Future of Text Mining and Unstructured Data Analysis

Nino Miljkovic

June 22, 2025

Introduction

Text mining has emerged as a critical methodology for extracting meaningful and actionable insights from unstructured data. Its use is becoming increasingly valuable as, according to recent reports, 80% of all data available is unstructured [1], and companies, institutions and businesses look to leverage this valuable asset in order to make the most beneficial data-driven decisions.
In that light, this essay aims to provide a comprehensive examination of text mining practices by exploring traditional methodologies alongside their latest implementation by means of artificial intelligence approaches that have opened up new possibilities.

Initially, I will introduce fundamental concepts and techniques used in text mining to establish foundational principles that govern effective and efficient text analysis. Subsequently, the discussion continues towards examining state-of-the-art AI models, particularly focusing on BERT (Bidirectional Encoder Representation from Transformers) and GPT (Generative Pretrained Transformer), which have revolutionized the field of artificial intelligence and have shown remarkable capabilities when it comes to analyzing unstructured data, drawing sentiments and meaning from it, processing natural language, and specifically in the case of GPT, generating natural language responses.

Furthermore, in order to illustrate the practical application of leveraging artificial intelligence for text mining, I decided to introduce a detailed example of how Perplexity's Open Research module can be utilized for scientific and scholarly research to improve workflows and increase efficiency. This example seeks to provide a tangible benefit of AI-assisted text mining by demonstrating how researchers can use this tool to navigate the vast troves of scientific and academic papers and resources.

Finally, the essay addresses various ethical and moral issues that might arise from using text mining in a traditional sense or through an AI platform. Some of the issues inherent to the nature of text mining involve privacy concerns, data bias, data ownership, and users' responsibilities and guidelines. This discussion aims to emphasize the importance of developing frameworks that balance the use and power of technology with ethical boundaries.

In summary, the main goal of this essay is to provide theoretical understanding along with the practical use of text mining, while remaining cognizant of the ethical concerns and best practices of this ever-evolving field while discussing its current state and future possibilities.

Text Mining and Unstructured Data

As mentioned in the introduction, we live in the world of Big Data, and 80% of that data is unstructured. This includes articles, research papers, survey results, customer reviews, emails, social media posts, as well as non-textual data (audio and video) of various forms and lengths. Whereas structured and organized data, usually contained in tabular datasets, can be analyzed using traditional statistical analysis approaches, unstructured text data requires special processes that are referred to as text mining.

The advancements in computational power and data storage, and the emergence of machine learning, has allowed researchers to utilize tools like Python's text mining libraries, deep learning modules, and natural language processing to uncover hidden patterns and insights. This process would have been impossible to perform at such a large scale in the past as the amount of data we analyze today would have been an insurmountable task before we had access to these modern tools and programs.
The goal of text mining is to convert unstructured text into a structured format from which valuable patterns, trends, concepts, or insights can be extracted, which can ultimately shape decisions across domains like business, research and public policy.[2]

Given the fact that we live in a data-driven world, having the ability to cleanly extract key insights and concepts from such an immense amount of information has become a priority not just for businesses, but also government agencies, medical centers, research laboratories and many other types of organizations. How big the text mining market has become is also indicated in its annual growth where it has gone from $7.05 billion in 2024 to $8.51 billion in 2025, marking a compound annual growth of 20.7%. These numbers are in reference to the on-premise and cloud-based text mining software that is used for everything from text analysis to fraud and spam detection, to various legal, medical, government and banking purposes. [3]

As eloquently detailed in our DSCI 6700 lectures, the actual process of text mining is composed of several parts: (A.Yu, lecture, 2025)

1. **Text processing**
   In this stage, raw text is cleaned and converted into a format that is suitable for analysis.

   a. Tokenization: splitting text into tokens, often individual words
   b. Lowercasing: to achieve uniformity
   c. Stopword Removal: eliminating common words like 'and', 'is', 'the'
   d. Stemming: reducing words to their base form by removing stems (*running* → *run*)

e. Lemmatization: reducing words to their dictionary form within context (*better* → *good*)

2. **Text Representation**
   These techniques transform the text into numerical formats for algorithms.

   a. Bag of Words (BoW): represents text by word frequency, ignores word order
   b. TF-IDF (Term Frequency-Inverse Document Frequency): quantifies a term's importance in a specific document by comparing its frequency in that document to its frequency in the entire corpus.

3. **Feature Engineering and Dimensionality Reduction**
   After vectorizing the text, features can be reduced.

   a. N-grams: captures sequences of N words. These can be bigrams, trigrams, and so on.
   b. LSA (Latent Semantic Analysis): reduces dimensions while preserving semantic relationships between terms and documents using techniques like Singular Value Decomposition (SVD). (A.Yu, lecture, 2025) [4][5]

Along with these methods, researchers also use Named Entity Recognition (NER) that identifies and categorizes entities in a text, such as names of people, organizations, locations, dates and similar, which becomes useful when a specific type of information needs to be retrieved. In addition, Sentiment Analysis can offer valuable information with regards to the sentiments expressed in a text, which provides another angle of viewing the data. The sentiment is usually classified as positive, negative or neutral. However, Sentiment Analysis can be lexicon-based or rely on machine learning models trained on labeled sentiment datasets, often requiring specialized libraries like NLTK or SpaCy.[6]

It is evident that most of these processes rely on specialized programming libraries and machine learning methods, and as those technologies evolve, the text mining processes are bound to evolve alongside them. Artificial intelligence, in particular, has had a significant impact on the field of text mining, by accelerating the text mining process from the initial collection of data through techniques like web scraping, to the automation of preprocessing and feature extraction, and finally to the application of deep learning models capable of uncovering deeper, more nuanced insights.

Therefore, a comprehensive understanding of the current state of text mining, and the direction in which it is heading, requires examining the role of cutting-edge AI models like

GPT and BERT, whose advanced language understanding capabilities have reshaped the way unstructured text is analyzed and interpreted.

Artificial Intelligence and its Impact on Text Mining

The introduction of deep learning models, particularly those based on the transformer architecture, has dramatically transformed the field of text mining. This shift began with the seminal paper "Attention is All You Need" by Vaswani et al. (2017), which introduced the transformer model and revolutionized how machines process language. At the heart of this innovation is the self-attention mechanism, which enables models to weigh the importance of different words in a sentence, regardless of their position, thus capturing long-range dependencies more effectively than previous architectures.

> *"...In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output." [7]*

Building upon this foundation, two of the most influential models emerged: BERT (Bidirectional Encoder Representation from Transformers), which was introduced by Google in 2018, and GPT (Generative Pretrained Transformer), developed by OpenAI.

The key feature of BERT is that it is bidirectional, which allows it to read sequences from left to right and right to left, allowing it to capture the context in a much improved way over the prior models. This has become especially relevant for the field of text mining and has significantly improved its tasks and processes like Named Entity Recognition (NER), semantic search, and question answering, as it increased the accuracy and depth of understanding when models process unstructured text.[9]

On the other hand, OpenAI's GPT models (particularly GPT-3 and GPT-4) are unidirectional autoregressive models designed primarily for language generation. While BERT excels at understanding and classifying text, GPT models have proven to be exceptional in generating contextually rich responses, and creative and coherent language.[8] When it comes to text mining, these capabilities of GPT models make them ideal for text summarization, content generation, sentiment analysis, and even zero-shot classification, where models perform tasks without task-specific training.

Together, these encoder(BERT) and decoder(GPT) architectures have redefined what is possible in text mining. Tasks that were once labor-intensive or simply not feasible, can now be handled efficiently and comprehensively by pre-trained models, often fine-tuned for specific domains and applications. Their ability to handle context, nuance, and

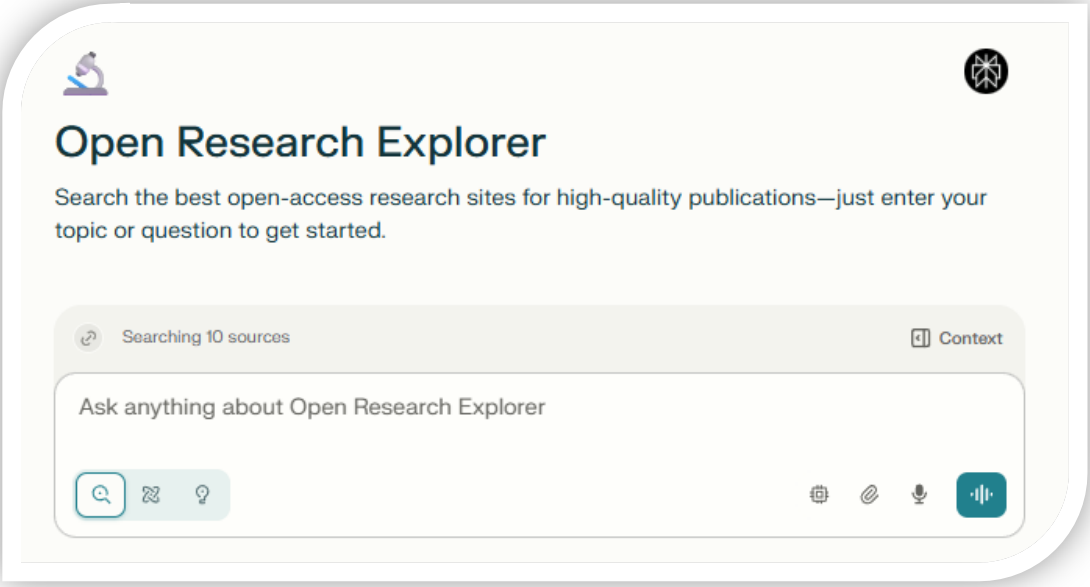
complexity, is slowly making their application an indispensable part in a variety of fields, ranging from legal document review and medical research to customer sentiment, market analysis, scientific research and similar tasks.

Case Study – Perplexity's Open Research Explorer

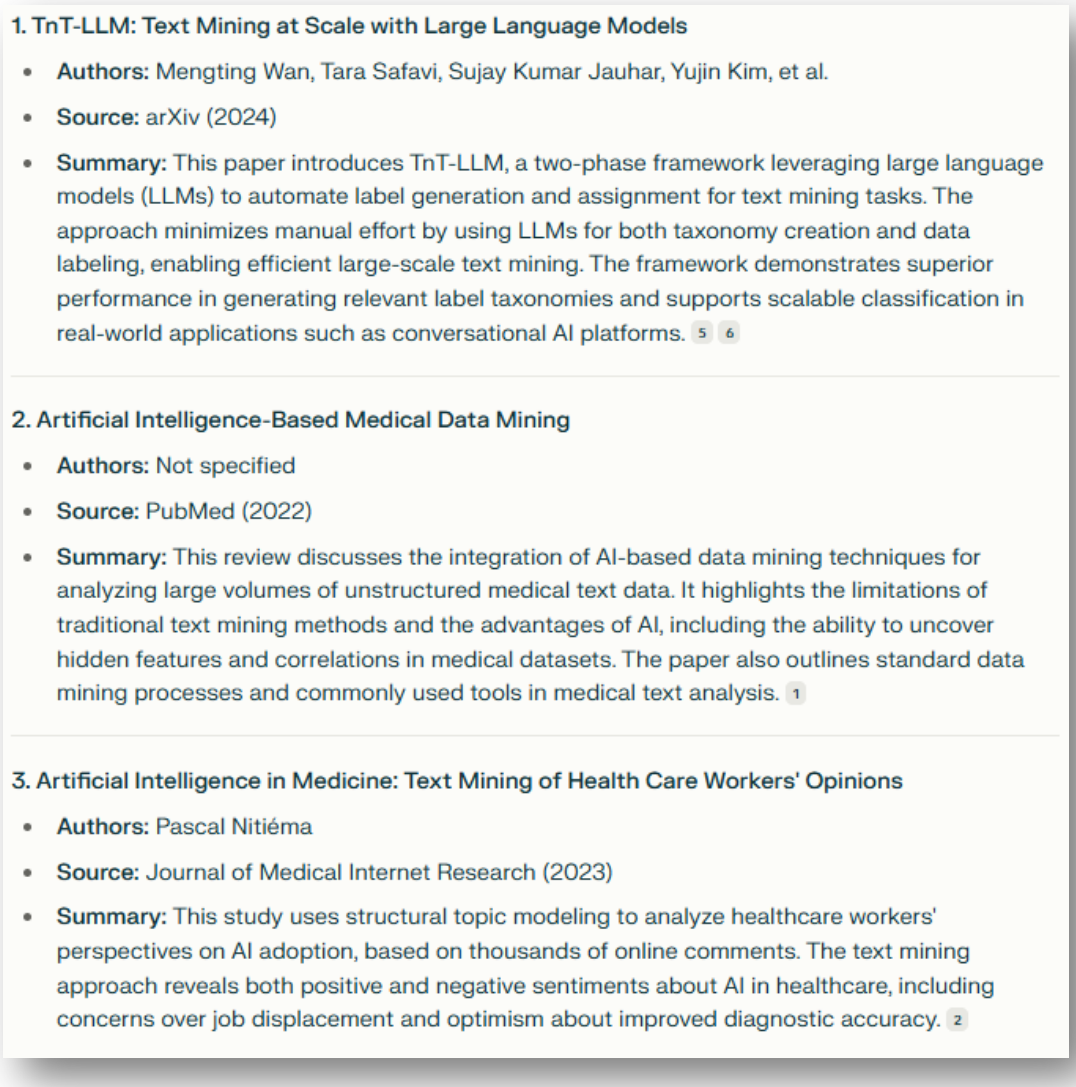
While BERT and GPT are integrated into a wide array of tools, systems and services, new platforms built upon these and other transformer models continue to emerge on the market. One such platform that has gained prominence and recognition, especially for its research functionality, is Perplexity.ai.

Perplexity's Deep Research system employs techniques such as advanced learning modeling, document retrieval and ranking, hierarchical summarization, probabilistic reasoning, iterative refinement, and even citation extraction. This advanced tool integrates autonomous reasoning with high-speed processing to generate detailed reports on specialized subjects. Benchmarks have shown that it often outperforms other systems in different research tasks. [10] [11]

The *Spaces* section at perplexity.ai allows users to organize threads and resources, collaborate on specific topics, and utilize pre-built templates. One such template, *Open Research Explorer,* is an incredibly powerful tool for academic and scientific research, providing fast and easy access to open-access publications on virtually any topic.



In order to explore its capabilites, I searched for “Text Mining and AI” using its interface and was provided with a curated list of open-access academic publications that explore the intersection of text mining and artificial intelligence. The output was a list of academic articles, each showing the author(s), source of the publication, and a brief summary, making it easy to assess relevance at a glance:



1. TnT-LLM: Text Mining at Scale with Large Language Models

- **Authors:** Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, et al.
- **Source:** arXiv (2024)
- **Summary:** This paper introduces TnT-LLM, a two-phase framework leveraging large language models (LLMs) to automate label generation and assignment for text mining tasks. The approach minimizes manual effort by using LLMs for both taxonomy creation and data labeling, enabling efficient large-scale text mining. The framework demonstrates superior performance in generating relevant label taxonomies and supports scalable classification in real-world applications such as conversational AI platforms. 5 6

2. Artificial Intelligence-Based Medical Data Mining

- **Authors:** Not specified
- **Source:** PubMed (2022)
- **Summary:** This review discusses the integration of AI-based data mining techniques for analyzing large volumes of unstructured medical text data. It highlights the limitations of traditional text mining methods and the advantages of AI, including the ability to uncover hidden features and correlations in medical datasets. The paper also outlines standard data mining processes and commonly used tools in medical text analysis. 1

3. Artificial Intelligence in Medicine: Text Mining of Health Care Workers' Opinions

- **Authors:** Pascal Nitiéma
- **Source:** Journal of Medical Internet Research (2023)
- **Summary:** This study uses structural topic modeling to analyze healthcare workers' perspectives on AI adoption, based on thousands of online comments. The text mining approach reveals both positive and negative sentiments about AI in healthcare, including concerns over job displacement and optimism about improved diagnostic accuracy. 2

Furthermore, another feature that sets this tool apart is its transparency and explainability. Perplexity’s Open Research Explorer documents all the steps taken to reach results in its *Steps* section, thus providing insight into the model’s reasoning and methodology. This view into its inner-working shows that this approach goes far beyond simple keyword matching and static search by leveraging iterative semantic search, transformer-based reasoning, and other protocols to deliver high-quality research material.

In addition, by streamlining the research process in this way, and enhancing accessibility to quality scholarly material, platforms like Perplexity.ai significantly accelerate the process of accessing and interacting with existing academic knowledge, allowing researchers to find relevant insights in a fraction of the time traditional methods would require.

Ethical Concerns and Future of Text Mining

As powerful and prevalent text mining has become, especially with the integration of transformer-based AI models, it is important to always consider the ethical implications that accompany its use. While it is true that these tools offer efficiency, insight, and access to important data, we must always be cognizant of the questions involving privacy, data ownership, bias, and transparency.

As text mining is often used in analyzing personal communication, reviews, social media posts or even government-connected data, data privacy concerns immediately arise. Without clear consent and transparency, mining such data can violate principles of privacy, data ownership, and raise other ethical concerns. Important care should especially be taken when web scraping information from a number of sources, in order to ensure that the collection of that data is permitted and that none of the aforementioned principles are harmed.

Additionally, performing AI-assisted text mining can blur the lines between fair use and exploitation of intellectual property if the sources of the data the models are trained on or that they provide are not fully transparent. This can also lead to another ethical concern, which is the bias in the output when performing tasks like sentiment analysis or content classification. If the input data reflects historical or cultural biases or inaccuracies, this will be apparent in the model's output as well. This is why responsible development and continuous auditing of text mining systems is crucial in order to mitigate these concerns and provide improved outcomes in the future.

Therefore, looking ahead, the future of text mining will not only be shaped by the advancements in artificial intelligence models, but also by our ability to embed ethical frameworks and guardrails into both the design and the application of their architecture. Emerging regulations, like EU's AI Act [12], have already started to address some of these questions by proposing data protections laws and AI use and development regulations.

That being said, the potential of text mining continues to expand. Some of the most exciting new features are multimodal and multilingual analyses that integrate not just text, but also video and audio, in a variety of languages, expanding the existing availability of data sources and with the added potential to enhance multicultural understanding. Also, as

transformer models become smaller and faster, text mining will become even more democratized by reaching smaller organizations, educators and individual researchers.[13]

Conclusion

In conclusion, text mining has become a vital component of modern data analysis through transforming unstructured text into actionable insights. From traditional methods to advanced AI models like BERT and GPT, the field has rapidly expanded across different fields, such as healthcare, law, marketing, research etc. Various tools, like Perplexity's Open Research Explorer, have made information more accessible and streamlined the discovery of knowledge and increased the scope of research in a significant way. Multimodal and multilingual models are revealing new possibilities and options that are yet to be fully explored.
However, we must always be aware of the ethical challenges these rapid advances possess and as we move forward the development of text mining must seek to balance innovation and efficiency with transparency, fairness and responsible data use.

References

1. The Business Research Company. (2024, March 6). *Unstructured data insights: Key statistics revealed*. The Business Research Company. Unstructured Data Insights: Key Statistics Revealed
2. IBM. (n.d.). *What is text mining?* IBM. What Is Text Mining? | IBM
3. The Business Research Company. (2025, January). *Text mining global market report*. The Business Research Company. https://www.thebusinessresearchcompany.com/report/text-mining-global-market-report
4. China, C. R. (2023, August 28). *Text mining use cases*. IBM. https://www.ibm.com/think/topics/text-mining-use-cases
5. Utkarsh. (2023, May 11). *Text mining tutorial*. Scaler. https://www.scaler.com/topics/data-mining-tutorial/Text-Mining/
6. Rafalsky, R. (2024, November 26). *6 Must-Know Python Sentiment Analysis Libraries*. https://www.netguru.com/blog/python-sentiment-analysis-libraries
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). *Attention is all you need*. https://arxiv.org/pdf/1706.03762
8. TowardsNLP.com. (2023, November 17). *BERT vs. GPT-3: Comparing Two Powerhouse Language Models*. https://www.towardsnlp.com/bert-vs-gpt-3-comparing-two-powerhouse-language-models/
9. Rogers, A., Kovaleva, O., Rumshisky, A. (2020, November 9) *A Primer in BERTology: What We Know About How BERT Works*. https://arxiv.org/pdf/2002.12327
10. Perplexity Team. (2025, February 14). https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research
11. Vaughan-Nichols, S. (2025, February 19). *What is Perplexity Deep Research, and how do you use it?* https://www.zdnet.com/article/what-is-perplexity-deep-research-and-how-do-you-use-it/
12. European Parliament. (2023, August 6). *EU AI Act: first regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
13. Masood, A. (2025, May 21). *The State of Embedding Technologies for Large Language Models — Trends, Taxonomies, Benchmarks, and Future Directions*. https://medium.com/%40adnanmasood/the-state-of-embedding-technologies-for-large-language-models-trends-taxonomies-benchmarks-and-95e5ec303f67