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-3, and RoBERTa. The study reports the results of the categorization and makes a comparison between the three models.

Introduction

The safety of data is a major concern to everyone who uses the internet. With the continuous advances in technology more of our life has moved online creating opportunities for cyber criminals to take advantage of. Therefore, the risks for cyber-attacks and data breaches are becoming more significant. One key challenge that cybersecurity professionals are facing is how to effectively detect and organize news articles related to cybersecurity events. With thousands of news articles published daily, categorizing them by topics and relevance is very time consuming and is prone to errors. 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, we are going to learn how to use some deep learning techniques. Specifically, we explore how text transfer deep learning models can be trained to automatically categorize and organize a large dataset of cybersecurity news articles. With the help of deep learning, our goal is to improve the accuracy and efficiency of categorization and provide better insight on the latest developments in the field.

To achieve our goals, we use a dataset of 3,732 cybersecurity news articles, which have been classified into four preset categories: cyberattack, data breach, malware, and vulnerability. With this dataset we will test out three text transformation deep learning models: BERT, GPT-3, and RoBERTa.

Our primary research question is how well do these models perform at automatically categorizing cybersecurity news articles? We will compare the results of the models and see which is the most effective, as well as find the strengths and weaknesses of each model. We hope that our findings will be able to help advance the field of deep learning for text classification and provide practical benefits to help cybersecurity professionals stay up to date on the latest developments in cybersecurity.

The methodology used in this study involves the following steps. First, we collected a dataset of 3,732 cybersecurity news articles from various sources and each was prelabeled into one of four categories: cyberattack, data breach, malware, or vulnerability. Second, we performed preprocessing steps on the dataset, including removing stop word, stemming, and converting all text to lowercase. Third, we used text transformation deep learning models (BERT, GPT-3 and RoBERTa) to classify the articles into their respective categories. We fine-tuned the models on our dataset and evaluated their performance by accuracy, precision, recall, and F1 score. Overall, our methodology aims to provide a robust and rigorous approach to the automatic categorization of cybersecurity news articles using deep learning techniques.

There has been some research done for text classification and organization of news articles. The studies used different deep learning models including Convolutional Neural

Networks (CNNs), Logistic Regression, and Long Short-Term memory (LSTM) networks, for automatic categorization of cyber security news articles.

A limitation that we noticed in a couple of these studies is that they often rely on smaller datasets that do not reflect the diversity and complexity relative to real-world news articles. Another gap in existing research is the lack of analysis of the feature importance and interpretability of deep learning models for text classification of cybersecurity of news articles. Deep learning models have shown they are capable of high accuracy computing, but it is difficult to know how they make their predictions. Knowing what the model considers for classification would help gain insights into the latest developments and trends in the field.

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 we want to implement it to automatically categorize cybersecurity news articles. We hope that with this research we can contribute to the development of accurate/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.

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.

One study by Menghan Zhang. (2021) used deep learning models to classify news articles into different event news and ordinary news categories. This study used a dataset of 8292 event news and ordinary news texts. From that set 6699 texts are randomly selected as the training set and 1593 were selected as the test set. This study

achieved an accuracy of at least 90% over all the models used. Some of these models were: CNN model, text-LSTM model, CNN-MLP model, CLSTM model, and the DCLSTM-MHP model. This study demonstrates the effectiveness deep learning models have for text classification on news articles.

Another study by Deepak Singh. (2021) used deep learning models to classify articles into five different categories: sport, business, politics, entertainment, and tech. This dataset contained 1490 entries. The top four models used were: Logistic Regression, Random Forest, Multinomial Naive Bayes, 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 can achieve high accuracy and F1 score while categorizing a smaller dataset.

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.

Dataset

The dataset that was used in this study consists of 3,732 cybersecurity news articles collected from various online sources between January 2020 and May 2022. The dataset was preprocessed to remove irrelevant information. The text was tokenized and normalized to lowercase, and stop words and punctuation was removed. The resulting dataset consisted of approximately 13 million tokens.

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

Table 1: Table 1 shows articles across the four categories in the dataset. As seen above, the data is unbalanced between the 4 classes.

The dataset is publicly available and can be accessed through the following link: https://data.mendeley.com/datasets/n7ntwwrtn5

Methodology

The code used to conduct this research is available on GitHub at https://github.com/nathanascott/MICS2023/tree/master/projects.

Dataset Preparation

The dataset used in this research consists of 3,732 cyber security news articles collected from various online news sources, such as ZDNet, Dark Reading, and Security Week, over a period of six months. The article was pre classified into four categories: cyberattack, data breach, malware, and vulnerability.

Preprocessing

Before feeding the text into the deep. Learning models, we preformed several preprocessing steps. First, we removed stop words and special characters from the text, Next, we tokenized the text into words and converted them to lowercase.

Model Selection and Training

We evaluated three state-of-the art text transfer deep learning models: BERT, GPT-3 and RoBERTa.

Model Evaluation

We evaluated the performance of each model on the testing set using precision, recall, and F1-score metrics. We also plotted the confusion matrix to visualize the models' predictions and identify the misclassified articles. Finally, we compared the performance of the models and discussed the results.

Software and Hardware:

We implemented the models using the PyTorch deep learning framework and ran the experiments on a MacBook Pro. (2.8GHz 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.

Results

During this experiment, we attempted to train three deep learning models which were BERT, RoBERTa, and GPT. Our goal was to compare the performance of the three models on how well they were able to classify our dataset. Due to our limited experience

with machine learning, we took a cautious approach to our study design and methodology. We acknowledge that our findings may be influenced by our beginner experience and our limited approach. Unfortunately, we were unable to obtain acceptable results for two of the models and obtained no results for our GPT model. Specifically:

1. BERT and RoBERTa models: We trained these two models using the simple transformers library from hugging face. After we were able to measure the performance of the models using accuracy, f1 score, precision, and recall. One interesting observation was that when using a subset of our dataset it would generate near perfect results but unfortunately while training on the complete dataset, both models produced inadequate results. For example, the BERT model only 0.188 for its accuracy. We repeated our experiment many times using different hyperparameters and preprocessing methods but were unable to improve the results for either model.
2. GPT model: We tried to train the GPT model using the same method as the other two models with the difference of using the transformers library by hugging face. The combination of the GPT model's complexity and our lack of experience and knowledge lead to significant challenges in implementing it. Given the importance of GPT in our research question, we acknowledge that our study is limited by the model's absence.

These disappointing results could suggest that the task of classifying the Hacker News dataset was challenging for these models. It could also suggest that our dataset may be insufficiently diverse, or the labels are too closely related making it hard to classify correctly. We acknowledge that the limitations of our experiment affected the results, and we plan to address these issues in future research.

Discussion

Our experiment aimed to compare the performance of three different deep learning models when classifying articles from the Hacker News dataset. While we hoped to obtain adequate results, we found that both BERT and RoBERTa produced subpar results, while the GPT model did not produce any results at all.

The lack of results for the GPT model came from our inexperience with machine learning. This is unfortunate given the model’s recent success in other natural language processing tasks. The subpar results of BERT and RoBERTa are likely due to the fine-tuning process that we implemented, however it is possible that these models may not be the most optimal for this specific task. Another reason could have been the unbalanced classes in the dataset. This could have affected the accuracy of the models.

It is worth noting that our experiment had several limitations that affected the results, however our experiment does highlight the need for careful consideration of the model and training arguments when working with deep learning algorithms. We are confident

that our experiment adds to a growing body of research on the strengths and limitations of deep learning algorithms in natural language processing tasks and points to areas for future research.

Finally, we believe our experiment supplies a useful starting point for future research. Future studies could benefit from using larger and more balanced classes for its dataset and implementing a wider range of model configurations. That said, we want to share our code for other users to build on for further research into the effectiveness of deep learning models in classifying cyber security related news articles.

Conclusion

While conducting this research we gained our first experience with machine learning algorithms. Although it was very challenging at times, this provided us with a valuable hands-on learning experience. Unfortunately, we did not achieve the results we were looking for, but we view this as a great opportunity to familiarize ourselves with this new technology.

In this study we compared the results of three machine learning algorithms when classifying articles from the Hacker News dataset. Our results indicate that BERT and RoBERTa obtained subpar results while the GPT model failed to record any results at all. These findings could suggest that these pre-trained models are not optimal for this task, or that more fine-tuning is required to achieve better results. Our experiment contributes to the research on the strengths and limitations of machine learning models in natural language processing and highlights the importance of careful model and parameter selection.

In conclusion, our study highlights the need for further research on the effectiveness of machine learning models on classifying cyber security related articles. While our results are limited, we believe that our work provides a useful starting point for future research. We hope our code is valuable to other researchers in this field and that our findings will help advance our understanding of natural language processing.

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. *Data.mendeley.com*, *1*. https://doi.org/10.17632/n7ntwwrtn5.1

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). Comparing BERT against traditional machine learning text classification. *ArXiv:2005.13012 [Cs, Stat]*. https://arxiv.org/abs/2005.13012

Medium. (n.d.). Medium. Retrieved March 19, 2023, from https://towardsdatascience.com/deep-learning-techniques-for-text-classification-78d9dc40bf7

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

Zhang, M. (2021). Applications of Deep Learning in News Text Classification. *Scientific Programming*, *2021*, 1–9. https://doi.org/10.1155/2021/6095354