

# Improving Cybersecurity Awareness: Tweet Classification using Multilingual Sentence Embeddings and Contextual Features

Anastasia Coto<sup>[0000-0001-8559-3658]</sup>, Carlo Bono<sup>[0000-0002-5734-1274]</sup>,  
Cinzia Cappiello<sup>[0000-0001-6062-5174]</sup>, and Barbara Pernici<sup>[0000-0002-2034-9774]</sup>

Politecnico di Milano - DEIB {firstname.lastname}@polimi.it

**Abstract.** The presence of professionals, interested general public, and cybersecurity intelligence accounts makes social media a valuable source for computer security awareness. By regularly capturing and analyzing the posts on emerging cyberthreats shared by these accounts, individuals and organizations can timely understand potential dangers and effectively implement mitigation strategies. However, retrieving relevant and informative posts from a social network without pulling data only from known and fixed accounts is very challenging due to the high percentage of posts containing only general remarks and other uninformative content. In this paper, we propose a novel approach based on building classifiers for selecting relevant tweets and categorizing them according to different types of vulnerabilities for a more effective use of this type of information. To accomplish this task, we designed a pipeline combining text classifiers in cascade, training them on manually labelled data. With active learning, we optimized the labelling process, reducing manual effort while enhancing the accuracy of the classification model. We experiment with FFNN architectures, combining language-agnostic sentence-level embeddings obtained with LASER with vectors describing past user activity extracted with User2Vec. To enhance the overall performance and prevent overfitting, architecture decisions and hyperparameter tuning were performed in a cross-validation setup. With an achieved accuracy of 87%, our approach offers effective classification of social media posts, empowering cybersecurity professionals to stay informed and take appropriate measures.

**Keywords:** social media analysis · posts classification · machine learning · security vulnerabilities.

## 1 Introduction

Nowadays, cybersecurity is one of the principal security concerns, mainly due to the rapid evolution of digital technology over the past few decades. With the majority of valuable data stored on accessible servers worldwide, cybercriminals target this information, posing significant threats to commercial companies, organizations, governments, and individuals. Specific security requirements have

been established to mitigate such risks, but smaller organizations often struggle to comply due to limited resources [5].

The potential of social media for extracting information about emergencies has been advocated in many crisis situations, and several approaches for extracting and analyzing posts have been developed in the literature, as illustrated in [13], and have been presented and discussed within the ISCRAM (Information Systems for Crisis Response And Management)<sup>1</sup> community. Specific platforms have been developed and used during crisis situations, and information has been extracted from general purpose sources, such as Twitter, Reddit, Facebook, and Flickr. Some social media, such as Reddit, host specialized channels in which thematic news are posted. Two factors have emerged as important for deriving information from general purpose social media: the wide availability of data, published by newspapers, specialized organizations, and the general public, and the timeliness of this data, since ongoing crisis events are reported within a very short time frame [19]. In the cybersecurity domain, this potential is also starting to be exploited: first of all, by posting on social networks warnings about vulnerabilities and new threats, both from official and specialized organizations and from the general public; furthermore, by analyzing social media posts to examine the perception of risks and the impact of threats from social media discussions.

On the other hand, using social media as a source of information poses several challenges. Most of the available tools are oriented to using the English language: in this way, more localized risks and threats could be missed. A general problem is also that, even when searching with specialized keywords, a large number of irrelevant posts are retrieved, which have to be filtered out in order to isolate useful information. In addition, usually each post is considered separately by the filtering tools, and in this way useful information about the posting behaviour of users is lost, while it could be exploited to understand valuable or irrelevant contributions better. All these problems are also identified as critical in other emergency situations, and there is a lack of a general approach to handle them. In some cases, crowdsourcing is proposed to support the classification of posts, but this might be too time-consuming in some scenarios, especially for cases [7, 6] in which timeliness is critical.

The goal of this paper is to propose a general approach to retrieve relevant posts in social media related to cybersecurity issues, focusing on two main contributions: i) selecting posts with relevant keywords and then classifying them based on a set of well defined vulnerabilities extracted from official sources: ii) exploring the contribution of including the user context as a subsidiary input. A crucial requirement is linked to the multi-lingual nature of social media posts. In fact, local organizations prioritize information posted in their country's language to better understand the risks they face locally. The proposed approach is multilingual, leveraging recent language-agnostic techniques for preprocessing the posts and training the classification models.

The scope of this research involves validating the functionality, performance, and usability of the proposed approach through experiments, using real-world

---

<sup>1</sup> <https://iscram.org>

data to test the algorithm, ensuring its effectiveness and adherence to the intended objectives.

The paper is structured as follows. In Section 2, we illustrate existing approaches to extract information from social media in the cybersecurity domain. In Section 3, we illustrate the underlying scenario motivating our work, which has the goal of providing awareness information for possible threats in a specific organization under development within the H2020 CS-AWARE-NEXT project<sup>2</sup>. In Section 4, we illustrate our approach to develop classification models for identifying posts related to vulnerabilities, discussing possible alternatives to be evaluated. Finally, in Section 5 we discuss the results obtained in a large case study in a multilingual context.

## 2 Related work

Official cybersecurity organizations and communities constantly share news and information about the evolution of the cyberspace and related security threats. The system departments use social media (blogs and forums), for finding information shared by other security companies [5]. Recently, social media has been widely used to collect and analyse data for the detection and characterization of events regarding various topics. Thanks to the immediate availability of real-time data, thematic information can be easily pulled.

The task of identifying cybersecurity-related data on social media has been widely investigated, focusing on detecting or predicting cyberattacks based on social media analysis. While some studies aim to extract suitable information related to cyber threats using NER algorithms [10], others build neural networks to classify relevant information [17, 18]. Research mainly focused on the textual contents of the posts, and has been mostly conducted on Twitter data, most likely because of its substantial data volume and the presence of an active community engaged in cybersecurity topics. A supervised learning approach to highlight contents related to cybersecurity is proposed in [12], based on sentiment analysis over a TF-IDF content representation. The work proposed in [8] focuses on the detection of denial-of-service attacks and evaluates a feed-forward neural network (FFNN) and an LDA model, deriving a fine-grained description of events with the latter enabling a further categorization by topic. The authors in [21] combine convolutional neural networks (CNN) and word embeddings to classify the presence and the severity of a cyber-threat; moreover, they propose a URL-based matching method to associate known Common Vulnerabilities and Exposures (CVE) to social media posts. [9] proposes a CNN approach to classify cybersecurity-related information from social media. An LSTM-based named entity recognition (NER) component highlights further relevant information contained in the posts. A classification method that leverages public repositories to evaluate if a post is related to cyber-threat intelligence is proposed in [15]. The authors in [14] combine similarity on word embeddings for content detection and

---

<sup>2</sup> <https://www.cs-aware-next.eu/>

community detection to identify the most relevant user groups in cyber-attacks. Frameworks analyzing Twitter data have been designed. One example is Cyber-Twitter [17], which studies cybersecurity tweets and issues timely threat alerts to security analysts based on an organization’s “system profile”. Alerts generated by CyberTwitter can then serve as input to various other security systems that can use them depending upon local organizational security policies.

As a contribution, we propose a two-stage classification method to filter relevant contents from social media streams, by first selecting relevant content and then applying a fine-grained categorization over known event types. The keywords for social media crawling are derived semi-automatically from validate data. The approach transparently supports multiple languages, both during training and at runtime. Finally, we take advantage of the centralization of the discussion around experts, by attaching the users’ post history as an additional input.

### 3 Scenario

The Horizon Europe CS-AWARE-NEXT project aims to provide organizations and local or regional supply networks with improved cybersecurity management capabilities. In particular, one main goal is to increase awareness about vulnerabilities, risks, and ongoing attacks.

To achieve such an objective, the ambition of CS-AWARE-NEXT is to collect data from several sources (internal and external to the organizations involved in the network), and develop an effective and efficient AI pipeline for analyzing data and getting useful insights. In particular, building on the results of the CS-AWARE H2020 project [5], which delivered excellent results in the detection of anomalies based on user-defined behaviour patterns derived during socio-technical analysis, CS-AWARE-NEXT is going to use AI / machine learning to correlate anomalous events detected within organisations, with context provided by threat intelligence, including the creation of contextualised mitigation and/or self-healing options. The CS-AWARE-NEXT pipeline is composed of the following steps:

1. *Data collection and cataloging*: The design of a proper data lake architecture is a prerequisite for ingesting internal and external sources. The internal sources include system log files and security-related documents, while external sources consider Web data such as structured cybersecurity threat intelligence or unstructured social media conversations.
2. *Data preparation and quality assurance*: it includes data transformation and cleaning components able to guarantee a certain data quality level. The goal is to detect and analyze only relevant and reliable data.
3. *AI data correlation models*: Building on the data quality assurance and cataloging activities, the AI models for data correlation and contextualisation can rely on effective and efficient access to the data. With the use of AI technologies such as deep learning and reinforcement learning, anomalies

and threat will be detected and valuable information that need to be shared with the organizations involved in the network will be identified.

In this paper, we focus on the social media analysis approach, aiming to select tweets related to cybersecurity vulnerabilities, attacks, and experiences and extract the information to be broadcasted to organizations to increase their awareness.

Currently, the CS-AWARE-NEXT platform (as shown in Figure 1) provides users with an interface that lists the relevant tweets and associates them with the architectural components to which their content refers. For example, if a tweet refers to a vulnerability in a specific application, the server containing that application is highlighted in the architecture representation. The tweets are extracted from selected specialized accounts that post news about vulnerabilities that have already been certified. We aim to improve the system by (i) extracting data monitoring the entire social media streams, as opposed to monitoring specific accounts only; (ii) using a multilingual approach also to consider local content and therefore extend the volume of input data; (iii) preprocessing tweets in a way to avoid generic or irrelevant text; (iv) finding information also about zero-day threats and not only certified vulnerabilities.

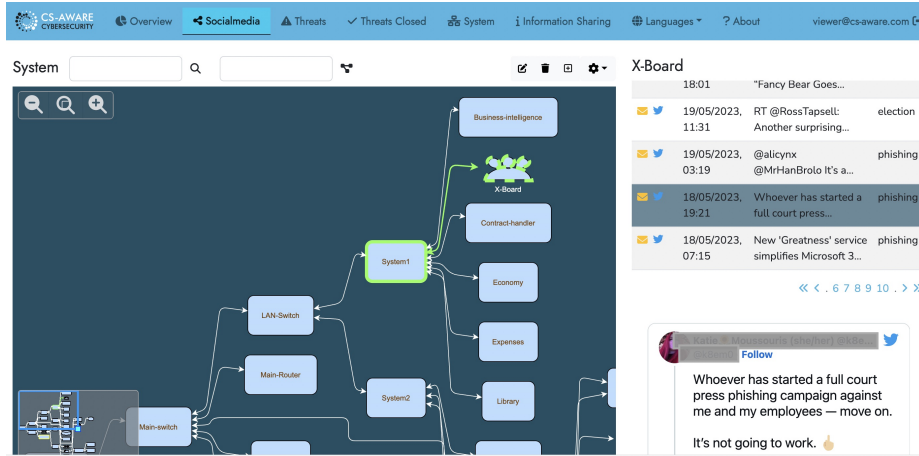


Fig. 1. CS-AWARE social media dashboard

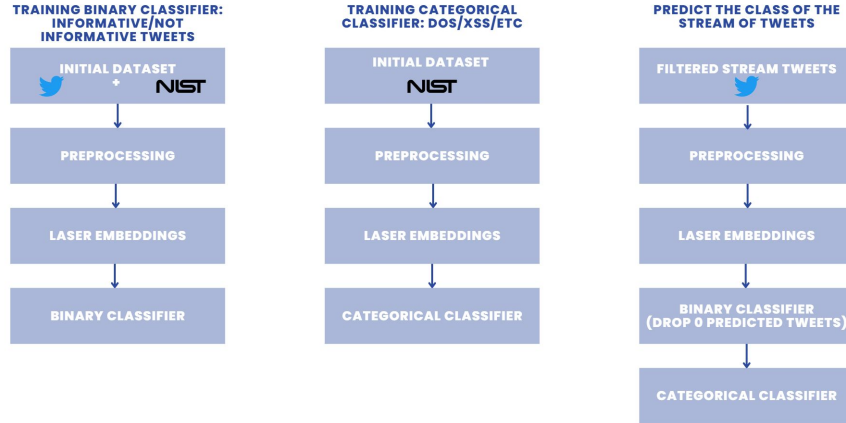
## 4 Approach

### 4.1 A Machine Learning Approach for Tweet Classification

The goal of the proposed approach is the automatic discovery and categorization of relevant posts from a social media stream without relying only on known

sources of information. The approach uses a binary filtering stage to select relevant posts, then assign a more fine-grained MITRE CVE<sup>3</sup> category to the contents deemed to be relevant, in order to pinpoint the specific type of vulnerability/threat. Moreover, we explicitly design the approach to be language-agnostic, meaning that the classification models are trained on multi-lingual data and are able to operate on many languages simultaneously, without relying on language-specific dictionaries or further configuration efforts. Since cybersecurity-related discussion on social media is highly centralized around experts, we exploit this characteristic by enriching the classifier inputs with representations of the user context, when available, condensing their past social media activity. To reach the intended objectives, a comprehensive architecture has been developed and validated using real-world data. To ensure the generalizability of the approach, multiple learning mechanisms are explored and selected through an iterative experimental process.

Consistently with approaches found in literature, we combine well-known domain data sources and posts crawled from Twitter to train the system. We utilize an active learning iterative approach to overcome limitations in the dataset generation, enabling quick iterations intended to update the classification models with fresher data. This approach is meant to be easily reproducible, and also enables model updating at runtime, as the cybersecurity landscape inevitably evolves.



**Fig. 2.** Tweet classification

<sup>3</sup> MITRE. Common vulnerabilities and exposures <https://cve.mitre.org/>

The overall approach is illustrated in *Figure 2 (on the right)* and consists of the following key steps: 1) Data collection using a relevant keyword dictionary, 2) Pre-processing by removing noise, irrelevant information, and inconsistencies, 3) Feature extraction using sentence embeddings [3] and representations to capture the user context [11], 4) Post filtering and classification with neural network classifiers.

At the design stage, a data collection of suitable “informative” / “not informative” tweets is performed. The quality of collected posts is critical in order to learn the proper concepts from data. The dataset preparation is discussed in Section 4.3. Moreover, ancillary data regarding vulnerability description reports are obtained. Preprocessing and representation choices are evaluated on the available data. Classifier design, training and optimization are jointly conducted on the embeddings extracted from the resulting datasets, using random K-fold cross-validation, with the number of folds equal to ten ( $k=10$ ).

We utilize established performance metrics such as accuracy, precision, recall, F1 score, AUC-ROC<sup>4</sup>, MCC<sup>5</sup>, and TT (Training Time) to assess the effectiveness of the approach and to determine the optimal configuration, as described in Section 5.

## 4.2 Vulnerabilities data sources

The primary data source considered for the training of the binary classifier is the National Vulnerability Database (NVD) created by the National Institute of Standards and Technology (NIST). The NIST vulnerability repository offers the possibility to download all the available CVE records archives grouped by year directly from their website [1]. For each record, a wide array of data fields is available, such as its category, description, severity score, affected product lists, additional references, the ability for community contributions, etc. The most significant data field for this study is represented by the CVE description, since it comprises accurate text information that is considered relevant and informative in relation to cybersecurity incidents.

Furthermore, social media streams are considered significant data sources from which information can be extracted. To ensure the relevance of the obtained data, pertinent keywords need to be selected for filtering the data stream. In this study, we derived relevant keywords from the NIST website, which contains a comprehensive cybersecurity vocabulary with terms and descriptions. Moreover, a semi-manual selection process was performed, by choosing the most prevalent cybersecurity vulnerability names (e.g., “buffer overflow”), types of cyber attacks (e.g., “malware”) and the most frequent occurrences of cybersecurity-related terms within the CVE descriptions (e.g. “data patch”, “remote code execution”, “software update”).

The chosen data source for classifying vulnerability and cyber-threat data by type is the CVE by Type dataset available on the CVE Details website [16].

<sup>4</sup> Area Under the Receiver Operating Characteristic Curve

<sup>5</sup> Matthews Correlation Coefficient

This dataset provides MITRE CVE data categorized into 12 different types and sorted by year, covering the period from 1999 to 2023. The vulnerabilities are classified using keyword matching and Common Weakness Enumeration (CWE) numbers when possible, but they are mostly based on keywords [16]. In the scope of this research, the 2022 CVEs were utilized because the descriptions are recent and the data volume is satisfying.

Besides posts obtained from live data stream, we also collected posts published by users belonging to a qualified user list. This list contains 36 users (see Appendix A for the detailed list) including Computer Emergency Response Team (CERT) accounts from various countries and individuals interested in cybersecurity. This is done in order to ensure that the training data contains multiple language tweets (e.g., “NationalCsirtCy”: Greek, “CSIRTPanama”: Spanish, “CSIRT\_MON”: Polish) belonging to the domain of interest. Some of the posts contain highly informative tweets (e.g., “csirt\_it”), while others share general information about cybersecurity that does not contain any specific vulnerability or threat (e.g., “irisscert”). This data is collected in order to enhance the classifier’s ability to efficiently discriminate informative tweets from non-informative ones even if the terminology within the text is similar.

### 4.3 Datasets

In order to build, train and validate the different proposed functionalities, three tailored datasets have been crafted. The datasets are publicly available on GitHub<sup>6</sup>. Social media posts constituting the datasets are extracted from Twitter, since at the date of this study it offered convenient APIs for data access, granting access to a suitable volume of data.

The first dataset combines NIST CVE descriptions and tweets that have been labelled as “informative” or “not informative” relative to cybersecurity events. An active learning approach has been used to prioritize the evaluation of tweets based on their probability of being relevant. The resulting dataset is composed as follows: NIST CVE descriptions: 8,020; labelled tweets: 11,444. Out of the labelled tweets, 9,000 were designated as “not informative” since they were sampled from an unfiltered stream and assumed to be negative cases. The remaining 4,919 labelled tweets were manually classified based on their text content. Regarding multilingualism, the dataset includes contents expressed in various languages, with English accounting for ~80% of the data, mainly as a consequence of including CVE descriptions from the NIST database. This significant imbalance may introduce a performance evaluation bias. The language distribution can be observed in Appendix B. An additional drawback of the dataset is the disparity between informative examples (CVE descriptions) and the random tweets, which are completely unrelated to the topic, making it unsuitable for learning a nuanced concept of informativeness.

<sup>6</sup> <https://github.com/AnastasiaCotov/Improving-Cybersecurity-Awareness-Tweet-Classification.git>



To overcome these limitations, and to be able to exploit the users’ post history, a second dataset has been built. This dataset contains only tweets, labelled with the same principle as the previous dataset. The data belongs to a selected list of 36 Twitter accounts representing institutions, experts and hobbyists. Some of these accounts post exclusively cybersecurity-related tweets, while others share diversified information. This dataset provides an “information context” for selected users, to be able to test to which extent the context given by a user’s post history matters for the classification task. The training set contains 3,023 tweets with 1,386 positive labels and 1,637 negative ones. The test set contains 560 instances, with 201 labelled as “informative” and 359 labelled as “not informative”. This dataset contains tweets from 18 languages. The main languages in Datasets I and II are provided in Appendix B.

The third dataset, which is aimed at a fine-grained categorical classification task, includes vulnerability descriptions extracted from NIST, along with tweets related to additional cybersecurity threats that are not specified in the NIST database, namely “malware”, “spam”, and “ransomware”. These tweets were obtained by performing a keyword search with the *twarc2* tool<sup>7</sup>, where the keywords represent the three classes to be added to the dataset (“spam”, “ransomware”, “malware”). Overall, this dataset comprises 24,184 samples belonging to 12 CVEs, and 6000 samples belonging to the additional cyber-threat types. The label distribution is balanced since each category is represented by approximately 2,000 samples<sup>8</sup>.

#### 4.4 Data quality, pre-processing and content representation

Basic preprocessing is applied to the collected data datasets. Tweets that are exact duplicates are considered only once. Tweets with extremely short text (less than 5 words) are not considered. Text pre-processing is performed taking into account the presence of multiple languages. In order to delete stopwords, the *stopwordsiso* Python library has been utilized, which contains stopwords from more than 50 languages. The entire corpus is converted to lowercase, then deleting all emojis and flag emojis, transport and map symbols, URLs, selected characters<sup>9</sup>, diacritics and Unicode characters. Text is normalized with stemming, which reduces the words to their root or stem form, thus reducing the vocabulary size. SnowballStemmer class provided in the NLTK library<sup>10</sup> supports stemming for multiple languages. It was noted that stemming the words results in higher accuracy of binary classifiers.

Selected and cleaned posts are then converted to sentence embeddings using Language-Agnostic SEntence Representations (LASER) using the official Python

<sup>7</sup> <https://github.com/DocNow/twarc>

<sup>8</sup> “Overflow”, “DoS”, “gain\_privileges”, “Memory\_Corruption”, “bypass”, “CSRF”, “Gain\_info”, “sql\_injection”, “XSS”, “Dir\_traversal”, “Exec\_code”, “File\_inclusion”, “ransomware”, “malware”, “spam”

<sup>9</sup> Like @, #, and other punctuation symbols

<sup>10</sup> <https://www.nltk.org/>

library[3] developed by Meta. In this way, each post is represented by a 1024-dimensional array, providing an implicit representation of the content semantics. This choice has the additional advantage of being applicable to a wide range of languages in a transparent way, with no additional configuration. By design, texts with corresponding semantics in different languages should be mapped to similar vectors in the embedding space.

Regarding user context, an additional vector representing the recent post history of the users is obtained by applying User2Vec [4], when applicable. To effectively capture the context provided by the authors’ post history, it is preferable to select an equal number of tweets per user as a test set. This enables us to test User2Vec’s ability to capture the author’s context in a sensible manner. To validate the hypothesis that accounts that usually post cybersecurity-related content will tend to share similar information in the future, the 20 tweets chosen from each user are selected to represent their most recent activity.

#### 4.5 Keywords for data ingestion

In order to derive a dictionary of keywords suitable for selecting relevant data from social media streams, an algorithm was run on the social media subset of the first dataset. This step is meant to enhance the first step of classification, namely relevance for the cybersecurity domain, while keeping control of the number of items to process downstream. The capability of each keyword in collecting cybersecurity-related posts is quantified by correlating its appearance in the dataset posts with its true relevance label. By eliminating the terms with low correlation values, the keyword list is selected (see Appendix A for the complete list). The resulting keywords can then be utilized at runtime for social media crawling.

#### 4.6 Classifiers

The core of the pipeline consists of two classifiers based on the textual content of the posts: a binary classifier that discriminates between “informative” and “not informative” tweets in the context of cybersecurity, and a second-stage classifier that assigns a fine-grained category label to the tweets marked as relevant by the first classifier. The training pipelines for the two classifiers are summarized in Fig. 2, which also shows the runtime architecture on the right. For the classification of tweets according to their cybersecurity relevance, various configurations of Feed Forward Neural Networks have been explored, comparing the results also with a Logistic Regression classifier built using the *sklearn* module [2].

The binary relevant tweets classifier has been trained with a combination of inputs, in order to assess the gain of using different information representations for the task. First, the LASER embeddings alone were utilized. Then, the same representations were concatenated with contextual information User2Vec embeddings. The training and evaluation of this classifier were performed using the second dataset described in Section

## 4.7 sec:datasets

The second-stage classifier is then applied to tweets that were marked as cybersecurity-related. Its task consists in classifying data into multiple categories that correspond to known vulnerabilities or cyberattack types. This classifier is trained on a combination of data related to the available CVE types and manually labelled tweets. The classifier is based on the third dataset described in Section 4.3.

For both models, the evaluation of the methodology is performed using a k-fold cross-validation setup. To estimate the generalization ability of the models, a further evaluation is conducted on an independent test set. The test set contains unseen data, as it comprises the more recent tweets that were not included in the training or validation data.

# 5 Results

## 5.1 Binary classifiers evaluation

The various configurations of FFNN and the Logistic Regression classifier were applied to the second dataset, considering also the user context in some of the experiments. The results are illustrated in *Table 1*, that reports the results of the experiments conducted together with the network’s layers, so that ‘laser\_bnl2’ corresponds to the model comprising only LASER sentence representation vectors, batch normalization, and L2 regularization method, ‘laser\_bnl1l2’ in addition to the previous ones uses also L1 regularization. The next models contain the LASER and User2Vec concatenated vectors, where ‘laser\_bnl2u2v’ has in its structure batch normalization layers and L2 regularization, the ‘laser\_bnl2u2v’ and ‘laser\_bnl2u2v’ in addition have the embedding vectors scaled. The next three models ‘laser\_bnl2u2v100’, ‘laser\_bnl2u2v200’ and ‘laser\_bnl2u2v1000’ differ in randomly dropped parameters: 100, 200, and 1,000 respectively. Finally, ‘tm\_laser’, ‘hm\_laser’, ‘tm\_laseru2v’, and ‘hm\_laseru2v’ models represent the tuned and hypermodels with LASER embeddings only, and next the combined embeddings with User2Vec.

In general, the models based on LASER achieve higher test accuracy scores. Moreover, it is clear that providing the user context (with User2Vec) improves the model performances. The model ‘laser\_bnl2u2v100’ achieves the best performance metrics overall, with an accuracy of 87%. It shows the ability of identifying relevant and informative tweets about cyber-threats more efficiently compared to its counterparts.

The model ‘tm\_laseru2v’, also performs well with competitive metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and MCC. However, it should be noted that the faster training time of ‘tm\_laseru2v’ is a result of the parameter tuning process rather than an intrinsic efficiency. When aiming at filtering and processing time performance together, the ‘laser\_bnl2u2v100’ variant could also be considered, since its balanced precision, recall, and F1 score, coupled

Model	Acc	Loss	Prec	Recall	F1	AUC	MCC	TT
laser_bnl2	0.85	2.89	0.83	0.84	0.84	0.84	0.67	82.78
laser_l2	0.76	0.72	0.67	0.70	0.66	0.70	0.42	53.12
laser_bnl1l2	0.83	5.69	0.82	0.83	0.82	0.83	0.64	87.27
laser_bnl1	0.83	5.87	0.82	0.83	0.82	0.83	0.65	83.49
laser_bnl2u2v	0.85	2.81	0.83	0.84	0.83	0.84	0.67	84.67
<b>laser_bnl2u2v100</b>	<b>0.87</b>	<b>3.04</b>	<b>0.86</b>	<b>0.87</b>	<b>0.86</b>	<b>0.87</b>	<b>0.73</b>	<b>84.69</b>
laser_bnl2u2v200	0.84	3.01	0.83	0.83	0.83	0.83	0.66	83.39
laser_bnl2u2v1000	0.85	3.55	0.83	0.85	0.84	0.85	0.68	80.51
tm_laser	0.86	0.80	0.85	0.86	0.85	0.86	0.70	10.03
hm_laser	0.84	0.82	0.82	0.83	0.83	0.83	0.66	9.78
<b>tm_laseru2v</b>	<b>0.87</b>	<b>0.81</b>	<b>0.85</b>	<b>0.86</b>	<b>0.85</b>	<b>0.86</b>	<b>0.71</b>	<b>3.71</b>
hm_laseru2v	0.86	0.85	0.85	0.85	0.85	0.85	0.70	3.95
LR_laser	0.83	5.48	0.82	0.83	0.83	0.83	0.65	0.66
LR_laser_u2v	0.84	5.49	0.83	0.84	0.83	0.84	0.66	0.65

**Table 1.** Evaluation metrics of the model deployed on the second dataset

with efficiency. The baseline LR classifier ‘LR\_laser\_u2v’ exhibits a slightly lower accuracy score of 0.84. However, it just requires 0.65 seconds for training, whereas the FFNN training times are roughly two orders of magnitude higher.

The outcomes come out in favor of the reliability of the models in predicting the target variable, highlighting the effectiveness of using LASER input vectors for the task, alone or in combination with User2Vec embeddings.

## 5.2 Categorical classifier evaluation

The categorical classifier was trained on the third dataset described in Section 4.3. The most recognizable cyber-threats/vulnerabilities are ‘spam’ and ‘file inclusion’ with a detection accuracy of 98.17% and 95.45% respectively. However, it is worth noticing that certain categories display lower accuracy, such as ‘buffer overflow’ with 79.14% and ‘bypass’ with 78.27%.

Some examples are illustrated in *Figure 3* to better understand the abilities of the model. In the figure, the first row is categorized as ‘buffer overflow’. However, the classifier jointly decided for membership to the denial of service (DoS) class. If the text is carefully analysed, it can be concluded that the model’s multiple categorization fits the result. Indeed, some incidents could be consequences of multiple root causes, such as various vulnerabilities in the system or a vulnerability that facilitates a particular type of attack. This example illustrates the advantages and expressiveness of the fine-grained approach herein proposed.

## 6 Concluding remarks

In the present paper, we introduced a multi-stage classifier approach aimed at a flexible and accurate identification of social media posts related to cyber-threats and vulnerabilities. The approach is able to classify the posts according

lazarus apt employed linux malware attacks linked 3cx supply chain attack north korea linked apt lazarus employed lir	13	[13]	['malware']
tenda ax12 v22 03 01 21 discovered stack buffer overflow function sub_422ce4 vulnerability attackers denial service s	0	[0, 1]	['Overflow', 'DoS']
edge windows 1703 attacker execute arbitrary code context current user edge handles objects memory aka edge mer	3	[3]	['Memory_Corrupti']
authorization bypass user controlled key github repository ionicabizau parse path prior	4	[4]	['bypass']
memory corruption video buffer overflow parsing asf clips snapdragon auto snapdragon compute snapdragon connec	0	[0, 3]	['Overflow', 'Memo']
naivas supermarket victim cyber attack attackers stole data attack carried ransomware type malicious software encry	12	[12]	['ransomware']

**Fig. 3.** A categorization example

to detailed vulnerability classes, defined using both official and custom sources. The proposed approach is independent from the input language, and it is able to leverage the contextual information related to the author of a post.

However, there are limitations to consider, such as the abundance and the variety of information available on social media, which poses challenges in building a flawless filtering algorithm, also due to the lack of specific open-source datasets. Resource constraints and changes in the social media APIs may also impact the data collection and analysis processes [20]. By acknowledging these limitations and constraints, this study also aims to provide a clear understanding of the research boundaries and ensure the appropriate interpretation and application of the results. Further investigations will focus on automatically analyzing the impact of newly emerging threats on the components of the architecture of specific organizations, in line with the CS-AWARE-NEXT scenario.

**Acknowledgements** This work has been supported by Horizon Europe research Project CS-AWARE-NEXT No. 101069543 and by the PNRR-PE-AI “FAIR” project funded by the NextGeneration EU program.

## References

1. National institute of standards and technology: National vulnerability database, <https://nvd.nist.gov/vuln>
2. Sklearn. <https://scikit-learn.org>
3. Zero-shot transfer across 93 languages: Open-sourcing enhanced laser library, <https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/>
4. Amir, S., Wallace, B., Lyu, H., Carvalho, P., Silva, M.: Modelling context with user embeddings for sarcasm detection in social media. pp. 167–177 (07 2016). <https://doi.org/10.18653/v1/K16-1017>
5. Andriessen, J., Schaberreiter, T., Papanikolaou, A., Röning, J.: Cybersecurity Awareness. Springer, 1 edn. (2022), advances in Information Security
6. Anjum, S., Verma, A., Dang, B., Gurari, D.: Exploring the use of deep learning with crowdsourcing to annotate images. Human Computation **8**(2), 76–106 (2021)
7. Bono, C., Mülâyim, M.O., Cappiello, C., Carman, M.J., Cerquides, J., Fernandez-Marquez, J.L., Mondardini, M.R., Ramalli, E., Pernici, B.: A citizen science approach for analyzing social media with crowdsourcing. IEEE Access **11**, 15329–15347 (2023)

8. Chambers, N., Fry, B., McMasters, J.: Detecting denial-of-service attacks from social media text: Applying nlp to computer security. pp. 1626–1635 (01 2018). <https://doi.org/10.18653/v1/N18-1147>
9. Dionísio, N., Alves, F., Ferreira, P., Bessani, A.: Cyberthreat detection from twitter using deep neural networks (07 2019). <https://doi.org/10.1109/IJCNN.2019.8852475>
10. Fang, Y., Gao, J., Liu, Z., Huang, C.: Detecting cyber threat event from twitter using idcnn and bilstm. *Applied Sciences* **10**, 5922 (08 2020). <https://doi.org/10.3390/app10175922>
11. Hallac, I., Makinist, S., Ay, B., Aydin, G.: user2vec: Social media user representation based on distributed document embeddings. pp. 1–5 (09 2019). <https://doi.org/10.1109/IDAP.2019.8875952>
12. Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J., Sanchez, V.: Social sentiment sensor in twitter for predicting cyber-attacks using  $\ell_1$  regularization. *Sensors* **18**(5) (2018). <https://doi.org/10.3390/s18051380>, <https://www.mdpi.com/1424-8220/18/5/1380>
13. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* **47**(4), 1–38 (2015)
14. Jeong-Ha Park, H.Y.K.: Cyberattack detection model using community detection and text analysis on social media. *ICT Express* **8**(4), 499–506 (2022). <https://doi.org/https://doi.org/10.1016/j.ict.2021.12.003>, <https://www.sciencedirect.com/science/article/pii/S2405959521001685>
15. Le, B.D., Wang, G., Nasim, M., Babar, A.: Gathering cyber threat intelligence from twitter using novelty classification (2019)
16. MITRE: Cve details: the ultimate security vulnerability data source. <https://www.cvedetails.com/>
17. Mittal, S., Das, P., Mulwad, V., Joshi, A., Finin, T.: Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities (08 2016). <https://doi.org/10.1109/ASONAM.2016.7752338>
18. Riebe, T., Wirth, T., Bayer, M., Kuehn, P.D., Kaufhold, M.A., Knauthe, V., Guthe, S., Reuter, C.: CySecAlert: An alert generation system for cyber security events using open source intelligence data. In: *International Conference on Information, Communications and Signal Processing* (2021)
19. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.* **25**(4), 919–931 (2013)
20. TwitterDev: The voice of the TwitterDev team and your official source for updates, news, and events, related to the TwitterAPI (2023), <https://twitter.com/TwitterDev/status/1621026986784337922>
21. Zong, S., Ritter, A., Mueller, G., Wright, E.: Analyzing the perceived severity of cybersecurity threats reported on social media. pp. 1380–1390 (01 2019). <https://doi.org/10.18653/v1/N19-1140>

## Appendix A: Filtered Keyword List, User List

keyword\_dict = [code execution, arbitrary code execution, authenticated attacker, escalation, phishing, buffer overflow, injection, escalation vulnerability, overflow vulnerability exists, injection vulnerability, privilege escalation, stack overflow, execution

vulnerability, ransomware, remote, execution, code injection, heap buffer overflow, remote code, malicious code, unauthenticated remote attacker, privilege vulnerability, sql injection vulnerability, sql injection, remote code execution, injection attack, command injection, malware, buffer overflow vulnerability, exploit heap corruption, overflow vulnerability, integer overflow, patch, execute arbitrary, execute arbitrary code, arbitrary code, exploit heap, vulnerability, code execution vulnerability, code, remote attacker, exploit]

```
user_list = ['Energy_CERT_at', 'ngCERTofficial', 'CSIRTGOB', 'CSIRT_MON',
'ProximusCSIRT', 'BanelcoCSIRT', 'CsirtEPN', 'CSIRTCV', 'csirtutpl', 'CCsirt', 'CSIRT-
SecAdvisor', 'CSIRT_Telconet', 'uuallan', 'gcluley', 'ncsc_no', 'mmorenog', 'Bleepin-
Computer', 'NWU_CSIRT', 'INTERPOL_Cyber', 'DIVDnl', 'CSIRTPanama', 'CSIRT-
Malta', 'csirtmu', 'swisscom_csirt', 'switchcert', 'UCT_CSIRT', 'AusCERT', 'Cert-
LAFD', 'lacnic_csirt', 'NationalCsirtCy', 'tfcirt', 'DSecRU', 'IlyaShabanov', 'rvision_pro',
'irisscert', 'csirt_it']
```

## Appendix B: Dataset I and Dataset II Language Distribution

Dataset	EN	IT	RU	FR	JA	Es	De	Others
Dataset I	17807	1685	396	167	150	136	87	698
Dataset II: training	1314	501	177	117	42	413	92	367
Dataset II: test	253	64	36	17	40	60	0	90

**Table 2.** Language Distribution within the datasets