CECILIA: Enhancing CSIRT Effectiveness with Transformer-Based Cyber Incident Classification

Juan José Delgado, Eduardo Fidalgo, Enrique Alegre,
Andrés Carofilis, Alicia Martínez-Mendoza
Department of Electrical, Systems, and Automation, Universidad de León, León, ES
jdelgs01@estudiantes.unileon.es,
eduardo.fidalgo, enrique.alegre, andres.carofilis, alicia.martinez@unileon.es

Abstract

This paper introduces an approach to improving incident response times by applying various Artificial Intelligence (AI) classification algorithms based on transformers to analyze the efficacy of these models in categorizing cyber incidents.

As a first contribution, we developed a cyber incident dataset, CECILIA-10C-900, collecting cyber incident reports from six qualified web sources. The contribution of creating a dataset on cyber incident detection is remarkable due to the scarcity of such datasets. Each incident has been tagged by hand according to the cyber incident taxonomy defined by the CERT (Computer Emergency Response Team) of the National Institute of Cybersecurity (INCIBE). This dataset is highly unbalanced, so we decided to unify the four least represented classes under the label "others", leaving a dataset with six categories (CECILIA-6C-900). With these reliable datasets, we performed a comparison of the best algorithms specifically for the cyber incident classification problem, evaluating eight different metrics on two conventional classifiers and six other transformer-based classifiers.

Our study highlights the importance of having a rapid classification mechanism for CSIRTs (Computer Security Incident Response Teams) and showcases the potential of machine learning algorithms to improve cyber defense mechanisms. The findings from our analysis provide valuable insights into the strengths and limitations of different classification techniques. It can be used in future work on cyber incident response strategies.

1 Introduction

There is a steady increase in cyber attacks worldwide, showing a clear need for better incident response methods. For example, in 2023, X-Force recorded the highest number of incidents in Europe in the last years, with an increase of 31% compared to 2022 (IBM X-Force Incident Response Services, 2024).

CSIRTs need to enhance their capacities to manage a growing number of cyber incidents, especially in the first step of the process: classifying the reported incidents. A good and fast classification makes it possible to follow each incident to the appropriate expert group and directly impacts improving the CSIRT response times.

The traditional automatic classification approach is based on incident reporting standardization. Still, it is difficult to achieve cyber incident reporting harmonization, that is, aligning different standards to work together more effectively without losing their individual characteristics (Brumfield, 2023). Therefore, multiple standards represent reporting information in diverse formats, making the task of classification difficult. To solve this problem we will work on classifying cyber incidents reported from various sources and without any prior standardization criteria using NLP-based classification techniques in general and transformer classification models in particular, having not found any study that applies transformers to the classification of cyber incidents. The obtained results may be helpful for future work in AI-assisted cyber incident classification processes.

This paper introduces CECILIA (**C**yb**E**rin**CI**dents c**L**assified **I**ncibe t**A**xonomy), a cyber incident dataset created using different cyber incident reports collected from six selected web sources and manual tagging according to INCIBE taxonomy[1]. We present two versions, CECILIA-6C-900 and CECILIA-10C-900, where cyber incidents are classified into six and ten categories, respectively. After that, we compute the baseline results for two traditional and six transformer-based approaches using CECILIA in the task of cyber incident classification.

The rest of the paper is divided as follows: Sec-

[1] https://www.incibe.es/incibe-cert/incidentes/taxonomia

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 186–195
July 29–30, 2024.

tion 2 analyzes the literature about incident classification using AI, cyber incident datasets, and multilabel classification with transformers. Section 3 describes our CECILIA-10C-900 dataset, and in Section 4, we apply different transformer-based algorithms to this dataset and discuss the results achieved. We also introduce CECILIA-6C-900 to avoid unbalancing problems and discuss again the new results obtained with this new dataset. In Section 5, we present our conclusions and future work.

2 Related work

Depending on the nature of the source, there are different approaches for AI-assisted cyber incident classification.

Andrade and Yoo (2019) established a cognitive security model called NOTAS-MH, considering several sources of information, such as those generated by humans, signals from a computer or network equipment, open-source information, sensing instruments, and geospatial systems. Sapienza et al. (2018) presented DISCOVER, an algorithm to predict cyber threats in online discussions using NLP. To test it, they used their own manually curated dataset of security warnings from experts' tweets, security blogs, and dark web forums, obtaining a precision of 84% on tweets, 59% on blogs, and 81% on average.

Another possible source is OSINT (Open Source Intelligence), which was used by Tundis et al. (2022), classifying incidents according to their risk with a parameter called "relevancy score". They made this process in four phases: source identification, feature selection, score definition, and model training. In model training, they used five regression algorithms: an Support Vector Machine (SVM) Regressor, a Random Forest Regressor, a Gradient Boosting Tree regression, an Extra Trees Regressor, and a Multi-Layer Perceptron, and applied them in a dataset with tweets and Twitter profiles chosen in a survey with security experts.

Other approach is the use of standardization for incident reporting. In this way, Posea et al. (2022) proposed a common European taxonomy for incident handling and reporting and Colome et al. (2019) proposed to work with incident information in Incident Object Description Exchange Format (R. Danyliw (CERT), J. Meijer (UNINET), 2007) format to provide some resolution guidelines using Case-Based Reasoning methods in their dataset with 259 different incidents collected from the security division of a commercial data center.

Abbiati et al. (2020) merged three different datasets from 2005 to 2018 derived from three websites: PRC (Privacy Rights Clearinghouse), which maintains a collection of data-breach records[2], ITRC (The Identity Theft Resource Center) provides a collection of data breaches on a yearly basis[3] and BLI (The Gemalto Data Breach Level Index) containing datasets of publicly disclosed data breaches[4]. D'Ambrosio et al. (2023) proposed the use of this dataset as future work in risk management using Bayesian decision methods, and Rafaiani et al. (2023) proposed the Cyber Risk Assessment method that combined probabilistic methods and SVM and tested it with this and other two datasets ((Upguard, 2023), (Ransomfeed, 2023)).

Transformers are an excellent option for NLP classification problems, specifically in cases with multiple output classes (multiclass classification), However, to date, no studies have been found on the classification of cyber incidents using transformer models. Therefore, we will approach this problem using a generic multiclass classification perspective. In this field, Dogra et al. (2022) reviewed the entire process of state-of-the-art text classification models, collecting the benefits and limitations of each model. In the case of transformers, they highlighted the advantage of attention in long sentences, but on the other hand, they are computer-intensive.

Li et al. (2022) presented a survey on text classification with different datasets, types of classification (sentiment analysis (SA), news classification (NC), topic labeling (TL), question answering, natural language inference (NLI), multi-label (ML) and others) and metrics for evaluation, finding that the best results for all the datasets were obtained for pre-trained-transformer-based models like BERT, RoBERTa, and XLNET.

Furthermore, Gasparetto et al. (2022) made a survey of text classification for different tasks (SA, TL, NC, QA, NLI, Named Entity Recognition and Syntactic Parsing, discussing the preprocessing, representation, and testing of seven algorithms (Naive Bayes, Linear SVM, FastText Classifier, BiLSTM, XML-CNN, Bert and XLM-R) with EnWiki-100 and RCV1-57 datasets and found that best results

[2] https://privacyrights.org/data-breaches

[3] https://www.idtheftcenter.org/publication/2022-data-breach-report/

[4] https://web.archive.org/web/20191115194239/https://www.breachlevelindex.com/ Gemalto was acquired by Thales, and this website is no longer maintained

Institution	URL
European Repository of Ciber Incidents	https://eurepoc.eu/dashboard
Council on Foreign Relations	https://www.cfr.org/cyber-operations/
Internet Corporation for Assigned Names and Numbers	https://www.icann.org
Center for Strategic and International Studies	https://www.csis.org/programs/
CISSM Cyber Attacks Database	https://cissm.liquifiedapps.com/
Open Web Application Security Project	https://owasp.org/

Table 1: URLs selected for cyber incident collection

were achieved with Transformer-based models, like BERT and XLNet. Jáñez-Martino et al. (2023) evaluated 16 pipelines combining four text representation techniques: Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words, Word2Vec and BERT, and four classifiers: SVM, Naive Bayes, Random Forest, and Logistic Regression to perform a topic-based class detection of malware in spam messages.

There are several works on IA applied to the classification of cyber incidents but none of them deals specifically with the problem of CSIRTs. There are two ways of working: the standardization of reports, which has the disadvantage that the report must be carried out by specialized personnel, and on the other hand the use of NLP techniques. In this case, traditional classifiers are applied and the scarcity of datasets with cyber incident reports is shown.

The novelty of the present study lies in the use of transformers for the classification of cyber incidents, because, to the best of our knowledge, no similar approach exists. To enable a comprehensive comparison with different types of transformers, it was also necessary to create a reliable dataset. This dataset has been labelled according to the INCIBE taxonomy, which is based on the taxonomy of ENISA, the European Union Agency for Cybersecurity (Security and Information, 2018).

3 CECILIA datasets

CECILIA datasets comprise 923 cyber incident reports collected from six selected sources and then manually curated and classified using INCIBE cyber incident taxonomy provided for incident reporting (Instituto Nacional de Ciberseguridad, 2020).

After conducting a search for potential websites containing cyber incident reports, we prioritized sources that provided comprehensive compilations of cyber incidents in PDF or CSV formats, each incorporating unique classification systems. We selected a set of six URLs based on the highest quality of their reports and the prestige of their institution. URLs selected are shown in Table 1 . Subsequently, we extracted the textual content from these documents and classified them according to the taxonomy provided by INCIBE. Since the cyber incidents were presented in an easily exportable text format, the samples were simply extracted literally and transferred to a new spreadsheet.

A cybersecurity expert and a labelling assistant with mutual supervision and consensus in difficult-to-label samples have done the labelling process. Explanations and examples provided by INCIBE[5] were used as criteria to perform the labeling. INCIBE taxonomy divides cyber incidents into 10 categories and 38 subcategories The main ten categories are the ones reflected in CECILIA-**10C**-900 version (10C stands for ten categories): abusive content (AC), malicious code (MC), information gathering (IG), intrusion attempts (IA), intrusions (I), availability (A), information content security (ICS), fraud (F), vulnerable (V) and others (O). In Fig. 1, we can observe the imbalanced distribution of the CECILIA-10C-900 dataset, where most samples belong to the ICS class. The emergence of this distribution may suggest that specific cyber incidents are less frequent in real-world environments. However, a deep study of the real-life distribution should be performed to avoid biased behavior. In Section 4.4, an alternative dataset, CECILIA-6C-900, is proposed to mitigate the issue of significant imbalance.

The dataset has three fields: Description, category, and subcategory. Incident descriptions are written in non-technical English and span between 103 and 4299 characters. Some samples of the CECILIA-10C-900 dataset are shown in Table 2.

4 Experimentation

This section describes the experimental setup, including the transformer-based models and evaluation metrics used to assess the performance of these models in classifying incidents according to INCIBE's taxonomy.

[5]An updated version can be consulted at https://github.com/enisaeu

Description	Category	Subcategory
An unknown actor took control of the Instagram account of the police authority of the German city of Brunswick during the night of 4-5 January 2024. The hijacked account with around 13,000 followers subsequently published suggestive ads, (...)	AC	Spam
The state-sponsored Iranian hacker group MYSTICDOME (also known as UNC1530, CHRONO KITTEN, STORM-0133) infected four cell phones in Israel with SOLODROID malware, Google′s Threat Analysis Group and Mandiant (...)	MC	Infected System
The financially-motivated group 'Scattered Spider' gained access to telecommunication and other business process outsourcing organization′s networks in December 2022, through SIM swapping. According to a report by Trellix from 17 August 2023, (...)	IG	Social Engineering
The Russian military intelligence service GRU exploited the Microsoft Exchange vulnerability ProxyShell to gain access to a Ukrainian target in January 2022 and subsequently wipe that target in February 2022 at the start of the war, (...)	IA	Exploitation of Known Vulnerabilities
Multiple APT groups with suspected state links to Iran (Charming Kitten and APT34) and China (Hafnium, Elderwood, and APT31) have exploited a critical vulnerability (CVE-2022-40684) in several Fortinet products prior to its public reporting, (...)	I	Application Compromise
North Korea has been hit by a massive cyber attack according to the declaration of a South Korean government official that also added the government of Seoul is investigating on the event denying every responsibility. Russia′s ITAR-TASS (...)	A	DDoS
Dynamite Panda breached the US-American health provider Community Health, and exfiltrated 4.5 Millions of confidential patient data. The attribution of Dynamite Panda is at that point unclear, some seeing them as cyber-criminals, (...)	ICS	Unauthorised Access
In 2021, the Chinese hacking group IndigoZebra impersonated the Afghan president in spear-phishing emails to infiltrate the National Security Council. This cyber attack is part of a larger campaign across Central Asia since 2014, (...)	F	Phishing
According to Bloomberg, a Chinese PLA unit managed to infiltrate the Chip production of the company SuperMicro, opening up entrance paths into the systems of important American companies, including Amazon and Google	V	Vulnerable System
Iranian hackers were identified in a report released Tuesday as the source of coordinated attacks against more than 50 targets in 16 countries, many of them corporate and government entities that manage critical energy, transportation, and medical services.	O	Uncategorised

Table 2: Example of CECILIA100-900 dataset samples. One sample of each category is shown.

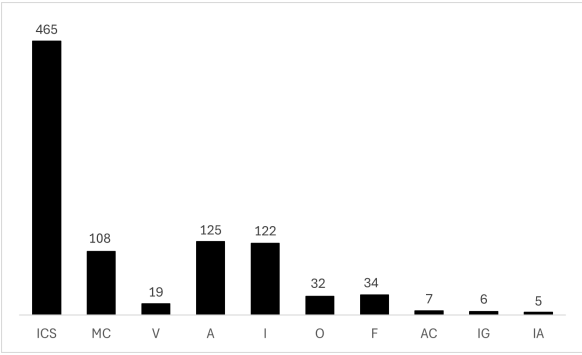

Figure 1: Class distribution in CECILIA-10C-900 dataset. Category of Information Content Security has almost 50% of the samples of CECILIA-10C-900, while Vulnerable (V), Abusive Content (AC), Information Gathering (IG), and Intrusion attempts (IA) contain each less than 20 incidents.

4.1 Models and evaluation metrics

The experiment was conducted using Simpletransformers[6] version 0.70.1. This Python library provides a high-level interface for easily utilizing Transformer models in NLP tasks and enables rapid and efficient AI application development with minimal required configuration. Using this library, we can compare various models under uniform conditions without additional configurations, parameters, or preprocessing tasks.

We selected six Transformer-based state-of-the-art models for our evaluation: DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), ELECTRA (Xu et al., 2020), Longformer (Beltagy et al., 2020) and MP-Net (Song et al., 2020) to apply to our dataset. The configuration for all the models is 6 epochs, and the maximum number of tokens is 512 using the default values for all hyperparameters.

Also, we computed baseline results using CECILIA with two traditional machine learning classifiers: Logistic Regression with TF-IDF feature extractor and K-Nearest Neighbor with Bag of Words

[6] http://simpletransformers.ai

(BoW). These classifiers are well-performed models in other cybersecurity text classification problems, such as malware detection using the text of spam emails (Redondo-Gutierrez et al., 2022). This test will be useful for comparing the performance of traditional classifiers with that of transformer-based models.

CECILIA-10C-900 contains 923 samples, which could be considered a small dataset for NLP tasks. Therefore, we use stratified K-Fold cross-validation with k=5 (5 splits) and data shuffled.

The cyber incident classification problem we address consists of selecting the category from INCIBE taxonomy that better represents each cyber incident. This is a multiclass classification problem, which requires adapting binary classification metrics to measure performance accurately and may also require the use of new metrics (Grandini et al., 2020). In this case, we evaluated the models with the following metrics:

- Accuracy: the total number of well-classified samples divided by the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

  TP is the number of True Positives, TN is the number of True Negatives, FP is the number of false positives, and FN is the number of false negatives.

- Variance: as we are working with k-fold cross validation, it is important to calculate also the variance value.

- Precision: defined as the True Positive elements divided by the total number of positively predicted.

$$Precission = \frac{TP}{TP + FP} \quad (2)$$

  In the case of multiple classes, we use both Precision weighted and Precision macro. For Precision weighted, we calculate metrics for each label and find their average weighted by the number of true instances for each label. This formula is more realistic for imbalanced datasets.

$$Prec - w = \sum_{i=1}^{N} w_i * Prec_i \quad (3)$$

  w is the weight of each class and N is the number of classes. For Precision macro, we only calculate the average of all precision values for each category.

$$Prec - m = \frac{\sum_{i=1}^{N} Prec_i}{N} \quad (4)$$

- Recall: the division of True Positive elements and the total number of positively classified units (True Positives and False Negatives)

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

  Also, we will calculate macro and weighted values for the Recall.

- F1-score: the harmonic mean of Precision and Recall

$$F1 - score = 2 * \left(\frac{precision * recall}{precision + recall}\right) \quad (6)$$

  Additionally, we will calculate macro and weighted values for F1-score.

- Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (7)$$

4.2 Results and discussion

The results of our experiment are collected in Table 3, where it can be seen that transformer-based models always perform better than traditional models in every metric calculated. The best results are obtained by the XLNet model in all the values (0.8385 accuracy and 0.7668 MCC), closely followed by the RoBERTa model (0.8245 accuracy and 0.7463 MCC).

Although ELECTRA achieves the lowest performance (0.7984 accuracy and 0.7059 of MCC) out of the transformer-based models, it still outperforms traditional classifiers. For BERT-based models, RoBERTa achieves the second-best results (0.8245 in accuracy and 0.7463 MCC) and DistilBERT remains above 80% of accuracy (0.8039 accuracy and 0.7151 of MCC) with a lower computational load.

4.3 Discussion

The advantages of XLNet, particularly its enhanced context understanding through a bidirectional approach, seem to be successful in improving BERT

Model	Accuracy	Variance	Prec-w	Prec-m	Recall-w	Recall-m	F1-score-w	F1-score-m	MCC
LR TF-IDF	0.7364	0.0000	0.7012	0.3960	0.7365	0.3646	0.7150	0.3740	0.6101
KNN BoW	0.5812	0.0000	0.4791	0.1824	0.5812	0.1950	0.5162	0.1815	0.3464
DistilBERT	0.8039	0.0007	0.7715	0.4251	0.8039	0.4165	0.7840	0.4147	0.7151
RoBERTa	0.8245	0.0008	0.8080	0.4767	0.8245	0.4814	0.8127	0.4724	0.7463
XLNet	**0.8385**	0.0007	**0.8250**	**0.4795**	**0.8385**	**0.4854**	**0.8272**	**0.4622**	**0.7668**
ELECTRA	0.7984	0.0006	0.7516	0.3659	0.7984	0.3798	0.7681	0.3613	0.7059
Longformer	0.8201	0.0006	0.7964	0.4462	0.8201	0.4559	0.8057	0.4467	0.7385
MPNet	0.8201	0.0007	0.7737	0.4232	0.8201	0.4410	0.7936	0.4259	0.7372

Table 3: Incident classification results over CECILIA-10C-900 dataset with two traditional classifiers and six transformer-based models. The best results are in bold.*-w and *-m stands for weighted and macro average in each metric

models like RoBERTa. Moreover, all metrics have similar values, so the model is efficient in all use cases. The choice of the key metric for this problem will depend on the impact of misclassifying a cyber incident. If those incidents not correctly classified are forwarded to their correct destination quickly, accuracy will provide the best performance whereas if a critical incident is incorrectly classified and the time to attention is important, F1-score will be a more appropriate metric.

However, while it was anticipated that the performance of MPNet, as it combines masking as BERT and permutation as XLNET would be in the range of XLNet and RoBERTa, it exhibits inferior results. This may be attributable to the limited dataset. Also, the advantages of LongFormer do not seem to be fully leveraged since the length of the samples under consideration is not sufficiently extensive.

Both Longformer and MPNet exhibit comparable outcomes. However, Longformer demonstrates superior performance in precision ($0,7964$ vs 0.7737 in weighted precision) and F1-score ($0,8057$ vs 0.7936 in weighted F1-score). This distinction suggests the importance of having long samples to minimize false positives.

Among the models with a more efficient computational load, DistilBERT exhibits the best performance (0.79 seconds per sample), followed by ELECTRA (far from DistilBERT with 1.52 seconds), Roberta (1.56 seconds), MPNet (1.61 seconds) and the last results are for XLNet ($2,96$ seconds) and Longformer (2.99 seconds). This may be attributed to having an unbalanced dataset. As the training dataset is highly unbalanced, we expected lower performance in terms of precision and recall. Quite satisfactory results were achieved with weighted values but were poor in macro values.

Values of MCC over 0.7 in all the models show a good general performance of all alternatives in cyber incident classification. Although fine-tuning mechanisms could improve the final values, our goal is to compare different methods and then focus on one of them for fine-tuning. We identify XLNet as the best-performing model and DistilBERT as the model with better results (0.8039 accuracy and 0.7840 F1-score weighted and lower computational costs (0.79 seconds per sample).

As XLNet obtained the best results, we will focus on it to get more information about its performance. As shown in Table 4, under-represented categories like V, IG, and IA never obtained a correct classification. Therefore, we can deduce that increasing the number of training samples or utilizing a balanced dataset will enable enhanced outcomes. This problem not only appears in XLNet but also in every model tested. Traditional methods also yield significantly low values in the macro-average and result in null classification for these categories.

	k=1	k=2	k=3	k=4	k=5
ICS	0.8829	0.8972	0.8969	0.8913	0.9197
MC	0.8837	0.8695	0.8936	0.7804	0.8500
A	0.9130	0.8936	0.9803	0.8800	0.9411
I	0.7636	0.6000	0.6086	0.7368	0.8000
O	0.3333	0.3333	0.4000	0.3333	0.5333
F	0.7272	0.8000	0.6667	0.8571	0.8235
AC	0.0000	0.0000	0.0000	0.5000	0.0000
V	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
IG	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
IA	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**

Table 4: F1-score values for each cross-validation split in XLNet model for every category. In each column, k represents the number of the split. As we can see, samples in the **last three categories** were never properly classified.

4.4 Balancing the dataset

To address the issues of high imbalance and poor performance in specific categories, AC, V, IG, and

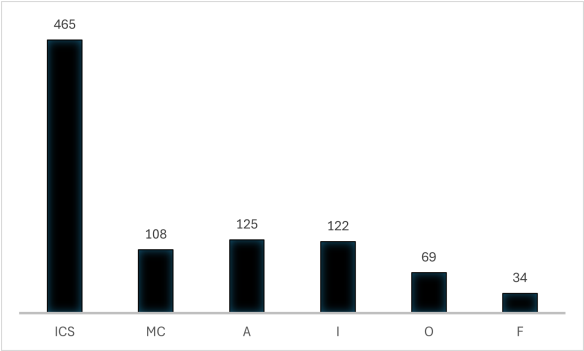

Figure 2: New class distribution in CECILIA-6C-900 dataset merging the four low representative categories with "others".

IA have been grouped under a category we named Others (O). The distribution of the modified dataset, called CECILIA-6C-900, is presented in Figure 2. Working with CECILIA-6C-900 might be helpful for training specific intelligent models that could detect cyber incidents of these four minority categories that a specific department of a CERT could later handle.

The results in this case are presented in table 5. Macro and weighted metrics have closer values (0.8245 F1-score weighted and 0.7476 in XLNET with CECILIA-6C-900 against 0.8272 and 0.4622 before) and the best values this time have been achieved by MPNet (0.8352 accuracy and 0.8273 F1-score-weighted), although its overall performance exhibits a slight decline in accuracy (0.8352 vs 0.8385) and MCC (0.7608 vs 0.7668). This could be attributed to the difficulty in classifying samples with heterogeneous themes under a single category. In this case, MPNet performs better, improving its results (0.8352 vs 0.8201 of accuracy). DistilBERT also improves their last values (0.8352 vs 0.8201 of accuracy), while RoBERTa (0.8169 vs 0.8080 of accuracy), XLNet(0.8256 vs 0.8385 of accuracy), and Electra (0.7865 vs 0.7984 of accuracy) are getting worse, and LongFormer (0.8166 vs 0.8201 of accuracy) remains at very similar values. These results can help us to assess the performance of different models in highly unbalanced datasets.

Again, if we analyze the best-performing model in each split of cross-validation, as we can see in table 6, now the F1-score for the less-occurrence categories has on the CECILIA-6C-900 dataset compared to the CECILIA-10C-900 dataset, achieving values ranging from 0.27 to 0.45, thereby enhancing the overall performance of the model.

5 Conclusions and future works

In this work, we have evaluated two traditional classifiers and six models based on transformers using the CECILIA-10C-900 dataset. The results show that transformer-based models outperform traditional classifiers.

The outstanding performance demonstrated by Transformer models strongly suggests that adopting this technology constitutes a promising strategy for the development of applications and services aimed at cyber incident classification. The ability of these models to capture complex contextual dependencies in extensive text sequences allows them to achieve high levels of accuracy in identifying and categorizing texts related to cybersecurity incidents.

The implementation of Transformers in cybersecurity expanded the ability to anticipate, detect, and respond more effectively to security threats, thereby contributing to the fortification of digital infrastructures against cyber attacks.

Given the evidence on the superior performance of Transformer-based models, developing applications and services focused on cyber incident classification, grounded in this technology, represents an appropriate approach for applying artificial intelligence to cybersecurity. This approach is justified not only by the demonstrated efficacy in precise text classification but also by the adaptability and scalability of Transformer models, which can be trained and fine-tuned to meet specific requirements in the field of cybersecurity.

Future research can be based on conducting further experiments by expanding and balancing the dataset used for training and evaluation. Augmenting the dataset can provide a more comprehensive representation of the linguistic and contextual diversity inherent to cybersecurity texts. This expansion is expected to enhance the model's ability to generalize from training to unseen data, thereby improving its robustness and reliability in real-world applications.

Additionally, addressing the issue of dataset imbalance can avoid bias toward the over-represented classes. By providing a richer and more balanced training foundation, the models are expected to achieve higher levels of performance in terms of accuracy and their capacity to handle a broader spectrum of cyber incident types.

Another possibility for improvement involves the completion of the dataset with all categories

Model	Accuracy	Variance	Prec-w	Prec-m	Recall-w	Recall-m	F1-score-w	F1-score-m	MCC
LR TF-IDF	0.7292	0.0000	0.7293	0.6665	0.7292	0.6024	0.7268	0.6252	0.6048
KNN BoW	0.5848	0.0000	0.5267	0.3881	0.5848	0.3372	0.5290	0.3255	0.3558
DiltilBERT	0.8093	0.0007	0.7994	0.7305	0.8093	0.7093	0.8013	0.7130	0.7205
RoBERTa	0.8169	0.0003	0.8186	0.7384	0.8169	0.7384	0.8151	0.7338	0.7350
XLNet	0.8266	0.0006	0.8300	0.7492	0.8231	0.7272	0.8265	0.7476	0.7490
ELECTRA	0.7865	0.0002	0.7699	0.6436	0.7865	0.6000	0.7667	0.5919	0.6887
Longformer	0.8201	0.0003	0.8166	0.7453	0.8201	0.7538	0.8167	0.7465	0.7399
MPNet	**0.8352**	0.0007	**0.8314**	**0.7741**	**0.8352**	**0.7489**	**0.8273**	**0.7494**	**0.7608**

Table 5: Incident classification results over CECILIA-6C-900 dataset after merging the four representative categories inside a category "others" using two traditional classifiers (LR+TF-IDF and kNN+BoW) and six transformer-based models. The best results are in bold.

	k=1	k=2	k=3	k=4	k=5
ICS	0.9297	0.8842	0.9312	0.8938	0.9109
MC	0.9047	0.8837	0.8837	0.7111	0.8571
A	0.8518	0.9615	0.8979	0.8846	0.9130
I	0.7547	0.6086	0.6037	0.7407	0.7142
F	1	0.6153	0.9230	0.7500	0.7692
O	0.2727	0.3846	0.4347	0.4545	0.3000

Table 6: F1-score values for each split of cross-validation in MPNet model for every category with CECILIA-6C-900 dataset. In each column, k represents the number of the split. "Other" category obtained lowest values in each split because by joining different classes the samples are heterogeneous and therefore more difficult to classify under the same category.

from the taxonomy of INCIBE currently not represented in CECILIA dataset. The current dataset, while extensive, does not fully cover all the groups of this taxonomy, resulting in certain types of cyber incidents being underrepresented or absent. By integrating these missing classes into the dataset, the model can be trained to recognize and classify a more complete spectrum of cyber incidents.

Finally, another way to improve future work involves enhancing the granularity of our classification approach, extending into subcategory precision. This refinement aims to yield a more detailed classification of cyber incidents.

Moreover, incorporating multi-label classification models or hierarchical classification structures can significantly improve the accuracy and performance of the classification model developed.

Limitations

To obtain the best possible comparison, we developed CECILIA-10C-900, a dataset of cyber incident reports that have been properly tagged and curated, although so far this dataset contains a limited set of 923 samples.

To this end, it is necessary to continue improving the dataset and obtaining the most reliable data possible from real cyber incident reports. In instances where the dataset appears highly imbalanced due to the infrequent occurrence of certain types of cyber incidents, the procedure of consolidating them under a single category has proven to be effective and may align with actual cyber incident response procedures. However, this work is challenging as this information is usually not public. Another potential path for future works may involve employing data augmentation techniques to mitigate the issue of categories with sparse samples.

Completing the dataset in alignment with the INCIBE taxonomy has significant implications for the practical application of the trained model. It would enable the model to work in real-world scenarios.

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 development of **fast and accurate systems** to classify cyber incidents in CSIRTs can contribute to improving their performance and, therefore, their incident response mechanisms.

The development of artificial intelligence (AI) models for classifying cyber incidents, particularly those utilizing Transformer architectures, carries significant ethical implications that warrant thorough consideration.

Bias: The dataset can contain biases related to incident types, geographic origins, or any other factors that could lead to unfair model outcomes in different fields of cyber incident classification.

Impact on Cybersecurity Workforce: We are mindful of the concerns related to automation and its potential impact on employment within the cybersecurity industry. Our intention is not to replace human experts but to augment their capabilities, enabling them to respond more effectively and ef-

ficiently to cyber threats. By automating routine tasks, we aim to free cybersecurity professionals to focus on more complex and strategic challenges.

Use of AI Technologies: We recognize the potential for misuse of AI technologies, including the possibility of adversarial attacks. We advocate for the ethical use of AI in cybersecurity, emphasizing its role in protecting individuals, organizations, and societies against cyber threats.

Data availability

The data used in this study will be publicly available under request.

Acknowledgements

This work was supported by the Strategic Project LUCIA granted to the University of León by the Spanish National Cybersecurity Institute (INCIBE) and funded by the Next Generation EU funds.

References

Giovanni Abbiati, Silvio Ranise, Antonio Schizzerotto, and Alberto Siena. 2020. Merging Datasets of CyberSecurity Incidents for Fun and Insight. *Frontiers in Big Data*, 3.

Roberto O. Andrade and Sang Guun Yoo. 2019. Cognitive security: A comprehensive study of cognitive science in cybersecurity. *Journal of Information Security and Applications*, 48.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

Cynthia Brumfield. 2023. Harmonization of cyber incident reporting to the federal government | homeland security. *CSO Online*.

Marcelo Colome, Raul Ceretta Nunes, and Luis Alvaro De Lima Silva. 2019. Case-based cybersecurity incident resolution. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, 2019-July:253–258.

Nicola D'Ambrosio, Gaetano Perrone, and Simon Pietro Romano. 2023. Including insider threats into risk management through Bayesian threat graph networks. *Computers Security*, 133:103410.

Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. 2022. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*, 2022.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A Survey on Text Classification Algorithms: From Text to Predictions. *Information 2022, Vol. 13, Page 83*, 13(2):83.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. 2020. Metrics for Multi-Class Classification: an Overview.

IBM X-Force Incident Response Services. 2024. X-Force Threat Intelligence Index 2024. Technical report, IBM.

Instituto Nacional de Ciberseguridad. 2020. Guía nacional de notificación y gestión de ciberincidentes.

Francisco Jáñez-Martino, Rocío Alaiz-Rodríguez, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. 2023. Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139:110226.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):31.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Vlad Posea, George Sharkov, Adrian Baumann, and Georgios Chatzichristos. 2022. Towards unified european cyber incident and crisis management ontology. *Information Security: An International Journal*, 53:33–44.

Y. Demchenko (University of Amsterdam) R. Danyliw (CERT), J. Meijer (UNINET). 2007. RFC 5070 - The Incident Object Description Exchange Format. Technical report, IETF Network Working Group.

Giulia Rafaiani, Massimo Battaglioni, Simone Compagnoni, Linda Senigagliesi, Franco Chiaraluce, and Marco Baldi. 2023. A Machine Learning-based Method for Cyber Risk Assessment. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2023-June:263–268.

Ransomfeed. 2023. DRM - Dashboard Ransomware Monitor. Accessed: 2023-10-02.

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. *DocEng 2022 - Proceedings of the 2022 ACM Symposium on Document Engineering*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining Online Chatter for Emerging Cyber Threats.

European Union Agency For Network Security and Information. 2018. Reference Incident Classification Taxonomy — ENISA.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *Advances in Neural Information Processing Systems*, 2020-December.

Andrea Tundis, Samuel Ruppert, and Max Mühlhäuser. 2022. A Feature-driven Method for Automating the Assessment of OSINT Cyber Threat Sources. *Computers & Security*, 113:102576.

Upguard. 2023. Breaches. Accessed: 2023-10-02.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2579–2591.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32.