https://doi.org/10.31891/2219-9365-2025-82-52
UDC 004.8

OLIANIN Denys
Ternopil Ivan Pului National Technical University
https://orcid.org/0009-0002-5266-3941
email: denys.olianin@gmail.com
TSYPRYK Halyna
Ternopil Ivan Pului National Technical University
https://orcid.org/0000-0002-8106-5628
email: tsupryk_h@tntu.edu.ua

OVERVIEW OF TRANSFORMERS ROLE IN DATA MINING FROM UNSTRUCTURED DATA

The rapid growth of Big Data has made it increasingly important to extract meaningful insights from unstructured sources such as text, audio, video, and emails. Traditional data mining techniques—like tokenization, clustering, classification, and association rule mining—have provided a basis for processing these complex data forms. However, they often struggle to capture the subtle semantic and contextual relationships that are inherent in unstructured data. In this article, we examine the limitations of these conventional methods and explore the impact of Transformer Neural Networks (TNNs) on unstructured data mining.

Transformer architectures have revolutionized the field by employing self-attention mechanisms and positional encodings, which allow for parallel processing of data. This new approach enables the creation of high-quality embeddings that capture both semantic and syntactic information. As a result, tasks such as sentiment analysis, topic modeling, and automated summarization are significantly enhanced. Additionally, integrating transformers into audio signal processing and email mining has led to notable improvements in automatic speech recognition and semantic analysis, effectively addressing some of the long-standing challenges in these areas. The findings discussed in this article highlight the potential of transformer-based approaches to not only overcome the limitations of traditional data mining methods but also to open the door to innovative applications across various fields. Future research directions include developing more computationally efficient transformer models and exploring hybrid approaches that combine traditional techniques with advanced neural architectures. These efforts will ultimately push the boundaries of what is possible in unstructured data mining.

Keywords: Big Data, Unstructured Data, Data Mining, Transformer Neural Networks (TNNs), Audio Signal Processing, Email Mining

ОЛЯНІН Денис, ЦУПРИК Галина
Тернопільський національний технічний університет імені Івана Пулюя

ОГЛЯД РОЛІ ТРАНСФОРМЕРНИХ НЕЙРОНИХ МЕРЕЖ У ВИДОБУВАНІ ІНФОРМАЦІЇ ІЗ НЕСТРУКТУРОВАНИХ ДАНИХ

Швидке зростання обсягів великих даних (Big Data) зумовило зростання важливості отримання значущих висновків з неструктурованих джерел, таких як текст, аудіо, відео та електронна пошта. Традиційні методи інтелектуального аналізу даних (майнінгу), такі як токенізація, кластеризація, класифікація та виявлення асоціативних правил, забезпечували базові можливості для обробки цих складних форм даних. Втім, ці методи часто не можуть повністю охопити тонкі семантичні та контекстуальні взаємозв'язки, притаманні неструктурованим даним. У цій статті ми аналізуємо обмеження традиційних методів і досліджуємо вплив нейронних мереж на основі трансформерів (Transformer Neural Networks, TNNs) на майнінг неструктурованих даних.

Архітектури трансформерів революціонізували сферу завдяки механізмам самоуваги (self-attention) та позиційного кодування (positional encoding), що дозволяють паралельно обробляти дані. Цей новий підхід забезпечує створення якісних вкладень (ембеддингів), які фіксують як семантичну, так і синтаксичну інформацію. Як наслідок, суттєво покращується якість виконання завдань, таких як аналіз настроїв (sentiment analysis), тематичне моделювання (topic modeling) і автоматичне реферування тексту (automated summarization). Крім того, інтеграція трансформерів в обробку аудіосигналів та аналіз електронної пошти призвела до помітних покращень в автоматичному розпізнаванні мовлення і семантичному аналізі, ефективно вирішуючи деякі давні проблеми в цих областях.

Результати, представлені в статті, демонструють потенціал підходів на основі трансформерів, які не тільки долають обмеження традиційних методів аналізу даних, але й відкривають шлях для інноваційних застосувань у різних сферах. Подальші напрями досліджень включають розробку більш ефективних з обчислювальної точки зору моделей трансформерів, а також вивчення гібридних підходів, які поєднують традиційні методики з передовими нейронними архітектурами. Ці зусилля врешті-решт дозволять суттєво розширити межі можливостей у сфері майнінгу неструктурованих даних.

Ключові слова: Великі Дані, неструктуровані дані, аналіз даних, трансформерні нейронні мережі, обробка аудіосигналів, аналіз електронної пошти.

Стаття надійшла до редакції / Received 19.04.2025
Прийнята до друку / Accepted 11.05.2025

INTRODUCTION

Big data is a term that existed in the past in 1990s and grew in popularity in the world thanks to different groups of people. [1-2] Now Big Data is very useful instrument for every corporation that wants to gather insights from their collected data as it has variety of technologies and technics for analyzing and visualization of data. [3]

We would like to get your attention to subset of Big Data - Data Mining and look at mining knowledge from unstructured data. Typically, data can be divided into three types:

- Structured data (RDMS, SQL)
- Semi-structured data (NOSQL, JSON, emails)
- Unstructured data (audio and video, unstructured text)

While gathering knowledge from structured data is straightforward, case with unstructured and semi-structured data is more complicated as there is no known beforehand structure behind the data.

However, transformer neural networks and LLM can help with this task. Many research show that TNN have demonstrated high effectiveness in various fields of applications. This neural architecture is effective in speech recognition tasks because it can effectively manage sequential data and capture long-range dependencies using self-attention mechanisms [4]. In computer vision, the same self-attention mechanism allows the model to integrate global contextual information and understand spatial relationships, thereby overcoming some limitations of traditional convolutional approaches [5]. Additionally, transformers have proven highly effective for extracting information from unstructured text. By generating rich, contextualized embeddings, they significantly enhance tasks such as classification and question answering[6]. Overall, these features highlight the transformer's versatility—its ability to process entire sequences in parallel and capture both local and global patterns contributes to robust performance across diverse domains, from speech and vision to complex text mining tasks.

UNSTRUCTURED DATA

First let us introduce you to unstructured data, unlike structured data, is challenging for ETL processing and knowledge comprehension. According to the International Data Corporation research unstructured data will account of 80% of whole data and will reach 179.6 ZB of global data generation. Another report made by NRoad[7] shows that only 58% of unstructured data is being analyzed after initial processing, which has potential in further analysis.

Unstructured data includes various documents, PDF files, audio, video, emails and social media posts and comments. Every document consists of unstructured text and images; however, it may imply some structure as headings and text that accords to it, same goes for PDF files. Videos on the other hand, are sequences of frames, each of which consists of 1 or 3 arrays of data, which are bytes that describe saturations in colors. Despite an announced structure, we must define images as unstructured data, due to the need of understanding what objects are shown and what context they hold to complete comprehending of the data.

Regardless of the complexity in describing and analyzing unstructured data it carries a lot of valuable knowledge. For example, various insights from medical reports, social reaction for some events in social network comments or social media posts. This has potential for both government and science in achieving progress in various fields.

TRADITIONAL DATA MINING TECHNICS FOR UNSTRUCTURED DATA

Unstructured data is a major challenge in data analysis because it does not follow a predefined format. Even though it is complex, many traditional data mining techniques have been adapted to work with it, forming the basis of current unstructured data analysis. While these older methods were useful at one time, they now act as a starting point for more advanced techniques like Transformer Neural Networks, which have transformed the field.

One of the most important steps when working with unstructured data is preprocessing. This process converts raw data into a format that can be analyzed. Common techniques include tokenization, stop word removal, and either stemming or lemmatization [7]. Tokenization breaks text into smaller parts, such as words or phrases, which become the building blocks for analysis [7]. Stop word removal eliminates common words like "and" or "the" that usually do not add meaningful information [8]. Stemming and lemmatization reduce words to their basic forms, grouping similar words together (for example, "running" and "ran") [9]. These steps are essential to improve data quality before applying further analytical methods.

Clustering is another widely used method in unstructured data mining. It groups similar data points together into clusters, helping to reveal hidden structures. Techniques such as k-means clustering and hierarchical clustering are commonly used [10, 11]. In k-means clustering, the data is divided into a specific number of clusters by reducing the variance within each cluster. Hierarchical clustering, on the other hand, builds a tree-like structure that shows how the data points relate to one another. Although these methods can uncover patterns, they often struggle with the high dimensionality and inherent complexity of unstructured data.

Classification techniques also play an important role in unstructured data mining by categorizing data into predefined groups. Decision trees provide a clear and understandable framework for classification [7]. Similarly, Naive Bayes classifiers use probabilistic models that assume feature independence to assign labels efficiently [9]. However, these methods sometimes fail to capture the deeper semantic relationships that are important for understanding complex datasets.

Association rule mining is another key method in traditional data mining. It finds patterns of items that occur together in datasets, which is particularly useful for analyzing unstructured text [7]. For example, this method can

identify frequently co-occurring terms in a text corpus, offering insights into common themes or trends. Despite its benefits, association rule mining often requires extensive parameter tuning to work effectively in different applications.

Finally, information extraction techniques aim to identify and pull out specific entities or relationships from unstructured data. One popular method is Named Entity Recognition (NER), which detects entities such as names, dates, or locations. Additionally, relationship extraction focuses on understanding how these entities are connected [8]. These techniques help convert unstructured data into a structured form, making it easier to analyze with traditional methods.

Audio Signal Processing in Data Mining

Audio data is a type of unstructured data that poses unique challenges for data mining because of its time-dependent nature, variations in frequency, and reliance on context. Traditional audio signal processing has focused on feature extraction techniques to convert raw audio signals into a more structured form. Methods such as Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) are commonly used to create compact representations that preserve important characteristics like pitch and tone. These extracted features are then fed into machine learning algorithms for tasks like speech recognition, speaker identification, and audio classification [12].

Earlier pipelines for automatic speech recognition (ASR) were based on hybrid systems that integrated several components, including acoustic models, language models, and pronunciation lexicons. Although these systems were effective, they required extensive engineering and were hampered by their inherent complexity and interdependent parts. Recent progress in neural architectures—especially end-to-end models—has greatly simplified the ASR pipeline and revolutionized audio signal processing [12].

e-mail Processing in Data Mining

Emails represent a common form of unstructured data, as they combine text, metadata, and attachments into a single, rich but complex source. Their semi-structured format, which includes metadata such as the subject, sender, and timestamps alongside free-form text, makes emails especially challenging for traditional data mining methods. Over time, researchers have developed a range of approaches to process email data and extract useful insights, focusing on tasks like spam detection, sentiment analysis, and topic modeling.

Spam filtering is one of the earliest and most widely used techniques in email mining. Its goal is to classify emails as either spam or legitimate messages. Traditional spam filters often employ machine learning algorithms such as Naive Bayes classifiers and support vector machines (SVMs). These methods work by extracting features from email content—including word frequencies, keywords from the subject line, and other metadata [7]. For example, Naive Bayes classifiers estimate the probability that an email is spam based on the presence of specific terms (e.g., "free" or "win"), and they have proven effective due to their simplicity and speed.

Another important task in email mining is sentiment analysis, which seeks to identify the emotional tone of email content. This technique is especially valuable in customer service, where understanding customer sentiment can guide better responses and improve service quality. Traditional sentiment analysis methods typically rely on lexicon-based approaches or simple machine learning models to categorize emails as positive, negative, or neutral [8]. However, these techniques often struggle to capture subtle nuances in language, such as sarcasm or context-dependent meanings.

Topic modeling is yet another crucial technique used to reveal the underlying themes or subjects discussed in emails. Traditional models like Latent Dirichlet Allocation (LDA) analyze how words co-occur within emails to identify clusters of related terms that represent different topics [9]. Although these models can successfully identify broad themes, they often require manual tuning and may not fully capture the semantic depth of the text.

TRANSFORMER NEURAL NETWORKS

Transformer Neural Networks first introduce in “Attention is all you need” [17] revolutionized the neural networks from traditional convoluted and sequential processing models to highly parallelizable, attention-driven frameworks that excel at capturing long-range dependencies and complex interactions within data.

Its main components allow the model to recognize and capture contextual relationships of the input data can work in parallel, providing scalability to ensure complex findings in the data.

Since transformers process input data in parallel rather than sequentially, positional encodings are introduced to retain information about the order of elements. This is critical for tasks where the sequence carries semantic meaning, such as language processing.

This capability is particularly beneficial when dealing with unstructured text, where context, idiomatic expressions, and long-range dependencies play crucial roles in understanding meaning.

Transformer models are especially strong because they can create high-quality embeddings—dense numerical representations of text. These embeddings capture both the meaning and structure of language, allowing data mining algorithms to recognize subtle patterns, detect anomalies, and group similar documents together. By

turning raw text into these rich representations, transformers help with tasks like topic modeling, sentiment analysis, and automated summarization, leading to more accurate and insightful data extraction.

In addition, transformers are well-suited for handling large-scale datasets thanks to their parallelizable architecture. This means that even very large collections of unstructured text can be processed quickly, making them ideal for real-time applications such as social media monitoring or customer feedback analysis. The process of pre-training and fine-tuning further boosts their performance. Models like BERT, GPT, and their variants are first trained on diverse text sources and then fine-tuned for specific tasks, reducing the need for extensive labeled data while maintaining high accuracy.

A significant breakthrough in audio signal processing is the incorporation of Transformer Neural Networks into automatic speech recognition (ASR) systems. Transformers use self-attention mechanisms to process entire audio sequences simultaneously, capturing the overall context and improving transcription accuracy. This approach is especially beneficial for long-form speech data, where traditional models may struggle to keep track of context. Recent research indicates that Transformer-based ASR systems perform better than older architectures, particularly in noisy and multilingual environments [10].

Although end-to-end models and Transformer-based approaches have greatly advanced the field, they do present challenges. These models typically require large amounts of labeled training data and considerable computational resources during training. Despite these hurdles, the move toward end-to-end and Transformer-based models marks a significant shift in how audio data is processed and analyzed.

Transformer Neural Networks have also transformed email data mining. They excel at capturing the sequential and contextual aspects of emails, making them ideal for understanding email threads, extracting semantic meaning, and performing complex tasks like entity recognition and relationship extraction. Pre-trained models such as BERT and GPT have shown impressive performance when analyzing unstructured text in emails. These models can process not only the email content but also their metadata and attachments, providing a comprehensive understanding of email datasets.

Overall, the integration of Transformers into email mining represents a major step forward, overcoming many limitations of traditional methods and enabling new applications. In the next section, we will explore how Transformer Neural Networks are redefining data mining for unstructured data, including their use in processing emails, audio, and other complex data sources.

CONCLUSIONS

As unstructured data—from text and audio to video and emails—continues to grow, the shortcomings of traditional data mining methods have become more evident. While these conventional techniques laid the groundwork for data analysis, they often fail to capture the deep semantic and contextual details that characterize unstructured information.

The introduction of Transformer Neural Networks, as described in "Attention is All You Need," marks a significant shift in this field. By using self-attention mechanisms and positional encodings, transformers can process data in parallel, manage long-range dependencies, and produce robust, high-quality embeddings. These features enable a more refined understanding of unstructured text, supporting advanced tasks such as topic modeling, sentiment analysis, and automated summarization. Moreover, incorporating transformers into audio processing and email mining has led to major improvements in automatic speech recognition and the semantic analysis of complex communication threads.

Looking ahead, Transformer Neural Networks are set to play a key role in unlocking the full potential of unstructured data. The progress achieved so far not only enhances our ability to extract meaningful insights but also opens the door to innovative applications across diverse fields—from business intelligence to scientific research.

References

1. Snijders, Chris, Uwe Matzat, Ulf-Dietrich Reips, 2012. "Big Data" : Big Gaps of Knowledge in the Field of Internet Science. In: International Journal of Internet Science. 2012, 7(1), pp. 1-5. eISSN 1662-5544
2. https://www.lightsondata.com/the-history-of-big-data/#:~:text=Some%20argue%20that%20it%20has,the%20O'Reilly%20Media%20group.
3. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.pdf
4. Latif S., Zaidi A., Cuayahuitl H., Shamshad F., Shoukat M., Qadir J. Transformers in Speech Processing: A Survey [electronic resource] / ArXiv. – 2023. – Available from: https://arxiv.org/pdf/2303.11607 (accessed: 02.02.2025).
5. Bai G., Guo H., Xiao C. Research on the application of transformer in computer vision [electronic resource] // J. Phys.: Conf. Ser. – 2023. – Vol. 2649, No. 1, 012033. – DOI: 10.1088/1742-6596/2649/1/012033. – Available from: https://www.researchgate.net/publication/376303699_Research_on_the_application_of_transformer_in_computer_vision (accessed: 07.12.2023).
6. Schmidt L., Weeds J., Higgins J. P. T. Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks [electronic resource] / arXiv. – 2020. – Journal reference: HEALTHINF 2020. – DOI: 10.48550/arXiv.2001.11268. – Available from: https://arxiv.org/abs/2001.11268 (accessed: 02.02.2025).
7. https://www.nroadcorp.com/free-whitepaper/
8. Kalambe, Y. S., Pratiba, D., & Shah, P. (2015). Big Data Mining Tools for Unstructured Data: A Review. International Journal of Innovative Technology and Research, 3(2), 2012–2017.

9. Linden, A., & Yarnold, P. R. (2016). Using data mining techniques to characterize participation in observational studies. Journal of Evaluation in Clinical Practice, 22(6), 835–843.
10. Saha, S., & Srivastava, J. (2014). An Exclusive Study on Unstructured Data Mining with Big Data. International Journal of Computer Applications, 94(11), 1–5.
11. Fonger, F., et al. (2024). Process Mining for Unstructured Data: Challenges and Research Directions. arXiv preprint arXiv:2401.13677.
12. Singhal, K. (2015). A Detailed Approach for Data Mining and Clustering of Unstructured Data. International Journal of Advanced Pharmaceutical Research and Review, 5(1), 28–45.
13. Gulati, A., et al. (2021). *A Comparative Study of End-to-End Architectures for Automatic Speech Recognition*. *arXiv preprint arXiv:2111.01690*. Available at arXiv.org.
14. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.
15. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.,N., Kaiser,Ł., Polosukhin,I. Attention is All You Need. Advances in Neural Information Processing Systems; 2017; Vol. 30, pp. 5998--6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.
16. Methods of constructing algorithms for comparative test statistical verification of mathematical models of bioobject responses to low-intensity stimuli / Bohdan Yavorskyy, Evhenia Yavorska, Halyna Tsupryk, Roman Kinash // Scientific Journal of TNTU. — Tern.: TNTU, 2023. — Vol 112. — No 4. — P. 82–90.