.). Oracle Help Center. Retrieved December 16, 2023, from https://docs.oracle.com/en/database/oracle/oracle-database/21/dwhsg/extraction-data-warehouses.html#GUID-A9A3D5CD-A34A-46BB-844A-76DFE119CE02

de Rosa, G. H., & Papa, J. P. (2021). A survey on text generation using generative adversarial networks. *Pattern Recognition, 119*, 108098. https://doi.org/10.1016/j.patcog.2021.108098

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Generative AI Market Size, Share and Industry Trends [2030]. (n.d.). Retrieved December 18, 2023, from https://www.fortunebusinessinsights.com/generative-ai-market-107837

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology, 117*(4), 500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Ling, M. H. (2023). ChatGPT (Feb 13 Version) is a Chinese Room. https://doi.org/10.48550/ARXIV.2304.12411

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys, 55*(9), 195:1-195:35. https://doi.org/10.1145/3560815

L. Sumathy, K., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues An Overview. *International Journal of Computer Applications, 80*(4), 29–32. https://doi.org/10.5120/13851-1685

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6), bbac409. https://doi.org/10.1093/bib/bbac409

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., & Xiong, C. (2022, March 25). CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. International Conference on Learning Representations. https://www.semanticscholar.org/paper/CodeGen%3A-An-Open-Large-Language-Model-for-Code-with-Nijkamp-Pang/38115e80d805fb0fb8f090dc88ced4b24be07878

OpenAI. (n.d.). Retrieved December 14, 2023, from https://openai.com/

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models (arXiv:2112.10752). arXiv. http://arxiv.org/abs/2112.10752

Shazeer, N. (2020). GLU Variants Improve Transformer (arXiv:2002.05202; Version 1). arXiv. https://doi.org/10.48550/arXiv.2002.05202

Similarweb. (n.d.). Chat.openai.com traffic analytics, ranking stats & tech stack. Retrieved December 14, 2023, from https://www.similarweb.com/website/chat.openai.com/

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). RoFormer: Enhanced Transformer with Rotary Position Embedding (arXiv:2104.09864; Version 5). arXiv. https://doi.org/10.48550/arXiv.2104.09864

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). Galactica: A Large Language Model for Science. https://doi.org/10.48550/ARXIV.2211.09085

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Touvron, H., Martin, L., & Stone, K. (n.d.). Llama 2: Open Foundation and Fine-Tuned Chat Models.

Tucker, H. (n.d.). Sprechen Sie Growth? How Duolingo Became A Hot Stock In 2023, Plus 99 More Mid-Cap Winners. Forbes. Retrieved December 16, 2023, from https://www.forbes.com/sites/hanktucker/2023/12/15/sprechen-sie-growth-how-duolingo-became-a-hot-stock-in-2023-plus-99-more-mid-cap-winners/

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is All you Need. Neural Information Processing

Systems. https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776

Wang, L., Zhao, Z., Liu, H., Pang, J., Qin, Y., & Wu, Q. (2023). A Review of Intelligent Music Generation Systems (arXiv:2211.09124). arXiv. http://arxiv.org/abs/2211.09124

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv. https://www.semanticscholar.org/paper/Chain-of-Thought-Prompting-Elicits-Reasoning-in-Wei-Wang/1b6e810ce0afd0dd093f789d2b2742d047e316d5

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT (arXiv:2302.11382). arXiv. http://arxiv.org/abs/2302.11382

Wolfram Plugin for ChatGPT. (n.d.). Retrieved December 15, 2023, from https://www.wolfram.com/wolfram-plugin-chatgpt/

Yao, Y., Xu, X., & Liu, Y. (2023). Large Language Model Unlearning. https://doi.org/10.48550/ARXIV.2310.10683