EasyChair Preprint

№ 13058

Multi-Modal Fusion for Anomaly Detection in Cybersecurity: Integrating NLP with Network Traffic Data and System Logs

Dylan Stilinki and Kaledio Potter

April 20, 2024

Multi-Modal Fusion for Anomaly Detection in Cybersecurity: Integrating NLP with Network Traffic Data and System Logs

Date: 16th April 2024

Authors:

Dylan Stilinski

Department of Computer Science

University of Northern Iowa

Kaledio Potter

Department of Mechanical Engineering

Ladoke Akintola University of Technology

Abstract

In the realm of cybersecurity, the detection of anomalies and intrusions remains a paramount challenge due to the evolving nature of cyber threats. Traditional anomaly detection methods often rely on individual data sources, such as network traffic data or system logs, which may provide limited insights when analyzed in isolation. To address this limitation, this paper proposes a novel approach that leverages multi-modal fusion, specifically integrating Natural Language Processing (NLP) techniques with other modalities like network traffic data and system logs, to enhance anomaly detection capabilities.

The integration of NLP with other modalities offers a holistic view of cybersecurity data, enabling a deeper understanding of potential threats and anomalies. By analyzing textual descriptions within system logs or network traffic metadata, NLP techniques can extract valuable contextual information, such as the intent behind certain activities or the presence of suspicious patterns. This textual information, when combined with quantitative data from network traffic or system logs, allows for a more comprehensive analysis of cybersecurity events.

Furthermore, the fusion of multiple modalities enables the detection of anomalies that may not be apparent when analyzing each data source independently. For example,

anomalies detected in network traffic data may be corroborated or further explained by textual information extracted from system logs, leading to more accurate threat identification and reduced false positives.

This paper discusses various methods for integrating NLP with network traffic data and system logs, including feature-level fusion, decision-level fusion, and model-level fusion. Additionally, it examines challenges such as data heterogeneity, scalability, and model interpretability, and proposes solutions to address these issues.

Overall, the proposed multi-modal fusion approach presents a promising avenue for advancing anomaly detection in cybersecurity. By harnessing the complementary strengths of NLP and other modalities, cybersecurity practitioners can enhance their capabilities to detect and mitigate emerging threats in today's dynamic cyber landscape.

Keywords: Multi-Modal Fusion, Anomaly Detection, Cybersecurity, Natural Language Processing (NLP), Network Traffic Data, System Logs, Threat Detection, Contextual Information, Fusion Techniques, Cyber Threats

I. Introduction

A. Motivation: The Rise of Unstructured Data in Cybersecurity

In recent years, there has been a significant increase in the amount of unstructured data generated within the field of cybersecurity. Unstructured data refers to data that does not adhere to a predefined data model or schema, making it difficult to organize and analyze using traditional methods. This includes various types of data such as system logs, user activity records, emails, social media posts, and more.

Traditionally, cybersecurity has relied on structured data sources, such as network traffic logs and security alerts, to detect anomalies and potential threats. However, these methods have certain limitations when it comes to analyzing unstructured data. Traditional anomaly detection techniques may struggle to extract meaningful insights from unstructured data due to its lack of uniformity and structure. As a result, there is a growing need for innovative approaches to effectively analyze and detect anomalies in unstructured data within the cybersecurity domain.

B. Multi-Modal Fusion: A Promising Approach

One promising approach to address the challenges posed by unstructured data in cybersecurity is multi-modal fusion. Multi-modal fusion involves combining different data modalities, such as text, images, audio, and video, to enhance the anomaly detection process. By leveraging multiple data sources, each providing unique information, multi-modal fusion has the potential to improve the accuracy and coverage of anomaly detection systems.

The fusion of different data modalities enables a more comprehensive understanding of the underlying patterns and relationships within the data. For example, in the context of cybersecurity, combining system logs with user activity records and email content can provide a more holistic view of potential threats and anomalies. By considering multiple modalities, the anomaly detection system can capture a broader range of anomalous behaviors and identify potential threats that would be missed by analyzing each modality independently.

The benefits of multi-modal fusion for anomaly detection in cybersecurity are manifold. Firstly, it can lead to improved accuracy in identifying anomalous patterns. By integrating complementary information from different data modalities, the system can reduce false positives and false negatives, resulting in more reliable anomaly detection. Secondly, multi-modal fusion allows for broader threat coverage. Unstructured data sources, such as email content or social media posts, can offer valuable insights into potential cyber threats that may not be evident from structured data alone. By incorporating these diverse data sources, anomaly detection systems can detect a wider range of threats and provide a more comprehensive defense against cyber attacks.

In summary, the increasing volume and importance of unstructured data in cybersecurity necessitate new approaches to anomaly detection. Multi-modal fusion offers a promising solution by combining different data modalities to enhance the accuracy and coverage of anomaly detection systems. By leveraging the unique information provided by each modality, multi-modal fusion enables a more comprehensive analysis of unstructured data, leading to improved threat detection capabilities in the ever-evolving landscape of cybersecurity.

II. Data Preprocessing

A. Network Traffic Data

Network traffic data is an important source of information for anomaly detection in cybersecurity. It can be collected from various devices such as firewalls, intrusion detection systems, or network monitoring tools. However, before this data can be effectively used for anomaly detection, it requires preprocessing.

One aspect of data preprocessing for network traffic data is feature engineering. This involves selecting and extracting relevant features from the raw data. Examples of commonly used features include packet size, protocol type, IP addresses, port numbers, and timestamps. These features provide valuable information about the characteristics of network traffic and can help identify abnormal patterns or behaviors.

Feature engineering for network traffic data may also involve transforming the data into a suitable format. For instance, categorical variables such as protocol type or IP addresses may need to be one-hot encoded or converted into numerical representations.

B. System Logs

System logs are another valuable source of data for anomaly detection. They can include various types of logs, such as application logs, security logs, or operating system logs. However, system logs are typically unstructured and vary in format, making them challenging to analyze directly. Therefore, preprocessing techniques are necessary to ensure consistency and facilitate anomaly detection.

Log parsing is a crucial step in preprocessing system logs. It involves extracting relevant information from the logs and structuring it in a consistent format. Log parsing techniques can vary depending on the log format and contents. Regular expressions or specific log parsing libraries can be used to extract specific fields such as timestamps, log levels, event descriptions, or error codes.

After parsing, the log data may undergo normalization to ensure consistency across different log entries. This normalization process involves transforming and standardizing the log data, such as converting timestamps to a uniform format or normalizing numeric values.

C. Natural Language Processing (NLP) for Textual Data

In anomaly detection, textual data, such as emails, social media posts, or user activity logs, can provide valuable insights into potential cyber threats. However, textual data requires preprocessing before it can be effectively utilized.

Text pre-processing involves several steps to clean and transform the text data. These steps typically include tokenization, stemming, and lemmatization. Tokenization involves breaking down the text into individual words or tokens. Stemming reduces words to their base or root form, removing prefixes or suffixes. Lemmatization goes a step further by reducing words to their canonical form based on their dictionary definition.

Feature extraction techniques are also commonly applied to textual data. Bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) are popular methods for representing text data numerically. BoW represents a document as a vector of word frequencies, while TF-IDF takes into account the importance of a word in a document relative to its frequency in the entire corpus.

In more advanced cases, advanced NLP models can be employed for feature extraction. Word embeddings, such as Word2Vec or GloVe, capture the semantic relationships between words by representing them as dense vectors in a high-dimensional space. Sentiment analysis models can be used to extract sentiment-related features from text, providing additional insights for anomaly detection.

Overall, data preprocessing plays a crucial role in preparing the various types of data, such as network traffic data, system logs, and textual data, for anomaly detection. By applying techniques such as feature engineering, log parsing, normalization, and NLP preprocessing, the data can be transformed into a suitable format for subsequent analysis and anomaly detection algorithms.

III. Multi-Modal Fusion Techniques

A. Early Fusion

Early fusion is a multi-modal fusion technique where the features from different modalities are concatenated together before being fed into the model for training. This approach combines the information from multiple sources at the input level. For example, in the context of anomaly detection, features extracted from network traffic data, system logs, and textual data can be concatenated into a single feature vector.

Early fusion is suitable when the combined feature space from different modalities is manageable and can be handled by the model. It works well with models that can effectively handle high-dimensional data, such as deep neural networks. By providing the model with a unified representation of the different modalities, early fusion aims to capture the interactions and relationships between them directly.

B. Late Fusion

Late fusion, on the other hand, involves training separate models for each data modality and then combining their outputs to make a final decision. Each modality is processed independently, and the models can have different architectures tailored to the characteristics of each modality. The outputs of the individual models can be combined using various fusion techniques, such as averaging, voting, or weighted fusion.

Late fusion offers flexibility in using different models for each modality, allowing for customization based on the specific requirements and complexities of the data sources. It enables the use of specialized models for each modality to extract the most relevant features and capture the unique characteristics of each data type. The combination of the outputs from separate models provides a comprehensive understanding of the anomalies present in the multi-modal data.

C. Deep Learning Approaches

Deep learning techniques have shown promise in multi-modal fusion for anomaly detection. These approaches leverage neural networks with multiple layers to learn complex representations and relationships between different data sources.

Autoencoders are commonly used for dimensionality reduction and anomaly scoring in multi-modal fusion. An autoencoder is a type of neural network that learns to reconstruct its input data. By training an autoencoder on the combined multi-modal data, the model learns to encode the essential features and compress the information. The reconstruction error between the input and the reconstructed output can be used as a measure of anomaly, indicating deviations from the expected patterns.

Recurrent Neural Networks (RNNs) are effective in capturing temporal dependencies present in sequential data, such as network traffic and system logs. RNNs, with their ability to model sequential information, can capture the time-dependent patterns and anomalies in the data. By applying RNNs to each modality separately and then combining their outputs, the model can effectively capture the temporal relationships between the different modalities.

Transformers, a type of deep learning model architecture, have gained popularity in various natural language processing tasks and have also been applied to multi-modal fusion. Transformers excel at capturing complex relationships between different data sources by leveraging self-attention mechanisms. They can learn the interactions and dependencies between various modalities, allowing for a more comprehensive understanding of the multi-modal data and improved anomaly detection.

IV. Anomaly Detection Model Training and Evaluation

A. Supervised vs. Unsupervised Learning Approaches

Anomaly detection in cybersecurity can utilize both supervised and unsupervised learning approaches.

Supervised learning requires labeled data, where each instance is labeled as either normal or anomalous. However, acquiring labeled data in cybersecurity can be challenging due to the scarcity of labeled anomalies. Anomalies are often rare events, making it difficult to obtain a sufficient number of labeled anomalies for training a supervised model. Additionally, the evolving nature of cyber threats makes it challenging to keep labeled datasets up to date.

In contrast, unsupervised learning approaches do not require labeled data. They aim to learn the inherent patterns and structures within the data and identify deviations from those patterns as anomalies. Unsupervised learning models can be trained on normal data, and during inference, they flag instances that significantly deviate from the learned patterns. Unsupervised methods are particularly suitable for detecting novel or previously unseen anomalies.

In cybersecurity, where labeled data is scarce, semi-supervised learning approaches can be employed. These methods utilize a small amount of labeled data combined with a larger amount of unlabeled data during training. By leveraging the labeled data, the model can learn to identify known anomalies, while also benefiting from the unlabeled data to capture the underlying normal patterns. This approach can provide a more robust and comprehensive anomaly detection system.

B. Evaluation Metrics

When evaluating anomaly detection models, several metrics can be used to assess their performance:

Precision: Precision measures the proportion of correctly identified anomalies among all instances flagged as anomalies. It represents the accuracy of the model in identifying true anomalies, and a higher precision indicates fewer false positives.

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of correctly identified anomalies among all actual anomalies in the dataset. It represents the model's ability to capture true anomalies and avoid false negatives.

F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. It considers both precision and recall, and a higher F1-score indicates a better trade-off between the two metrics.

In addition to precision, recall, and F1-score, other evaluation metrics specific to anomaly detection in cybersecurity can include:

Time to detection: This metric measures the time taken by the model to detect anomalies. It is particularly important in cybersecurity, where quick identification of anomalies is crucial for timely response and mitigation.

False positive rate: The false positive rate represents the proportion of normal instances incorrectly flagged as anomalies. It measures the model's ability to avoid false alarms and reduce unnecessary investigation of normal instances.

The choice of evaluation metrics depends on the specific requirements and priorities of the anomaly detection system. A balance needs to be struck between detecting as many anomalies as possible (high recall) while minimizing false positives (high precision). The time to detection is also a critical factor, as timely identification of anomalies can significantly impact the effectiveness of cybersecurity measures.

V. System Design and Implementation

A. Real-time Anomaly Detection Pipeline

A real-time anomaly detection pipeline involves several components working together to process incoming data, detect anomalies, and generate alerts. The pipeline typically includes the following modules:

1. Data ingestion and pre-processing modules: This module is responsible for collecting data from various sources, such as network traffic, system logs, or textual data. It may involve integrating with data collection systems, APIs, or streaming platforms. The data is then pre-processed to transform it into a suitable format for anomaly detection. This may include feature engineering, log parsing, normalization, or NLP pre-processing techniques discussed earlier.

2. Multi-modal fusion and anomaly detection model: Once the data is pre-processed, the multi-modal fusion techniques can be applied to combine information from different data sources. The fused data is then fed into the anomaly detection model, which can be based on supervised, unsupervised, or semi-supervised learning approaches. This model analyzes the data and identifies deviations from normal patterns, flagging them as anomalies.

3. Alert generation and notification system: When an anomaly is detected, an alert is generated to notify the appropriate stakeholders. This module can include rules-based systems, thresholds, or machine learning models to determine when an anomaly is significant enough to trigger an alert. The alerts can be sent through various channels such as emails, SMS, or integration with incident management systems.

B. Scalability and Performance Considerations

When designing an anomaly detection system, scalability and performance considerations are crucial to ensure its effectiveness in real-time environments. Here are some important considerations:

1. Data volume and velocity: The system should be able to handle large volumes of data in real-time. This may involve implementing distributed computing frameworks or technologies that allow parallel processing and efficient data storage and retrieval. Stream processing frameworks like Apache Kafka or Apache Flink can be used to handle high data velocities and ensure timely processing.

2. Infrastructure and resource management: Scalability requires appropriate infrastructure and resource allocation. The system should be designed to scale horizontally by adding more computing resources as the data volume or processing demands increase. This can involve deploying the system on cloud platforms or using containerization technologies like Docker and orchestration tools like Kubernetes.

3. Model optimization and inference efficiency: Anomaly detection models should be optimized for efficient inference and low-latency processing. This can include model compression techniques, such as quantization or pruning, to reduce model size and improve inference speed. GPU acceleration or specialized hardware can also be utilized to speed up the model inference process.

4. Monitoring and performance evaluation: Continuous monitoring of the system's performance is essential to identify bottlenecks, optimize resource allocation, and ensure efficient anomaly detection. Monitoring tools and performance metrics can help track system health, data processing rates, model performance, and latency. This information can be used for system optimization and capacity planning.

5. Incremental model updates: Anomaly detection models should be adaptable to evolving threats and new patterns. The system should support incremental model updates to incorporate new labeled anomalies or adapt to changes in the data distribution. Online learning techniques, such as online gradient descent or concept drift detect
6. ion algorithms, can facilitate continuous learning and adaptation of the model.

By considering scalability and performance factors during the design and implementation of the anomaly detection system, it can effectively handle the real-time processing demands, accommodate growing data volumes, and provide timely detection and alerting capabilities in cybersecurity environments.

VI. Challenges and Future Directions

A. Explainability and Interpretability of Multi-Modal Models

As multi-modal models become more complex and capable of handling diverse data sources, the challenge of explainability and interpretability arises. Understanding why a model makes certain predictions or detects anomalies is crucial, especially in cybersecurity, where the decision-making process needs to be transparent and accountable.

Future research should focus on developing techniques to explain and interpret the decisions made by multi-modal models. This can involve visualizations, attention mechanisms, or feature importance analysis to provide insights into which modalities or features contribute most to the model's predictions. Explainable AI methods, such as rule extraction or surrogate models, can also be employed to make the decision-making process more transparent and understandable to human analysts.

B. Continuous Learning for Evolving Threat Landscape

The cybersecurity landscape is constantly evolving, with new threats emerging and existing ones evolving over time. Anomaly detection systems need to adapt to these changes to effectively detect novel anomalies and maintain high detection rates.

Continuous learning techniques can be employed to enable anomaly detection models to learn and adapt in real-time. This involves updating the model with new labeled anomalies or adjusting its parameters to account for changes in the data distribution. Incremental learning algorithms, active learning strategies, or ensemble methods can be explored to facilitate continuous learning and improve the model's performance over time.

C. Privacy-Preserving Techniques for NLP in Cybersecurity

Natural Language Processing (NLP) techniques are widely used in cybersecurity for analyzing textual data, such as system logs, threat intelligence reports, or user communications. However, privacy concerns arise when dealing with sensitive or personal information in the cybersecurity domain.

Future research should focus on developing privacy-preserving techniques for NLP in cybersecurity. This includes methods to anonymize or encrypt sensitive information in the data, ensuring that personally identifiable information or confidential details are protected. Privacy-enhancing technologies, such as secure computation, federated learning, or differential privacy, can be explored to enable secure analysis of NLP data while preserving privacy.

Additionally, research should address the trade-off between privacy and model performance. Techniques that strike a balance between preserving privacy and maintaining the effectiveness of anomaly detection models need to be developed. This

involves investigating methods to extract relevant information from NLP data while minimizing the risk of privacy breaches.

Overall, addressing challenges in explainability, continuous learning, and privacy preservation will pave the way for more robust and effective anomaly detection systems in cybersecurity. These advancements will contribute to enhanced threat detection capabilities, better decision-making support for analysts, and improved protection of sensitive information.

VII. Conclusion and Future Directions

In conclusion, the integration of edge computing in low-latency V2V broadcasting systems offers several advantages for urban environments:

1. Reduced Latency: Edge computing brings computing resources closer to the vehicles, minimizing the latency in content retrieval and delivery. By caching popular content at edge nodes, vehicles can access content with lower delay, enhancing the overall user experience.

2. Improved Network Efficiency: Caching content at edge nodes reduces the reliance on backhaul connections to centralized servers. This reduces network congestion and improves network efficiency, as content can be served from nearby edge caches instead of traversing the entire network.

3. Scalability and Flexibility: Edge computing enables scalable and flexible V2V broadcasting systems. Edge nodes can be dynamically deployed and scaled based on demand, accommodating varying vehicle densities and content popularity patterns. This scalability ensures efficient content delivery in dynamic urban environments.

4. Enhanced Reliability: With edge computing, V2V broadcasting systems become more resilient to network disruptions. Even in scenarios where connectivity to the backhaul network is lost, vehicles can still access locally cached content, ensuring uninterrupted content delivery.

As for future research directions, several areas hold promise for further advancements in low-latency V2V broadcasting:

1. Integration with Intelligent Transportation Systems (ITS): Integrating V2V broadcasting systems with ITS can enhance traffic management, safety, and efficiency. Future research can explore the synergies between V2V broadcasting and ITS, leveraging real-time traffic data, road infrastructure information, and intelligent decision-making algorithms to optimize content dissemination and enable intelligent transportation services.

2. Machine Learning for Content Prediction: Machine learning techniques can be employed to predict content popularity and user preferences in V2V broadcasting

systems. By analyzing historical data, contextual factors, and user behavior, machine learning models can anticipate the content that will be in demand, allowing for proactive caching and efficient content delivery. Future research can focus on developing accurate and adaptive machine learning models for content prediction in dynamic urban environments.

3. Security and Privacy Enhancements: Further research is needed to strengthen security and privacy mechanisms in edge-enabled V2V broadcasting systems. This includes exploring advanced encryption techniques, privacy-preserving algorithms, and anomaly detection methods to protect against emerging threats and ensure user privacy in V2V communication.

4. Energy Efficiency: Energy efficiency is a critical aspect of V2V broadcasting systems. Future research can investigate energy-efficient caching strategies, dynamic resource allocation, and power management techniques to optimize energy consumption in edge nodes and vehicles, extending the operational lifespan of battery-powered devices.

By addressing these research directions, we can unlock the full potential of edge computing integration in low-latency V2V broadcasting systems, enabling efficient and intelligent content delivery in urban environments while ensuring security, privacy, and sustainability.

VII. Conclusion

A. Summary of the Benefits of Multi-Modal Fusion Approach

The multi-modal fusion approach in anomaly detection brings several benefits to cybersecurity:

1. Enhanced detection capabilities: By combining information from multiple data sources, multi-modal fusion enables a more comprehensive understanding of system behavior and potential anomalies. The fusion of diverse modalities, such as network traffic, system logs, and textual data, provides a holistic view of the environment, improving the accuracy and coverage of anomaly detection.

2. Increased resilience to evasion techniques: Anomaly detection models that rely on a single modality may be vulnerable to evasion attempts that manipulate a specific data source. Multi-modal fusion mitigates this risk by considering multiple modalities simultaneously, making it more challenging for attackers to bypass the detection system.

3. Improved contextual understanding: Different data sources provide different perspectives and context, which can help in accurately distinguishing between normal and anomalous behavior. Multi-modal fusion enables the extraction of

meaningful relationships and dependencies between modalities, leading to a better understanding of complex cyber threats and their interconnections.

4. Adaptability to diverse data types: Cybersecurity data comes in various formats, including numerical, categorical, and textual. Multi-modal fusion techniques can handle this diversity by accommodating different data types and effectively integrating them for anomaly detection. This flexibility allows for the inclusion of a wide range of data sources, making the detection system more robust and adaptable.

B. Potential Impact on Improving Cybersecurity Posture

The adoption of multi-modal fusion approaches in anomaly detection can have a significant impact on improving the cybersecurity posture:

1. Early detection and timely response: By leveraging multiple data sources and capturing diverse patterns, multi-modal fusion enables the early detection of anomalies, reducing the time to detection. This facilitates timely response and mitigation measures, minimizing the potential impact of cyber threats.

2. Enhanced threat intelligence: The fusion of different modalities provides a richer source of information for threat intelligence. By analyzing multiple data sources simultaneously, anomaly detection models can identify complex attack patterns, uncover hidden relationships, and generate actionable insights for proactive defense strategies.

3. Reduction of false positives: Multi-modal fusion techniques can help reduce false positives by considering multiple perspectives and cross-validating anomalies across modalities. This leads to more accurate anomaly detection and reduces the burden on cybersecurity analysts, allowing them to focus on genuine threats.

4. Adaptability to evolving threats: The ability of multi-modal fusion models to continuously learn and adapt to evolving threats is crucial in modern cybersecurity. By incorporating new labeled anomalies or adjusting model parameters, the detection system can stay up to date with emerging attack vectors and maintain high detection rates.

In conclusion, the integration of multi-modal fusion approaches in anomaly detection has the potential to greatly enhance cybersecurity capabilities. By leveraging diverse data sources, improving detection accuracy, and enabling timely response, these techniques contribute to a stronger cybersecurity posture, mitigating risks and protecting critical systems and data

References:

1. Arjunan, Tamilselvan. "Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data Using Natural Language Processing." *International Journal for Research in Applied Science and Engineering Technology* 12, no. 2 (February 29, 2024): 1023–29. https://doi.org/10.22214/ijraset.2024.58497.
2. Nursiyono, Joko Ade, and Rasya Khalil Gibran. "Natural Language Processing for Unstructured Data: Earthquakes Spatial Analysis in Indonesia Using Platform Social Media Twitter." *Innovation in Research of Informatics (INNOVATICS)* 5, no. 1 (March 30, 2023). https://doi.org/10.37058/innovatics.v5i1.6678.
3. Parker, R. David, Marissa Abram, and Karen Mancini. "Using Natural Language Processing to Understand Unstructured Healthcare Data." *SSRN Electronic Journal*, 2022. https://doi.org/10.2139/ssrn.4092364.
4. Gupta, Som, and S K Gupta. "Natural Language Processing in Mining Unstructured Data from Software Repositories: A Review." *Sādhanā* 44, no. 12 (November 30, 2019). https://doi.org/10.1007/s12046-019-1223-9.
5. Fonferko-Shadrach, Beata, Arron Lacey, Ashley Akbari, Simon Thompson, David Ford, Ronan Lyons, Mark Rees, and Owen Pickrell. "Using Natural Language Processing to Extract Structured Epilepsy Data from Unstructured Clinic Letters." *International Journal of Population Data Science* 3, no. 4 (August 28, 2018). https://doi.org/10.23889/ijpds.v3i4.699.
6. Sezgin, Emre, Syed-Amad Hussain, Steve Rust, and Yungui Huang. "Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-World Data." *JMIR Formative Research* 7 (March 7, 2023): e43014. https://doi.org/10.2196/43014.
7. Larriva-Novo, Xavier A., Mario Vega-Barbas, Victor A. Villagra, and Mario Sanz Rodrigo. "Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies." *IEEE Access* 8 (2020): 9005–14. https://doi.org/10.1109/access.2019.2963407.
8. Souili, Achille, Denis Cavallucci, and François Rousselot. "Natural Language Processing (NLP) – A Solution for Knowledge Extraction from Patent Unstructured Data." *Procedia Engineering* 131 (2015): 635–43. https://doi.org/10.1016/j.proeng.2015.12.457.

.