Comparative Analysis of Natural Language Processing Models for Malware Spam Email Identification

Francisco Jáñez-Martino, Eduardo Fidalgo,
Rocío Alaiz-Rodríguez, Andrés Carofilis
Alicia Martínez-Mendoza
Department of Electrical, Systems, and Automation, Universidad de León, León, ES
Researcher at INCIBE (Spanish National Cybersecurity Institute), León, ES
francisco.janez, eduardo.fidalgo, andres.carofilis, alicia.martinez, rocio.alaiz@unileon.es

Abstract

Spam email is one of the main vectors of cyberattacks containing scams and spreading malware. Spam emails can contain malicious and external links and attachments with hidden malicious code. Hence, cybersecurity experts seek to detect this type of email to provide earlier and more detailed warnings for organizations and users. This work is based on a binary classification system (with and without malware) and evaluates models that have achieved high performance in other natural language applications, such as fastText, BERT, RoBERTa, DistilBERT, XLM-RoBERTa, and Large Language Models such as LLaMA and Mistral. Using the Spam Email Malware Detection (SEMD-600) dataset, we compare these models regarding precision, recall, F1 score, accuracy, and runtime. DistilBERT emerges as the most suitable option, achieving a recall of 0.792 and a runtime of 1.612 ms per email.

1 Introduction

Spam email has been a challenge since the creation of email services. Spam is known as a synonym for annoying and unwanted emails, which result in a loss of time and productivity for users. Moreover, spam is currently one of the most common sources for incoming scams (Jáñez-Martino et al., 2023), and also a frequent medium to spread malicious files like ransomware, viruses, and malware. Malicious files can take control of the devices for a harmful and undesirable effect on host machines (Cohen et al., 2018). Criminals often demand financial rewards from individuals or organizations to release the infected devices.

Cybersecurity organizations develop anti-spam filters focusing on fraudulent activities such as phishing or spoofing (Gallo et al., 2021). However, little work has been done to detect those spam emails with highly suspicious indicators that may contain malware, either through external links or attached files (Jáñez-Martino et al., 2023). Filtering these emails may enhance the identification by Computer Security Incident Response Teams (CSIRT), cybersecurity companies like the Spanish National Cybersecurity Institute (INCIBE), or users, as well as alerting and providing insight for further investigation.

Additionally, spammers, users who send spam emails, counteract this type of system through various sophisticated strategies like introducing obfuscated words. Consequently, there is a back-and-forth battle between both parties, which causes a deterioration of datasets and models trained with them over the years (Jáñez-Martino et al., 2023). This adversarial dataset shift leads developers to update the filters with newer data constantly. The lack of public and annotated data hinders the periodic update of anti-spam systems for some trending and malicious scams. Nevertheless, the rise of Natural Language Processing (NLP) models such as Transformers (Vaswani et al., 2017) or Large Language Models (LLMs) (Naveed et al., 2024) allows the specialization of pre-trained models using a smaller number of examples. These models may enhance and accelerate the adaptation of filters to new trends.

In this context, we propose to evaluate a selection of the most used NLP models to detect spam emails with suspicious files from traditional pipelines to the application of Transformers and LLMs. Following the work of Redondo-Gutierrez et al. (2022), we classify spam email using only the textual information, i.e., through a text classification approach, as either with or without malware files. Due to the lack of a publicly available dataset, we leverage the previous dataset built by Redondo-Gutierrez et al. (2022) to obtain the performance results. This small dataset allows us to provide evidence for our hypothesis. Finally, this work can offer an initial recommendation about the most suitable model and its configuration that cyberse-

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 59–63
July 29–30, 2024.

curity companies may use if they would decide to implement this filter.

The rest of the paper is organized as follows. Section 2 reviews the background of malware detection, especially in spam emails. Section 3 explains the Spam Email Malware Detection (SEMD-600) dataset and the seven classifiers to be evaluated. Section 4 presents the evaluation and discussion of the classifier performance. Section 5 sums up the contributions of our work and identifies future work.

2 Background

Malware detection has been studied in the literature (Mehta et al., 2024) in recent years using NLP techniques by exploring different learning machines such as Support Vector Machine (SVM) or Long Short-Term Memory (LSTM) following a hybrid approach. Alam (2021) enhanced the techniques to make accessible the potentially malicious code for NLP techniques, in particular semantic similarities. These works aimed at detecting malware in several environments, like Android applications, by directly analyzing the code. However, transferring this methodology to spam email can increase the analysis runtime, as spam emails usually contain multiple potentially malicious resources, URLs, or attachments.

Although some works in the literature focus on detecting malware in files, we only focus on detecting spam emails containing such files. Delving into spam email, Abu Qbeitah and Aldwairi (2018) dynamically analyzed the automatic anomaly detection and active signature generation based on the observed behavior of new malware in phishing emails. Cohen et al. (2018) investigated malware propagation patterns to define features to spot malicious webmail attachments. While Arivudainambi et al. (2019) focused on surveillance against malware by developing a robust traffic classification system, using Principal Component Analysis (PCA) and Artificial Neural Network (ANN). Nevertheless, we aim to leverage quick and secure analysis of the textual information to process the largest possible number of spam emails.

The work of Redondo-Gutierrez et al. (2022) laid the foundation for targeted detection of spam emails with malware content. They sought to analyze the textual information from the email to avoid opening the potential malicious resource throughout a binary text classification. In this way, they proposed a faster and more secure system to detect these emails and a custom and novel dataset available on request (SEMD-600). Despite the novelty of the work, they only carried out the challenge through a traditional approach using Term Frequency - Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) as vectorizers and SVM, Logistic Regression (LR) and Random Forest as classifiers. Thus, exploring trends and current alternatives can improve the performance of the system, considering they achieved their best performance using TF-IDF along with LR.

During the latest years, there has been a rise in the NLP from Word Embedding based models such as Word2Vect and FastText, the attention-based models — Transformers — from BERT and RoBERTa to LLMs like ChatGPT (Palanivinayagam et al., 2023). The attention-based models represent the state-of-the-art in most NLP applications, including text classification. Transformers achieved high overall performance in text classification using pre-trained models as a single pipeline containing all stages of preprocessing, feature extraction, selection, and classification.

3 Methodology

In this paper, we follow a text classification approach to classify spam emails based only on their textual content, focusing specifically on whether they contain malware or not. Redondo-Gutierrez et al. (2022) also adopted this perspective in their work; thereby, we take their work as the baseline to compare their best model with Transformers and LLMs and using their custom and only publicly available dataset in the literature, Spam Email Malware Detection 600 (SEMD-600) [1].

Redondo-Gutierrez et al. (2022) built SEMD-600 using VirusTotal reports to find spam emails with malware. They obtained the resources, i.e., spam emails, from the public repository Spam Archive of Bruce Guenter [2]. Authors randomly selected examples between January 2021 and April 2022, building a dataset comprising 300 spam emails with malware and 300 without malware, written in English only.

Due to the rise of attention-based models and the recent emergence of LLMs, we compare the best model based on traditional techniques (TF-IDF

[1] https://gvis.unileon.es/datasets-semd-600/ retrieved June 2024

[2] http://untroubled.org/spam/ retrieved June 2024

with LR) from (Redondo-Gutierrez et al., 2022) against the most popular current methods in the task of text classification. By doing this, we are providing new baseline results for the task of spam malware detection using only the text of the email. These encompass a Word Embedding solution — FastText —, four early attention models based on BERT architecture (Ameer et al., 2023) — BERT, RoBERTa, DistilBERT, and XLM-RoBERTa on their base version — and two well-known LLMs — LLaMA, built by Meta, and Mistral[3] on their 7B version.

4 Experimentation

4.1 Configuration

We conducted the experiments on a computer with 128 GB of RAM, two Intel Xeon E5-2630v3 processors of 2.4 GHz, and two Nvidia Titan Xp. We used the following Python packages for coding, training and evaluating the models: simpletransformers[4], transformers [5] and fastText [6].

For the BERT, ROBERTa, DistilBERT, XLM-RoBERTa, and LLMs, we chose 512 tokens as the maximum number and 0.00001 for the learning rate while keeping the remaining parameters at their default values. We fine-tuned on text classification each model during 10 epochs using 8 as the training batch. Regarding fastText, we kept the default parameters, training the model for 200 epochs from scratch.

We calculated the precision, recall, F1-Score, accuracy, and runtime in ms per email of each model. Due to the small size of the dataset, we followed a 5-fold cross-validation evaluation.

4.2 Results and discussion

We aim to detect as many spam emails with malware as possible; therefore, we consider recall the most relevant metric for this problem. Table 1 shows the overall results, where RoBERTa and DistilBERT achieved the highest performance with a recall of 0.792. However, DistilBERT also overcame RoBERTa in terms of precision and, consequently, F1-Score, making it a more suitable option for this task.

The model based on TF-IDF and LR of Redondo-Gutierrez et al. (2022) achieved higher recall than BERT and XLM-RoBERTa, the largest model. The complexity of these models and the task may affect negatively due to the spam features, and simpler models like DistilBERT can leverage that. In general, despite the small number of examples, we can say that transformers captured the contextual relationship between words similarly and detected specific patterns of spam language, while FastText stands out as the worst option among those examined. This may be because this model is based on word embeddings and follows a hierarchical classification of the words. These properties may not fully capture the language complexity and features of spam emails.

Finally, the LLMs obtained lower results. It is worth noting that their precision is slightly higher than recall, contrary to the behavior observed in Transformers. The LLMs may capture better those emails with fairly malware features, mistaking in those close to the negative class. This may confirm that the larger models perform lower for this task.

Model	**P**	**R**	**F1**	**Acc**
TF-IDF-LR	0.768	0.763	0.763	76.4
FastText	0.730	0.643	0.681	68.7
BERT	0.733	0.734	0.730	71.9
RoBERTa	0.743	**0.792**	0.766	74.8
DistilBERT	0.774	**0.792**	0.781	77.0
XLM-RoBERTa	0.718	0.780	0.746	72.4
LLaMA	0.620	0.594	0.606	59.8
Mistral	0.653	0.593	0.621	62.4

Table 1: Evaluation of baseline results **(TF-IDF-LR)** from the previous work (Redondo-Gutierrez et al., 2022) for spam malware detection against the one-word embedding model **FastText**), four attention models, and two LLMs in terms of **P**recision, **R**ecall, **F1**-Score and **Acc**uracy.

We also provided a runtime analysis (Fig. 1), as spam email is a big data challenge, and detection speed plays an essential role. The results show that the FastText model and the traditional pipeline (TF-IDF-LR) achieved the fastest runtime, analyzing an email in 0.164 ms and 0.278 ms, respectively. DistilBERT is the fastest attention model with 1.612 ms per email, making it the most recommendable option. The results confirm that both DistilBERT and FastText have a significant advantage in terms of speed.

We avoided including the LLMs runtime in the

[3] `https://mistral.ai/` retrieved June 2024

[4] `https://simpletransformers.ai/` retrieved June 2024

[5] `https://pypi.org/project/transformers/` retrieved June 2024

[6] `https://pypi.org/project/fasttext/` retrieved June 2024

picture due to their longer processing times compared to others. LLaMA and Mistral had a runtime of 200.12 ms and 88.28 ms per email, respectively.

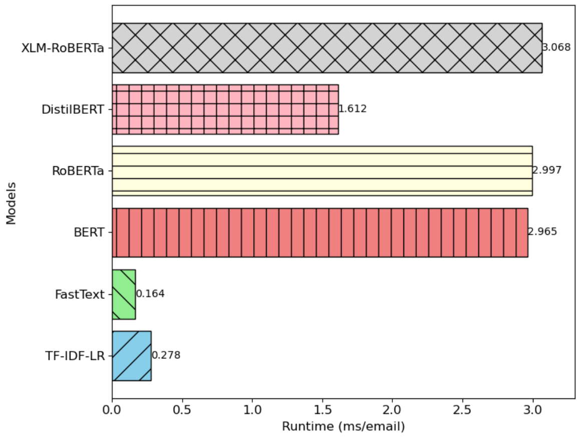

Figure 1: Evaluation of the models in terms of execution times. The results are in milliseconds (ms) per email.

5 Conclusions

In this work, we evaluated a set of one-word embedding and six attention-based models (two out of six are LLMs) against the results obtained by Redondo-Gutierrez et al. (2022) using traditional techniques to detect spam emails containing malware. We followed a binary classification (with or without malware) and trained our model in the SEMD-600 dataset. This small dataset can help determine the effectiveness of using a pre-trained model with few examples.

The results show that DistilBERT achieved the highest recall and was the third fastest model. Although DistilBERT outperformed the previous best model, the overall recall was less than 0.800, indicating a wide range of improvements. The performance gap between state-of-the-art NLP models and more traditional models is not as wide as initially expected, and considering the easy portability of the traditional models, they prove to be a suitable option for cybersecurity organizations.

For future work, it would be interesting to evaluate different sets of parameters in Transformer models and extract features and patterns common in spam emails with malware. In addition, extending the number of examples in both classes (with or without malware) of the SEMD-600 dataset can help to determine if the size of the dataset plays a crucial role in this task.

Limitations

In this work, we have evaluated one traditional classifier, one-word embedding, four Transformers, and two LLMs on the SEMD-600 dataset. The results show a wide range of improvement since any model can surpass 0.800 of recall. We can try to find the most suitable parameter combination per model because we used the same configuration for every model. Moreover, we can conduct a feature analysis to understand patterns of spam emails that can enhance the performance of the models. Due to the spam language, we think a preprocessing stage delves into the obfuscated words and other textual strategies to mislead classifiers. Finally, there was no other dataset and we only tested the models on a small dataset. For future work, we aim to increase the number of examples.

Ethics Statement

This work can contribute to **society and human well-being** and **avoid harm**: by ensuring the safety and security of individuals and organizations who may otherwise fall victim to cyber threats. The **robust system** to detect malware in spam emails can mitigate the negative consequences of being infected, such as data breaches, financial loss, and damage to reputation. Moreover, it provides further **accurate information** about the risks of spam emails that help users **without any discrimination**.

Acknowledgements

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

Mohammad Abu Qbeitah and Monther Aldwairi. 2018. Dynamic malware analysis of phishing emails.

Shahid Alam. 2021. Applying natural language processing for detecting malicious patterns in android applications. *Forensic Science International: Digital Investigation*, 39:301270.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

D Arivudainambi, K A Varun, Sibi Chakkaravarthy, and Pandu Visu. 2019. Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Computer Communications*, 147.

Yehonatan Cohen, Danny Hendler, and Amir Rubin. 2018. Detection of malicious webmail attachments based on propagation patterns. *Knowledge-Based Systems*, 141:67–79.

Luigi Gallo, Alessandro Maiello, Alessio Botta, and Giorgio Ventre. 2021. 2 years in the anti-phishing group of a large company. *Computers & Security*, 105:102259.

Francisco Jáñez-Martino, Rocío Alaiz-Rodríguez, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. 2023. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 56(2):1145–1173.

Ritik Mehta, Olha Jurečková, and Mark Stamp. 2024. A natural language processing approach to malware classification. *Journal of Computer Virology and Hacking Techniques*, 20(1):173–184.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

Ashokkumar Palanivinayagam, Claude Ziad El-Bayeh, and Robertas Damaševičius. 2023. Twenty years of machine-learning-based text classification: A systematic review. *Algorithms*, 16(5).

Luis Ángel Redondo-Gutierrez, Francisco Jáñez Martino, Eduardo Fidalgo, Enrique Alegre, Víctor González-Castro, and Rocío Alaiz-Rodríguez. 2022. Detecting malware using text documents extracted from spam email through machine learning. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, DocEng '22.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.