Published in the Proceedings of the 2023 Midwest Instruction and Computing Symposium (MICS), April 2023

Automated Categorization of Cybersecurity News Articles through State-of-the-Art Text Transfer Deep Learning Models

Aden Scott, JT Snow, Muhammad Abusaqer

Department of Math and Computer Science

Minot State University

Minot, ND, USA

nathan.a.scott@minotstateu.edu, joshua.snow@minotstateu.edu,
muhammad.abusaqer@minotstateu.edu

Abstract

The rapid increase in cybersecurity events highlights the need for effective detection and organization of news articles that report such incidents. This research investigates deep learning techniques to categorize and organize a large dataset. The study uses a dataset of 3732 cybersecurity news articles classified into four categories: cyberattack, data breach, malware, and vulnerability. The time-consuming and error-prone task of manual categorization can be efficiently and accurately accomplished with deep learning methods. Moreover, the use of deep learning for the categorization and organization of news articles can provide insights into trends and the latest developments in the field.

This study applies the latest deep learning models for categorizing cybersecurity news articles automatically. Specifically, the research evaluates the performance of three state-of-the-art text transformation deep learning models on the cybersecurity news dataset, including BERT, GPT, and RoBERTa. The study reports the results of the categorization and compares the three models.

This paper's primary contribution is applying the latest deep learning models for categorizing and organizing cybersecurity news articles to show performance comparison of the models, highlighting the need for further research into deep learning for text classification in the cybersecurity domain. The results of this study could also help develop tools to assist cybersecurity professionals in keeping up to date with the latest developments in the field.

1 Introduction

The safety of data is a major concern for all internet users. As technology advances, more aspects of life move online, creating opportunities for cybercriminals to exploit. Consequently, the risks of cyberattacks and data breaches are increasing. Cybersecurity professionals face the challenge of effectively detecting and organizing news articles related to cybersecurity events. With thousands of news articles published daily, categorizing them by topic and relevance is time-consuming and error-prone. With all the news on cybersecurity emerging every day, it is difficult to always stay up to date on the latest trends because there is so much information to process.

To address this issue, the authors explore the use of deep learning techniques, specifically focusing on how text transfer deep learning models can be trained to automatically categorize and organize a large dataset of cybersecurity news articles. The goal is to improve the accuracy and efficiency of categorization and provide better insights into the latest developments in the field.

The authors use a dataset consisting of 3,732 cybersecurity news articles, pre-classified into four categories: cyberattack, data breach, malware, and vulnerability. Text preprocessing was applied to the dataset to prepare it for experimentation. The study's methodology encompasses experimentation on the dataset using text transformation deep learning models (BERT, GPT, and RoBERTa) for article classification in automatically categorizing these articles. The authors assess the performance of these models using accuracy, precision, recall, and F1 score as evaluation metrics

Previous research has employed different deep learning models, such as Convolutional Neural Networks (CNNs), Logistic Regression, and Long Short-Term Memory (LSTM) networks, for the automatic categorization of cybersecurity news articles.

Given the increasing importance of cybersecurity in today's world and the growing volume of news articles on cybersecurity events, there is a need for efficient and effective methods for organizing and categorizing these articles. Deep learning in text classification has proven to be effective, which is why authors want to implement it to automatically categorize cybersecurity news articles. The authors hope that this research contributes to the development of accurate and efficient methods for organizing and analyzing cybersecurity news and help cybersecurity professionals stay up to date on the latest threats and developments in the field.

2 Literature Review

Many studies have proven that deep learning is a powerful tool for text classification but is usually applied in a broader domain. Rather than specifically cybersecurity news, many are categories of all news.

A study by (Zhang, 2021) explored the use of deep learning models for classifying news articles into distinct categories, such as event news and ordinary news, addressing challenges related to lengthy text data length and feature extraction difficulties. Traditional approaches, which relied on single-word vectors, considered only the relationship between words while neglecting the crucial relationship between words and categories. Zhang developed a customized DCLSTM-MLP model, combining deep learning algorithms like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP). This model processes word vector and word dispersion information simultaneously to capture the relationship between words and categories. The study achieved an accuracy of at least 90% using models like CNN, text-LSTM, CNN-MLP, CLSTM, and DCLSTM-MHP. Building upon Zhang's work, the authors of the present study focus on classifying cybersecurity news articles using deep learning text transformers. These transformers have demonstrated remarkable success across various natural language processing tasks, offering enhanced classification performance and adaptability to the rapidly evolving cybersecurity landscape.

Arora, et al. (2022) conducted a study on the performance comparison of different machine learning algorithms for Hindi news classification. Text classification is the process of categorizing text into predefined classes. Before applying machine learning algorithms to the extracted text, the authors implemented preprocessing and feature engineering techniques such as count vectorizer, TF-IDF, and word2vec. Preprocessing in this context is challenging due to the presence of multisensory words, conjunctions, punctuations, and special characters in the Hindi language. The model they developed accepts Hindi news headlines from predefined categories like Entertainment, Sports, Tech, and Lifestyle news. After preprocessing, the corpus size containing unique words was 54,44,997. Out of the different combinations tested, the multinomial Naïve Bayes algorithm with count vectorizer achieved the highest accuracy of 85.47%. This study demonstrates the potential of using machine learning algorithms for news classification in languages other than English, which can be informative for research on the classification of cybersecurity news articles using deep learning text transformers.

Another study by Deepak Singh (2021) used deep learning models to classify articles into five different categories: sports, business, politics, entertainment, and tech. This dataset contained 1490 entries. The top four models used were: Logistic Regression, Random Forest, Multinomial Naive Bayes, and Support Vector Classifier. All of which had an average accuracy of 97% as well as an average F1 score of 97%. The results showed that deep learning models could achieve high accuracy and F1 score while categorizing a smaller dataset.

González-Carvajal and Garrido-Merchán (2021) conducted a study comparing the performance of BERT, a state-of-the-art machine learning model, against traditional machine learning text classification approaches using TF-IDF vocabulary. BERT has gained popularity in recent years due to its ability to handle a wide range of NLP tasks, including supervised text classification, without human supervision. The authors aimed to provide empirical evidence supporting or refuting the use of BERT as a default method in NLP tasks. Their experiments demonstrated the superiority of BERT over traditional methods and its independence from features such as the language of the text, adding empirical evidence supporting the use of BERT as a default technique for NLP problems.

This finding highlights the potential of using advanced models like BERT for the classification of cybersecurity news articles using deep learning text transformers.

There are limitations in the existing research, like the lack of standardization in datasets and evaluation metrics. Most studies use different datasets and evaluation metrics, making it difficult to compare the results. Another limitation is the lack of transparency in deep learning models. Deep learning models are often seen as “black boxes,” and it is difficult to understand how they get their classifications.

Overall, the existing research proves the potential of text classification and organization deep learning can have in the cybersecurity domain. With that said, more research is needed to standardize datasets and evaluation metrics.

3 Methodology

In the research, the authors used The Hacker News Dataset from Mendeley Data (Ahmed et al., 2021). We planned to use this dataset to train and evaluate the three deep learning models on how well they classified the evaluation set.

3.1 Dataset and Preparation

The Hacker News Dataset that was used in this study consists of 3,732 cybersecurity news articles collected from thehackernews.com website. The dataset was preprocessed to remove irrelevant information. The articles are pre-classified into four categories: cyberattack, data breach, malware, and vulnerability. These categories were not completely balanced. The number of labels of Cyberattacks and Data breaches is less than Malware and Vulnerability. Having this unbalanced data set could lead to biased models, but authors will measure this using specific metrics. This is illustrated in Table 1 below.

Category	Number of Articles
Cyberattack	364
Data breach	699
Malware	1327
Vulnerability	1352

Table 1: The number of articles across the four categories in the dataset.

3.2 Dataset Preprocessing

Prior to inputting the text into the deep learning models, the authors carried out several preprocessing steps. They first removed stop words and special characters from the text and converted them to lowercase. Then tokenized the text.

3.3 Model Selection and Training

In terms of model selection and training, the authors evaluated three state-of-the-art text transfer deep learning models: BERT, RoBERTa, and GPT. They aimed to obtain a diverse set of results from these three popular deep learning algorithms for evaluation purposes.

3.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model introduced by Devlin et al. (2019) that revolutionized NLP. Its bidirectional context representation allows it to learn both the left and right context of a word, leading to significant performance improvements in various NLP tasks (Devlin et al., 2019).

3.3.2 RoBERTa

RoBERTa (Robustly Optimized BERT) is an adaptation of BERT by (2019) that refines BERT’s pre-training methodology and data processing. RoBERTa’s modifications result in improved performance on downstream tasks (Liu et al., 2019).

3.3.3 GPT

GPT (Generative Pre-trained Transformer) is a transformer-based model developed by OpenAI that focuses on learning the left context of a word. GPT is pre-trained on a large-scale unsupervised language modeling task and fine-tuned for specific NLP tasks (Radford et al., n.d.).

3.4 Software and Hardware

The authors implemented the models using the PyTorch deep learning framework and ran the experiments on a MacBook Pro. (2.8 GHz Quad-Core Intel Core i7). We used Python 3.10 and several Python libraries, such as Pandas, NumPy, and Scikit-learn, for data preprocessing and analysis.

3.5 Model Evaluation

The performance of each model on the testing set was evaluated using the accuracy, precision, recall, and F1-score metrics. We believed that the combinations of these metrics would give us a good evaluation of the performance of the models. Finally, the authors compared the performance of the models and discussed the results.

3.5.1 Accuracy

Accuracy measures the percentage of correctly predicted instances out of all instances. The mathematical representation for this calculation is accuracy = (true positives + true negatives) / (true positives + false Positives + true negatives + false negatives). This is the simplest metric and could be misleading with an imbalanced dataset (Kelleher et al., 2020).

3.5.2 Precision

Precision measures the percentage of correctly predicted positive instances out of all instances that were predicted as positive. The mathematical representation for this calculation is precision = true positives / (true positives + false positives) (Kelleher et al., 2020) (Sokolova & Lapalme, 2009).

3.5.3 Recall

Recall measures the percentage of correctly predicted positive instances out of all actual positive instances. The mathematical representation for this calculation is recall = 2 * (precision * recall) / (precision + recall) (Sokolova & Lapalme, 2009).

3.5.4 F-1 Score

The F-1 score combines precision and recall scores into a single number. The mathematical representation for this calculation is F-1 Score = true positives / (true positives + false negatives). The F-1 score is a good metric to use when the classes are imbalanced since it is the mean of precision and recall (Chicco & Jurman, 2020) (Powers, 2020).

4 Results

In their research, the authors evaluated the results of three deep learning models: BERT, RoBERTa, and GPT. The performance of each model was measured based on accuracy, precision, recall, and F1-score. Upon completion of the training for all three models, the authors reported the best results for each, as shown in Figure 1.

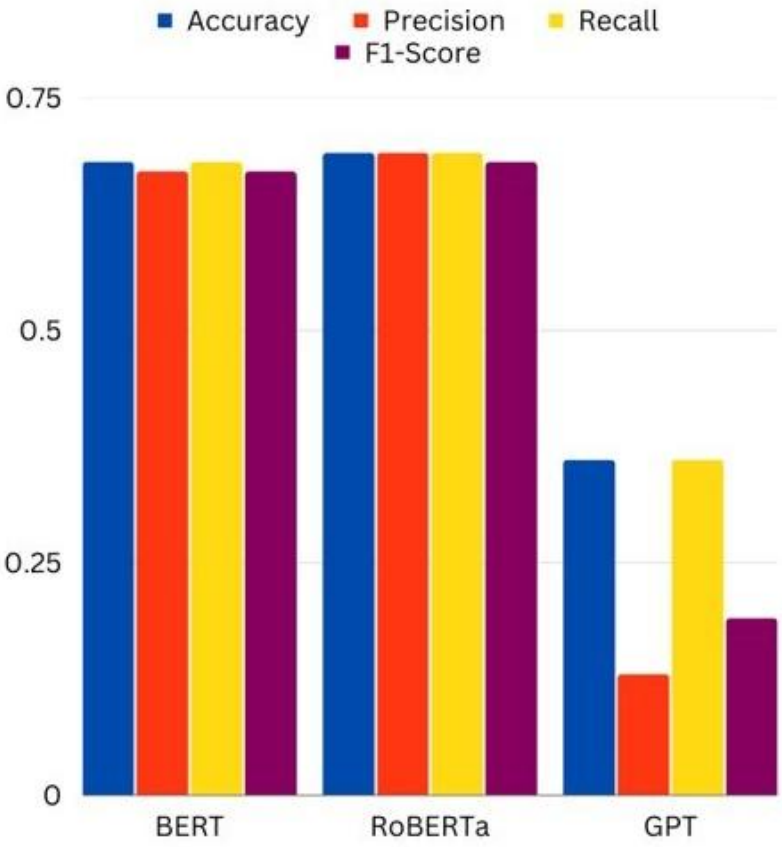

Figure 1: Evaluation results for deep learning models.

5 Discussion

The authors' experiment aimed to compare the performance of three different deep learning models when classifying articles from the Hacker News dataset.

The results point to RoBERTa being the best model for this specific task. RoBERTa performed slightly better than BERT with an accuracy of 0.69, precision of 0.69, recall of 0.69, and F1-Score of 0.68. *These results show the potential of using these large pre-trained language models to classify news articles in the cyber security domain*.

In contrast, the GPT model did not perform as well as the other two, with an accuracy of 0.36, precision of 0.13, recall of 0.36, and F1-Score of 0.19. *The results suggest that GPT struggled with this specific task and is better suited for other natural language tasks like text prediction or generating text*.

6 Future work

Future research in this area, as suggested by the authors, could enhance model accuracy by employing larger, more diverse datasets for training, as well as exploring various pre-trained language models to find the optimal architecture for this specific task. Additional experimentation with these algorithms may facilitate the development of tools for rapidly detecting trends in the cybersecurity domain, thereby enabling companies to bolster security and respond to emerging trends more promptly.

7 Conclusion

In this study, the authors compared the results of three text transformers' deep learning algorithms for classifying articles from the Hacker News dataset. The findings indicate that RoBERTa outperforms the other two models tested for this task. The experiment contributes to the understanding of the strengths and limitations of machine learning models in natural language processing, emphasizing the need for further research into these models, particularly within the cybersecurity domain. The authors also hope their work inspires additional research into predicting trends in the cybersecurity field.

References

Ahmed, M. F., Anwar, M. T., Tanvir, S., Saha, R., Shoumo, S. Z. H., Hossain, M. S., & Rasel, A. A. (2021). *Cybersecurity News Article Dataset*. *1*. https://doi.org/10.17632/n7ntwwrtn5.1

Arora, M., Dhingra, B., Gupta, D., & Singh, D. (2022). Performance Comparison of Different Machine Learning Algorithms on Hindi News Classification. In A. Khanna, D. Gupta, S. Bhattacharyya, A. E. Hassanien, S. Anand, & A. Jaiswal (Eds.), *International Conference on Innovative Computing and Communications* (pp. 323–333). Springer. https://doi.org/10.1007/978-981-16-2597-8_27

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). *Comparing BERT against traditional machine learning text classification* (arXiv:2005.13012). arXiv. https://doi.org/10.48550/arXiv.2005.13012

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pre-training Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation* (arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). *Improving Language Understanding by Generative Pre-Training*.

Singh, D. (2021, December 27). Text Classification of News Articles. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Zhang, M. (2021). Applications of Deep Learning in News Text Classification. *Scientific Programming*, *2021*, e6095354. https://doi.org/10.1155/2021/6095354