1.Dariusz WÓJCIK[1,2], 2.Dariusz MAJEREK[3]

[1]Netrix S.A., Research & Development Centre, Lublin, Poland, [2]University of Economics and Innovation in Lublin, Poland
[3]Lublin University of Technology, Poland
ORCID: 1. 0000-0002-4200-3432, 2. 0000-0002-0035-7187

doi:10.15199/48.2023.12.69

Uncovering insights with kanguage modeling. Analyzing Przeglad Elektrotechniczny with GPT

Abstract. *Recent advancements in natural language processing (NLP) have led to the development of powerful language models, which have significantly impacted various applications such as text classification, sentiment analysis, and language translation. This paper explores the application of the GPT-2 large language model to analyze the text of articles published in the 2023 edition of the Przegląd Elektrotechniczny journal. The model was fine-tuned on the entire text corpus, and a word analysis was performed to identify the most frequent words and their relationship to the articles. Our method uncovers insights into the most common topics, themes, and ideas discussed in the journal, offering valuable information for the Editorial Board and researchers to better comprehend the state of electrical engineering in 2023*

Streszczenie. *Ostatnie postępy w przetwarzaniu języka naturalnego (NLP) doprowadziły do opracowania potężnych modeli językowych, które znacząco wpłynęły na różne zastosowania, takie jak klasyfikacja tekstu, analiza nastrojów i tłumaczenie językowe. W niniejszym artykule zbadano zastosowanie dużego modelu językowego GPT-2 do analizy tekstu artykułów opublikowanych w wydaniu czasopisma Przegląd Elektrotechniczny z 2023 roku. Model został dostrojony do całego korpusu tekstowego i przeprowadzono analizę słów w celu zidentyfikowania najczęściej występujących słów i ich związku z artykułami. Nasza metoda odkrywa najczęstsze tematy, motywy i pomysły omawiane w czasopiśmie, oferując cenne informacje dla redakcji i naukowców, aby lepiej zrozumieć stan inżynierii elektrycznej w 2023 roku (**Odkrywanie spostrzeżeń za pomocą modelowania języka: Analiza przeglądu elektrotechnicznego za pomocą GPT**).*

Keywords: Please insert 3 – 4 keywords or phrases.
Słowa kluczowe: please use Google Translation.

Introduction

Many algorithms are used to analyze data from many different sources [1-11]. The rapid advancements in natural language processing (NLP) have led to the development of numerous robust language models, such as Transformers, Language Model (LM), and Generative Pre-trained Transformer (GPT) [12-13]. These models have dramatically impacted various applications, including text classification, sentiment analysis, and language translation. The key to their success lies in their ability to learn contextual relationships between words and generate high-quality text outputs. Generative Pre-trained Transformer (GPT) models, part of the LM family, are constructed on transformer architecture and extensively pre-trained on vast amounts of textual data. As illustrated in Fig. 1, these models can have various sizes and can be fine-tuned for various downstream tasks, such as language modelling, text creation, and machine translation [14-15]. GPT models have established themselves as an essential standard in NLP by achieving state-of-the-art performance on numerous benchmark datasets.

In this paper, we investigate the text from a journal using a specific type of language model, the GPT-2 Large Language Model (LLM). We aim to fine-tune the GPT-2 LLM on the entire text corpus of the Przegląd Elektrotechniczny journal's 2023 edition and perform an in-depth word analysis to identify the most frequent words and their relationship with the articles. By leveraging advanced NLP techniques, our findings intend to provide valuable insights into the journal's content, revealing prevalent topics, themes, and ideas discussed in the articles.

Methods

We began our process by obtaining a pre-trained GPT-2 model with 774 million parameters from the Hugging Face library [16]. We then collected PDFs of all articles from the journal website, converted them to text format, and cleaned the data by removing irrelevant characters and symbols and transforming the text to lowercase. To analyze the text data from the articles, we employed the following method:

1. Data Collection and Preprocessing: We obtained the article texts from a website, from which we extracted the text data for further cleaning. This entailed the removal of unwanted characters, symbols, and line breaks while retaining the alphanumeric data. We also eliminated the "REFERENCES" section from the text to focus our analysis on the main content of the articles. After cleaning, the text data was converted into a list.
2. Dataset Preparation: Using the cleaned text data, we constructed a Hugging Face Dataset object. This conversion facilitated the application of tokenizer and training functions from the transformers library.
3. Model Selection, Tokenization, and Traning: We employed the GPT-2 Large pre-trained model along with its associated tokenizer. The tokenizer was configured to pad the input sequences from the right, with the padding token being set as the eos_token.
4. After fine-tuning, we prompted the model with a description from the journal website and generated 512 tokens 20 times, yielding 10240 tokens, maintaining a 2.0 temperature and top_p as 0.9 settings to promote diverse and creative outputs. We collected all generated tokens and visualized them in a Word Cloud, as shown in Fig. 2, where more frequent words appeared larger.

Sentence embeddings a sentence similarity

Sentence embeddings are numerical representations of sentences in a high-dimensional vector space. These embeddings are designed to capture a sentence's semantic and contextual information, allowing for meaningful comparisons and analyses of text data. In essence, sentence embeddings transform text into numerical vectors that can be used for various natural language processing (NLP) tasks such as text classification, similarity measurement, and clustering.

Using a pre-trained language model, we created sentence embeddings for a set of generated sentences. The process involved the following steps:

Tokenization: The generated

1. Sentences, were tokenized. Tokenization breaks down the sentences into individual tokens or subwords, preparing them for processing by the language model.

2. **Model Inference**: The tokenized sentences were fed into the pre-trained language model. During model inference, the model processed the input tokens and produced a range of outputs, including hidden states that encapsulate the contextual information of each token in the input.
3. **Hidden States**: The code extracted the hidden states from the model's output. These hidden states represent the internal representations of the sentences at different layers of the language model. In this case, the hidden states from the last layer were used, capturing the sentences' most refined and abstract representations.
4. **Mean Pooling**: To obtain a single vector representation for each sentence, mean pooling was applied to the hidden states of the last layer. Mean pooling involves taking the average of all token embeddings in the sentence. This process condenses the contextual information into a single vector, creating a sentence embedding.

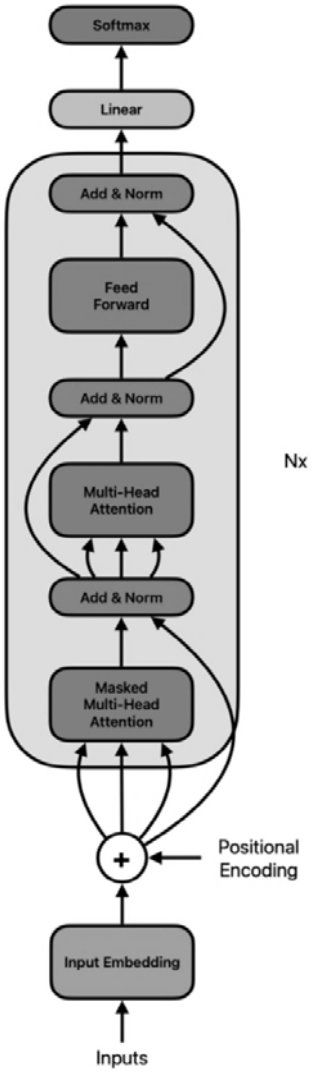

Fig. 1. The decoder from the transformer model - a building block of GPT [1].

The resulting sentence embeddings are now numerical representations that encode the semantics and context of the generated sentences. These embeddings were utilized for measuring semantic similarity, identifying patterns, and making inferences based on the content and context of the sentences.

We computed the cosine similarity between the embeddings of the generated sentences and the original sentences. Cosine similarity is a measure of similarity between two vectors, with values ranging from 0 (completely dissimilar) to 1 (identical).

In the context of text analysis, it calculates the similarity between two sentence embeddings, A and B, where A and B are vectors in the embedding space. The cosine similarity between these vectors is computed as follows:

$$cossim(A,B) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

where $A \cdot B$ means the inner product of A, B.

This metric is frequently used in natural language processing to assess the semantic similarity between text fragments. The cosine similarity was computed for each pair of generated and original sentence embeddings, resulting in a similarity matrix.

A similarity heatmap was created using the matplotlib library to represent the computed cosine similarities visually. The heatmap displayed in Fig. 3 shows the pairwise cosine similarity scores as a grayscale-coded grid. The x-axis of the heatmap represented the original sentences, while the y-axis corresponded to the generated sentences. The grayscale intensity indicated the degree of similarity between each pair of sentences.

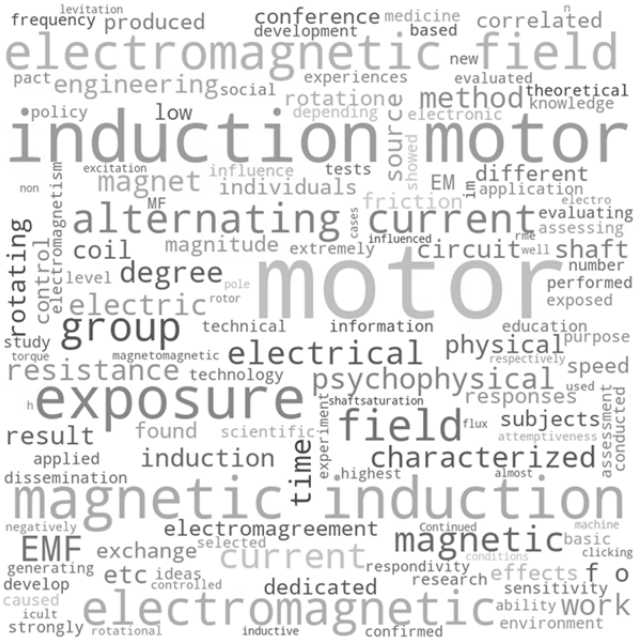

Fig. 2. Word cloud created from text generated by the fine-tuned model.

Results

In analysis of the text from the journal, we have obtained a collection of significant words and phrases that provide valuable insights into the predominant themes and trends within the journal's content. These results illuminate key aspects of the field and contribute to a better understanding of the research landscape in electrical engineering. Below are the main keywords from the word cloud

1. The term "Motor" signifies a fundamental component in electrical engineering, indicating a focus on various types of electric motors. Electric motors are essential for converting electrical energy into mechanical energy, and they have widespread applications in various industries and technologies.
2. "Induction Motor" points to a specific type of electric motor commonly used in industrial and commercial applications. This result suggests a particular emphasis on

the study and development of induction motors within the electrical engineering domain.
3. "Alternating Current" is a critical concept in electrical engineering. AC represents the type of electrical current that periodically changes direction, and it is the standard form of electricity used in households and most power transmission systems. This result indicates a focus on AC electrical systems and related research.
4. "Magnetic Induction" refers to the process of inducing an electromotive force (EMF) in a conductor through changes in magnetic flux. This concept is fundamental to the operation of transformers and generators, illustrating a focus on electromagnetic principles and their applications.
5. The term "Exposure" suggests considering the potential exposure of individuals or equipment to electromagnetic fields or radiation. This keyword may indicate research on safety and health concerns associated with electromagnetic fields.
6. “Electromagnetic Field" represents a central concept in electrical engineering, describing the spatial distribution of electromagnetic forces and effects. This result highlights a focus on the characterization and analysis of electromagnetic fields in various contexts.

Collectively, these keywords illustrate a diverse and multidimensional research landscape within electrical engineering. The prominence of terms related to motors, electromagnetic fields, and magnetic induction emphasises fundamental electrical principles and practical applications. Additionally, including "Exposure" suggests considering safety and health aspects associated with electromagnetic technologies. These results provide a valuable snapshot of the key areas of interest and exploration within the field, facilitating a deeper understanding of the research focus in electrical engineering.

The cosine similarity analysis (shown as a heatmap in Fig. 3) between generated sentences and sentences extracted from journal articles revealed intriguing patterns in the relationship among these sentences. Notably, a distinct cluster of generated sentences, located in the bottom-right corner of the cosine similarity heatmap (ranging from 200 to 228), exhibited high similarity to each other. These sentences also displayed a higher degree of similarity to the sentences extracted from the journal articles. This means that the training of the model succeeded, and it allows the generation of sentences within the topic of the journal.

The phenomenon worth noting in these findings is the presence of articles in which all sentences exhibit high similarity throughout the entire article. Such uniformity in the similarity scores across sentences within an article suggests a strong thematic coherence within those articles. It implies that the content of these articles might predominantly revolve around a specific central topic or idea, with minimal variation in the language used to express concepts.

This observation could signify several possibilities:
1. **Focused Research**: Articles with similar sentences might be centered around a single, tightly focused research topic. Researchers within the field may collectively explore and build upon a specific area of interest or a particular aspect of electrical engineering.
2. **Technical Jargon**: It's plausible that the consistent use of technical jargon or specialized terminology within the articles contributes to the high similarity.
3. **Methodological Consistency**: Articles exhibiting uniform sentence similarity may adhere to a common research methodology or approach. This could indicate a standardized approach to experimentation or analysis within certain subfields.
4. **Editorial Policies**: The similarity in sentences may also be influenced by the journal's editorial policies. Journals may encourage authors to adhere to specific writing styles or guidelines, resulting in a more homogenous expression of research findings.

Discussion

In this study, we harnessed the formidable capabilities of the GPT-2 Large Language Model to delve into the realm of text analysis within the specialized domain of electrical engineering. Our endeavour encompassed fine-tuning a large language model to ensure its adaptability and suitability for the targeted journal's content. Subsequently, we embarked on a comprehensive word analysis, delving deep into the journal's corpus to unearth the most pertinent keywords. These keywords, which emerged through meticulous analysis, provide a crucial insight into the current trends and focal points within the field of electrical engineering.

Our results testify to the potential prowess of pre-trained language models in text analysis. By drawing on the inherent ability of GPT-2 to understand and generate text, we successfully identified the essential topics and themes that permeate the journal's articles. This discovery sheds light on the multifaceted landscape of electrical engineering and underscores the immense potential of such models to extract meaningful insights from a vast sea of textual information.

Furthermore, the findings of our study possess practical significance, particularly for the Editorial Board of the journal under investigation. The identified keywords and themes can act as a compass, guiding the editorial decision-making process and enabling the Board to steer the journal toward emerging trends and crucial developments within the field. This approach maintains the journal's relevance and ensures that it remains at the forefront of disseminating the latest research in electrical engineering.

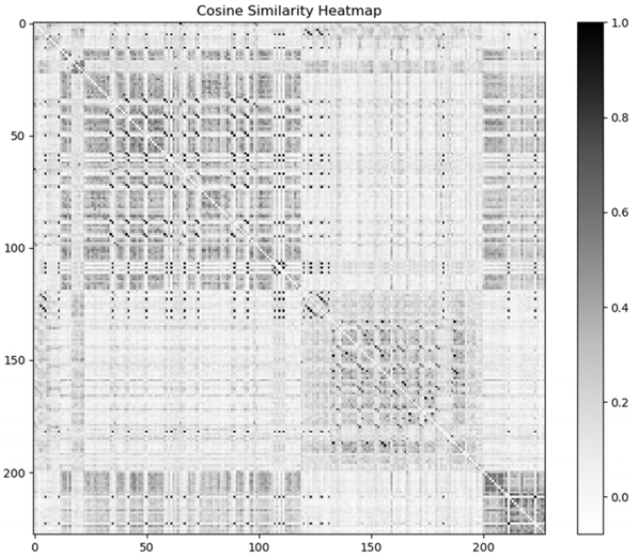

Figure 3 Word cloud created from text generated by the fine-tuned model.

Summary

In summary, our research underscores the formidable utility of advanced language models like GPT-2 when applied to the nuanced landscape of text analysis within specialized domains such as electrical engineering. Through this application, we have unveiled a rich tapestry of insights that possess the potential to stimulate and reshape scientific progress within the field.

Crucially, our findings reveal intriguing patterns in sentence similarity, offering a glimpse into the dynamic nature of research within electrical engineering. The presence of distinct clusters of sentences, some demonstrating high coherence and others marked by variability, hints at the multifaceted character of the field. We observe articles where sentences harmonize, reflecting a singular thematic focus, while others exhibit diversity, suggesting a spectrum of research exploration.

This nuanced understanding of the research landscape underscores the pivotal role of language models as indispensable tools in the researcher's arsenal. These models facilitate knowledge discovery, interpretation, and dissemination, transcending traditional boundaries. Integrating such models into research methodologies becomes advantageous and imperative in an era marked by the evolution of science and technology.

Our study advances the notion that advanced language models are not mere tools but catalysts for innovation and discovery. By harnessing their capabilities, we augment our comprehension of intricate domains like electrical engineering and pave the way toward a future defined by limitless innovation and scholarly advancement.

***Authors**: dr inż. Dariusz Wójcik, Akademia WSEI, Wydział transportu i informatyki, ul. Projektowa 4, 20-209 Lublin, E-mail: dariusz.wojcik@wsei.lublin.pl; dr Dariusz Majerek, Politechnika Lubelska, Wydział Podstaw Techniki, ul. Nadbystrzycka 38, 20-618 Lublin, E-mail: d.majerek@pollub.pl*

REFERENCES

[1] Gnaś, D., Adamkiewicz, P., Indoor localization system using UWB, Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, 12 (2022), No. 1, 15-19.

[2] Styła, M., Adamkiewicz, P., Optimisation of commercial building management processes using user behaviour analysis systems supported by computational intelligence and RTI, Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, 12 (2022), No 1, 28-35.

[3] Kropidłowska P., Irzmańska E. Korzeniewska E., Tomczyk M., Jurczyk-Kowalska M., Evaluation of laser texturing in fabricating cut-resistant surfaces for protective gloves, Textile Research Journal, 93 (2023), No. 9-10), 1917–1927.

[4] Koulountzios P., Aghajanian S., Rymarczyk T., Koiranen T., Soleimani M., An Ultrasound Tomography Method for Monitoring CO2 Capture Process Involving Stirring and CaCO3 Precipitation, Sensors, 21 (2021), No. 21, 6995.

[5] Kłosowski G, Rymarczyk T, Niderla K, Kulisz M, Skowron Ł, Soleimani M., Using an LSTM network to monitor industrial reactors using electrical capacitance and impedance tomography – a hybrid approach. Eksploatacja i Niezawodnosc – Maintenance and Reliability, 25 (2023), No. 1, 11.

[6] Kłosowski G., Rymarczyk T., Kania K., Świć A., Cieplak T., Maintenance of industrial reactors supported by deep learning driven ultrasound tomography, Eksploatacja i Niezawodnosc – Maintenance and Reliability; 22 (2020), No 1, 138–147.

[7] Rymarczyk T., Kłosowski G., Hoła A., Sikora J., Tchórzewski P., Skowron Ł., Optimising the Use of Machine Learning Algorithms in Electrical Tomography of Building Walls: Pixel Oriented Ensemble Approach, Measurement, 188 (2022), 110581.

[8] Pawłowski S., Plewako J., Korzeniewska E., Field Modeling of the Influence of Defects Caused by Bending of Conductive Textronic Layers on Their Electrical Conductivity, Sensors, 23 (2023), No. 3, 1487.

[9] Rymarczyk T., Kłosowski G., Hoła A., Hoła J., Sikora J., Tchórzewski P., Skowron Ł., Historical Buildings Dampness Analysis Using Electrical Tomography and Machine Learning Algorithms, Energies, 14 (2021), No. 5, 1307.

[10] Kłosowski G., Rymarczyk T., Niderla K., Rzemieniak M., Dmowski A., Maj M., Comparison of Machine Learning Methods for Image Reconstruction Using the LSTM Classifier in Industrial Electrical Tomography, Energies 2021, 14 (2021), No. 21, 7269.

[11] Koulountzios P., Rymarczyk T., Soleimani M., A triple-modality ultrasound computed tomography based on full-waveform data for industrial processes, IEEE Sensors Journal, 21 (2021), No. 18, 20896-20909.

[12] Vaswani, A., et al., Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA,* (2017), *6000–6010.*

[13] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, (2019), 4171–4186.

[14] Radford A. et al., Improving Language Understanding by Generative Pre-Training, online: https://openai.com/research/language-unsupervised, (2018).

[15] Brown, T., et al., Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., (2020), 1877–1901.

[16] Wolf, T., et al, HuggingFace's transformers: State-of-the-art natural language processing. ArXiv preprint arXiv:1910.03771, (2019).