EasyChair Preprint

№ 9777

Artificial Intelligence in Data Analysis for Open-Source Investigations

Teodor-Cristian Radoi

February 26, 2023

Artificial Intelligence in Data Analysis for Open-Source Investigations

Teodor-Cristian Rădoi
Cybersecurity Department
Web Vortex
Bucharest, Romania
radoi.teodor.cristian@gmail.com

***Abstract*— Open source investigations are challenging due to the vast amounts of data to verify and the likelihood of encountering incorrect data. To address these issues, an agent needs a tool that can process data in real-time and simplify the processing of large volumes of information. This can be achieved by integrating a GPT model that learns from the data it processes. The study introduces an OSINT platform that uses a GPT model to enhance the efficiency of open source investigations.**

The OSINT platform developed by us organizes data in a hierarchical graph, with a particularly interesting feature: the integration of a GPT model, which allows the user to process large data faster and more easily. To communicate with this GPT model, the user may chat with a virtual agent in natural language to give data processing commands.

The study assessed different natural language processing models, including BERT and GPT models, and focused on the benefits of pretraining, fine-tuning, and generative models for open source investigations. GPT models have an advantage in pretraining, allowing them to capture complex relationships between words and phrases. This pretraining makes the models customizable for specific tasks, providing investigators with a powerful tool for analyzing text data.

The generative nature of GPT models is a key advantage for OSINT investigations, as it allows the model to generate human-like text for analyzing data. Fine-tuning is also critical, as it enables investigators to train the model on specific topics and customize it to their needs. By using natural language processing models in open source investigations, investigators can generate more accurate and reliable results while reducing the time and effort required for data analysis. Overall, this work highlights the importance of incorporating natural language processing models in OSINT investigations and provides a foundation for future research in this field.

Keywords— GPT, BERT, Artificial Intelligence, Open Source Intelligence, Information

I. Introduction

Open source investigations require a significant amount of time and effort from investigators, as they must carefully analyze vast amounts of information found on the internet and in various media sources. While tools like OSINT frameworks, Maltego, social media, and government websites can assist investigators in their work, they also present a couple of significant challenges: dealing with data overload and verifying the accuracy of the data.

For instance, if an investigator is researching a company's director and comes across ten articles containing relevant information, they must sift through each one individually, verify the accuracy of any mentioned individuals, and ensure that no false information is present. This is a time-consuming and daunting task that can be overwhelming, especially when dealing with large amounts of data.

The emergence of artificial intelligence and the development of advanced natural language processing models like GPT (Generative Pre-trained Transformer) have the potential to revolutionize open source investigations. GPT models are based on generative transformers and are trained on vast amounts of data, making them capable of filtering, summarizing, and assisting investigators in processing and analyzing vast amounts of data quickly and accurately.

By leveraging the power of GPT models, investigators can streamline their research process, save time and resources, and enhance the accuracy and reliability of their findings. In short, GPT has the potential to transform the way open source investigations are conducted, making them faster, more efficient, and more effective than ever before.

II. Transformers

Transformers are a class of neural network models that have revolutionized natural language processing tasks, including language translation, question-answering, and text generation. The origins of transformer models can be traced back to a paper by Vaswani in 2017, where they introduced the transformer architecture for sequence-to-sequence modeling tasks. Prior to the introduction of transformers, most state-of-the-art language models relied on recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

Transformers are unique in that they do not rely on sequential processing and have the ability to parallelize computations, making them highly efficient for processing large amounts of data. The transformer model is based on an encoder-decoder architecture, where the encoder processes the input sequence and the decoder generates the output sequence.

The transformer encoder-decoder architecture consists of multiple layers of self-attention and feedforward neural networks. The input sequence is first embedded into a high-dimensional vector space, and then passed through a series of encoder layers. Each encoder layer consists of two sub-layers: a multi-head self-attention mechanism and a position-wise feedforward neural network.

The multi-head self-attention mechanism allows the model to attend to different parts of the input sequence to compute a weighted sum of the values, which is used to generate a context vector. The feedforward neural network applies a non-linear transformation to the context vector to create a new representation of the input sequence.

The decoder, which is similar to the encoder, consists of two sub-layers: masked multi-head self-attention and encoder-decoder attention. The masked multi-head self-attention mechanism ensures that the decoder can only attend to positions before the current position, while the encoder-decoder attention mechanism allows the decoder to attend to the encoder's output.

The transformer model's self-attention mechanism is the key to its success, allowing it to capture dependencies between different parts of the input sequence without relying on sequential processing. This property enables transformer models to efficiently process large amounts of data and achieve state-of-the-art performance on various natural language processing tasks.

III. Self-Attention Layers

Self-attention is a mechanism that allows the transformer model to weigh the importance of different positions in the input sequence when computing the representation of a given position. This mechanism enables the model to focus on the most relevant parts of the input sequence while ignoring irrelevant or redundant information, which is crucial for natural language processing tasks.[1]

In a self-attention layer, the input sequence is transformed into three vectors: the query vector, the key vector, and the value vector. These vectors are then used to compute a weighted sum of the values, where the weights are determined by the similarity between the query and key vectors. The resulting weighted sum is known as the context vector and is used to compute the representation of each position in the input sequence.

What makes self-attention so special is its ability to capture long-range dependencies between different parts of the input sequence without relying on sequential processing. This is achieved by computing the similarity between the query and key vectors for all positions in the input sequence in parallel. This property makes the transformer model highly efficient for processing large amounts of data and enables it to capture complex relationships between different parts of the input sequence that are difficult to capture using traditional neural network architectures.[2]

Moreover, self-attention enables the transformer model to capture context-dependent dependencies between different parts of the input sequence. This is achieved by allowing the query vector to vary based on the current position, enabling the model to focus on different parts of the input sequence depending on the current context. This property is especially important for natural language processing tasks, where the meaning of a word can vary depending on the surrounding words and the overall context. By capturing context-dependent dependencies between different parts of the input sequence, the transformer model can generate more accurate and meaningful representations of natural language.

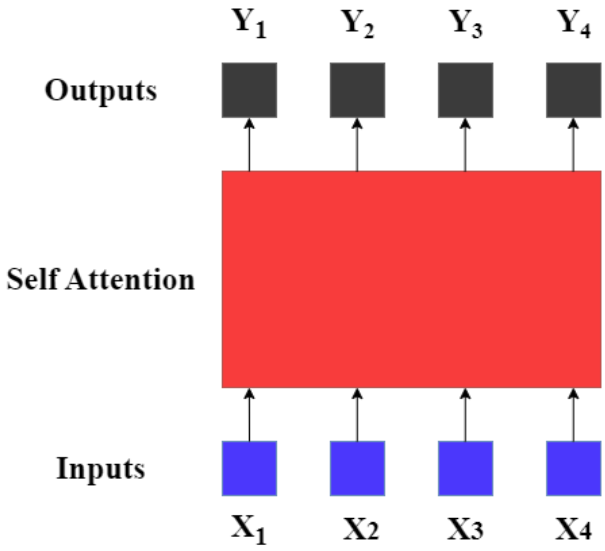

Figure 1. Self-Attention Layer Example

$$y_i = \sum w_{ij} * x_j \tag{1}$$

$$w'_{ij} = x_i * x_j \tag{2}$$

$$w_{ij} = \frac{exp(w'_{ij})}{\sum exp(w'_{ij})} \tag{3}$$

(1) Prediction equation, note that we can predict in every step, exactly like RNNs.

(2) Dot product between two inputs will give us a temporary value used to compute the weight.

(3) The weight equation where we use the temporary values resulted before, notice that it is the same equation as softmax.

The inputs are first embedded into a high-dimensional vector space, and then passed through multiple layers of self-attention and feedforward neural networks. The self-attention mechanism computes a set of weights w_{ij} for each input position x_i, which determines how much attention should be paid to other positions x_j when generating the output y_j.

The attention weights are computed using a dot product between a query vector q_i, which represents the current input position, and a set of key vectors k_j, which represent all the other input positions. The resulting scores are then normalized using a *softmax* function, which ensures that the weights sum up to 1. The final context vector is computed as a weighted sum of the value vectors v_j, where the weights are given by the attention weights.

The computed context vector is then passed through a feedforward neural network to generate the final output y_i. This process is repeated for each output position, with the weights and context vectors computed independently for each position. By allowing each output position to attend to different parts of the input sequence, the transformer model can capture complex dependencies between input and output positions, making it highly effective for a wide range of natural language processing tasks.

In summary, the equations used in transformers allow the model to dynamically determine the importance of different input positions when generating each output position. By computing attention weights based on the similarity between query and key vectors, the model can capture complex relationships between input and output positions, making it a powerful tool for natural language processing.

IV. BERT AS INVESTIGATION AGENT

Bidirectional Encoder Representations from Transformers (BERT) is a natural language processing model developed by Google in 2018. BERT is based on the transformer architecture and has achieved state-of-the-art performance on various natural language processing tasks, including language translation, question-answering, and text classification. The model is trained on large amounts of text data, allowing it to capture complex relationships between different words and phrases.[4]

One of the key features of BERT is its ability to perform bidirectional processing, meaning it can take into account the context of a word or phrase by analyzing the words that come before and after it. This is achieved by training the model on a masked language modeling task, where a certain percentage of words in a sentence are randomly masked, and the model is trained to predict the masked words based on the context of the surrounding words.[2]

BERT also utilizes a next sentence prediction task, where the model is trained to predict whether two sentences are likely to appear in sequence. This task allows the model to capture relationships between different sentences and enables it to generate more coherent and meaningful outputs.

In the context of open source investigations, BERT has the potential to improve the accuracy and efficiency of information retrieval and analysis. By leveraging the power of BERT, investigators can quickly sift through large amounts of textual data, extract relevant information, and identify relationships between different pieces of information. This can help investigators to generate more accurate and comprehensive reports in a shorter amount of time.

Some of the key usage of BERT include:

- *Language Translation:* BERT can be used to translate text from one language to another by training the model on a large parallel corpus of text in both languages. BERT can capture the subtle nuances of language and context, enabling it to generate accurate translations.[3]
- *Text Classification:* BERT can be used to classify text into different categories, such as sentiment analysis, spam detection, and topic modeling. BERT's ability to capture context and relationships between words makes it highly effective for these types of tasks.
- *Question Answering:* BERT can be used to answer natural language questions by training the model on a large corpus of text and the corresponding questions and answers. BERT's ability to capture the relationships between words and phrases enables it to generate accurate answers to a wide range of questions.
- *Named Entity Recognition:* BERT can be used to identify and extract named entities from text data, such as people, organizations, and locations. BERT's ability to capture context and relationships between words makes it highly effective for this type of task.

The ability of BERT to capture the subtle nuances of language and context is particularly valuable for OSINT investigations, where investigators often encounter text data in foreign languages. By training BERT on a large corpus of parallel text in both languages, investigators can use the model to translate articles and other text data with high accuracy, enabling them to quickly extract relevant information and gain a deeper understanding of the topic or event being investigated.

In addition to language translation, BERT can also be used for text classification, question answering, and named entity recognition in OSINT investigations. For example, investigators can use BERT to classify social media posts or news articles based on their sentiment, enabling them to identify trends or potential threats related to a particular topic or event. BERT can also be used to extract named entities, such as people, organizations, or locations, from text data, which can help investigators to identify important players and relationships in a given context.

Overall, BERT's ability to capture complex relationships between words and phrases in natural language makes it a valuable tool for OSINT investigations. By leveraging the power of BERT, investigators can quickly and accurately process large amounts of textual data, extract relevant information, and identify relationships between different pieces of information, which can help them to generate more accurate and comprehensive reports in a shorter amount of time.[4]

Although it has proven to be a successful model, it has several disadvantages:

- *Lack of Interpretability:* Another potential disadvantage of BERT is that it can be difficult to interpret how the model is making its predictions. The model is based on complex mathematical operations and high-dimensional vector spaces, which can make it challenging to understand how it is making decisions.
- *Not Always the Best Model:* While BERT has achieved state-of-the-art performance on a wide range of natural language processing tasks, it is not always the best model for every task. Depending on the specific requirements of a given task or domain, other models may be more effective or efficient.
- *Language-Specific:* While BERT can be trained on multiple languages, it is designed for specific languages and may not be as effective or accurate for languages that it was not trained on. This can limit its applicability in certain contexts or for investigating topics that are primarily discussed in non-English languages.

The lack of interpretability in BERT can make it challenging for investigators to understand how the model is making its predictions, which can have implications for the accuracy and credibility of the results in OSINT investigations. In order to ensure the reliability of the analysis and conclusions drawn from BERT-generated output, investigators must have a clear understanding of how the model is arriving at its predictions.

For example, if a BERT-based sentiment analysis model identifies a particular social media post as being positive, but

the investigator cannot understand why the model has made this determination, it can be difficult to determine the reliability of the result. It may be that the model is identifying subtle nuances in the language that the investigator is not aware of, but it may also be the case that the model is making a mistake or being influenced by external factors.

In order to mitigate the lack of interpretability in BERT, investigators can use techniques like visualization tools, sensitivity analysis, and data perturbation to understand how the model is making its predictions. By gaining a deeper understanding of the factors that are influencing the model's output, investigators can have greater confidence in the accuracy and reliability of the results.

The language-specific limitation of BERT can be a potential challenge for OSINT investigations, as it can limit the model's applicability for investigating topics that are primarily discussed in non-English languages. While BERT can be trained on multiple languages, it is designed for specific languages and may not be as effective or accurate for languages that it was not trained on.

This can be especially challenging in cases where investigators are analyzing text data in languages that are not commonly used or where there are limited resources available for training and fine-tuning BERT. In such cases, investigators may need to rely on other models or techniques for analyzing the text data, which may be less accurate or efficient than BERT.

Moreover, the language-specific limitation of BERT can also have implications for cross-lingual analysis and the identification of relationships and trends across different languages. For example, if investigators are analyzing text data in multiple languages to identify common themes or patterns, the limitations of BERT for certain languages can limit the accuracy and comprehensiveness of the analysis.

To mitigate the language-specific limitation of BERT in OSINT investigations, investigators can consider training and fine-tuning the model on specific languages or using other models and techniques that are designed for multilingual analysis. Additionally, investigators can collaborate with experts in specific languages or regions to ensure that their analysis is accurate and comprehensive.

V. GPT AS INVESTIGATION AGENT

Generative Pre-trained Transformer (GPT) is a natural language processing model developed by OpenAI. The model is based on the transformer architecture and is capable of generating human-like text by predicting the next word in a sentence. GPT has achieved state-of-the-art performance on various natural language processing tasks, including text generation, language translation, and question-answering. The model is trained on large amounts of text data, allowing it to capture complex relationships between different words and phrases.

The GPT model has been released in two versions so far, GPT-2 and GPT-3. GPT-2, released in 2019, was trained on a massive corpus of text data and is capable of generating high-quality human-like text. However, due to concerns about the potential misuse of the model for generating fake or misleading information, OpenAI chose not to release the full version of the model to the public.

GPT-3, released in 2020, is the most powerful version of the model to date and has achieved state-of-the-art performance on a wide range of natural language processing tasks. GPT-3 is capable of generating highly coherent and convincing text, making it a promising tool for language translation, text classification, and other applications in OSINT investigations. Despite its impressive performance, however, GPT-3 has also raised concerns about the potential misuse of the model for generating fake or misleading information.

There are several key factors that differentiate GPT from other natural language processing models:

- *Generative:* One of the main features of GPT is that it is a generative model, meaning that it is capable of generating human-like text. This is in contrast to other models that are designed for tasks like classification or question-answering, where the model produces a single output based on the input.[5]
- *Unsupervised Learning[6]:* GPT is trained using unsupervised learning, meaning that it is not given explicit labels or annotations during training. Instead, the model is trained to predict the next word in a sequence of text data, allowing it to capture complex relationships between words and phrases.
- *Pre-trained:* GPT is a pre-trained model, meaning that it is trained on a large corpus of text data before being fine-tuned for a specific task. This pre-training allows the model to capture a wide range of linguistic knowledge and enables it to be adapted to different applications and domains.[5]
- *Large Scale:* GPT is trained on a massive corpus of text data, allowing it to capture complex relationships and patterns in natural language. The large scale of the model also enables it to generate highly coherent and convincing text, making it a promising tool for a wide range of natural language processing tasks.
- *Contextual:* GPT is designed to capture the context and relationships between different words and phrases in natural language. This enables the model to generate text that is highly coherent and consistent with the surrounding context, making it a powerful tool for language translation, text classification, and other applications in OSINT investigations.

One of the key advantages of GPT for OSINT investigations is that the model is pre-trained and ready to be fine-tuned for specific applications and domains. This means that investigators can teach the model to think critically and process data as a real OSINT investigator, by providing it with large amounts of text data related to the target or topic being investigated.

Moreover, because GPT is capable of learning unsupervised[6], investigators can train the model using large articles about the targets and previous reports created by the user agent. This can provide the model with a rich source of data and enable it to capture complex relationships and patterns in the data, which can help it to make more accurate and informed predictions.

By leveraging the power of GPT in this way, investigators can create an artificial agent that is capable of processing and analyzing large amounts of text data, identifying relevant

information, and making informed predictions based on the available evidence. This can help investigators to reduce the amount of time and effort required for data analysis and enable them to generate more accurate and comprehensive reports in a shorter amount of time. Ultimately, the ability to fine-tune GPT for specific OSINT applications can help to improve the accuracy and reliability of the analysis, making it a valuable tool for OSINT investigations.

VI. OSINT Platform with GPT Agent

The OSINT platform that we have developed is a SaaS product that is structured as a web application, providing users with a graph-based UI for data visualization and analysis. The platform is designed to support hierarchical data structuring, with each node in the graph representing a piece of data such as an article, website, or IP address. These data nodes are connected by one-way edges, creating a tree-like structure that supports the organization of information during an investigation.

In order to enrich the data on the platform, we have implemented a transform process. A transform is a custom operation that can be applied to a node in order to extract additional data or perform advanced processing. These transforms are implemented using Recon-ng modules, a powerful OSINT framework that enables us to scrape data from multiple sources and search engines.

The application is designed to run on a PHP backend infrastructure, which provides a scalable and flexible framework for processing and managing data. The GPT module is implemented as a chatbot, enabling users to interact with the platform using natural language commands. For example, a user could select a set of articles related to a particular company[7] and ask the bot to "Get me the persons involved in these articles." The bot would respond with a custom transform that would generate a new node as the child of one of the selected nodes, containing information on the individuals involved in the articles.

One significant problem that we have encountered in the development of our OSINT platform is the need for communication with third-party servers, such as those operated by OpenAI. This creates a potential security risk, as sensitive data could be transmitted outside the client's area or network. In certain OSINT investigations, such as those involving government agencies, the need to keep sensitive data within the client's network is of paramount importance[8]. Transmitting such data to a third-party server is simply not an option.

To mitigate this issue, we have explored the possibility of running a GPT model on either the client's server or on our own server. However, this solution is not feasible at the moment, as running a GPT model requires a significant amount of computing resources and would be prohibitively expensive for most clients. Therefore, we have implemented a robust security protocol to ensure that data is transmitted securely and only to trusted servers.

It is worth noting that in certain governmental OSINT investigations, such as those conducted by intelligence agencies, the sensitivity of the data being analyzed is even higher. In these cases, the risks associated with transmitting data to third-party servers are even greater. As such, it is critical for OSINT platforms to have strong security protocols in place and to be designed to operate entirely within a secure network. This can help to ensure the integrity and confidentiality of the data being analyzed, while also minimizing the risks associated with the transmission of sensitive data.

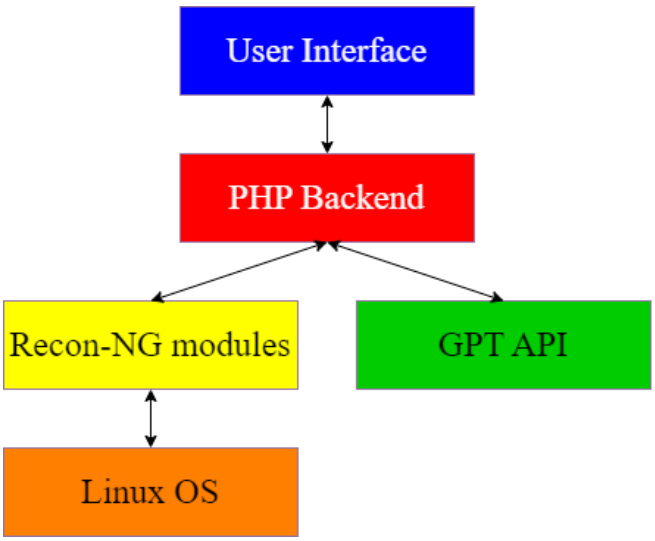

Figure 2. Web Application Architecture

Overall, the OSINT platform that we have developed provides a powerful and flexible tool for investigating and analyzing large amounts of data. By leveraging the power of Recon-ng and the flexibility of the GPT model, we are able to provide users with an intuitive and effective platform for processing and analyzing text data. With its scalable architecture and advanced features, the platform represents a valuable tool for investigators looking to extract insights and identify patterns in large amounts of text data in OSINT investigations.

VII. Conclusion

In conclusion, the GPT model is a superior natural language processing tool for OSINT investigations due to its generative architecture. Unlike other models that are designed for specific tasks like classification or question-answering, GPT is a generative model that is capable of generating human-like text. This makes it a powerful tool for text generation, language translation, and other applications in OSINT investigations.

The generative nature of GPT enables investigators to process and analyze large amounts of text data, identify patterns and relationships, and generate coherent and convincing text. This can help investigators to extract insights and identify important information from open sources, reducing the amount of time and effort required for data analysis. Additionally, GPT's pretraining, fine-tuning, and unsupervised learning capabilities enable investigators to train the model on specific domains and topics, further enhancing its accuracy and effectiveness for OSINT investigations.

Overall, the generative architecture of GPT is a critical advantage for OSINT investigations, allowing investigators to process large amounts of text data, identify patterns and relationships, and generate coherent and convincing text. By leveraging the power of GPT, investigators can improve the

accuracy and reliability of their analysis, enabling them to generate more comprehensive and insightful reports in a shorter amount of time. As such, GPT represents a valuable tool for investigators looking to extract insights and identify patterns in large amounts of text data in OSINT investigations.

REFERENCES

[1] Lewis Tunstall, Leandro von Werra, Thomas Wolf, "Natural Language Processing with Transformers, Revised Edition" published by O'Reilly Media.

[2] Sudharsan Ravichandiran, "Getting Started with Google BERT" published by Packt Publishing.

[3] Denis Rothman, Antonio Gulli, "Transformers fro Natural Language Processing – Second Edition" published by Packt Publishing.

[4] Shashank Mohan Jain, "Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve the Problem" published by Apress.

[5] Sinan Ozdemir, "Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks" published by Addison-Wesley Professional

[6] Sandra Kublik, Shubham Saboo, "GPT-3" published by Packt Publishing

[7] Nihad A. Hassan, Rami Hijazi, "Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence" published Apres

[8] Joe Gray, "Practical Social Engineering" published No Starch Press