POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Anastasia Cotov

Advisor: Prof. Barbara Pernici

Co-advisor: Carlo Alberto Bono, Cinzia Cappiello

Academic year: 2022-2023

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, creating threats to commercial companies, organizations, governments, and individuals.
Using social media for information extraction poses several challenges. Many available tools are primarily designed for the English language, potentially missing localized risks and threats. Filtering out irrelevant posts is also a challenge, even when using specialized keywords. Existing tools often overlook user posting behaviour, and a comprehensive approach to address these challenges is lacking. Crowdsourcing has been suggested as a solution, but it may be time-consuming, especially in critical cases where timeliness is crucial [1].
In the cybersecurity domain, the potential in extracting cybersecurity information from social media is 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.
The goal of this study is to propose a general approach to retrieve relevant posts in social media related to cybersecurity issues, focusing on two main contributions: i) selecting relevant posts using specialized keywords and then classifying these posts 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.

2. Related work

Official cybersecurity organizations and communities constantly share news and information about the evolution of cyberspace and related security threats. The system departments use social media, for finding information shared by

other security companies [2].
The identification of cybersecurity-related data on social media has been extensively studied, with a focus on detecting and predicting cyberattacks through social network analysis. Some researches aim to extract suitable information related to cyber threats using NER algorithms and they build neural networks to classify relevant information. The authors in [3] combine convolutional neural networks (CNN) and word embeddings to classify cybersecurity-related posts. Further, they design an LSTM-based named entity recognition (NER) component that highlights meaningful information contained in the posts.
Word embeddings were also used by authors in [4]. They combined similarity on word vector representations for content detection and community detection to identify the most relevant user groups in cyber-attacks.
As a contribution, we propose a two-stage classification method to filter relevant content from social media streams, by selecting informative posts and then applying a fine-grained categorization over known event types. But before that relevant keywords should be selected for social media crawling. These keywords are derived semi-automatically from validated 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 and develop an effective and efficient AI pipeline for analyzing data. In particular, 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.
Currently, the CS-AWARE-NEXT platform provides users with an interface that lists the relevant tweets and associates them with the architectural components to which their content refers. The tweets are extracted from selected specialized accounts that post news about vulnerabilities that have already been certified.
The project aims to improve the system by (i) extracting data monitoring the entire social media streams; (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 not only about certified vulnerabilities.

4. Approach

4.1. A Machine Learning Approach for Tweet Classification

The goal of the proposed approach is the automatic 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[1] 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 able to operate on many languages simultaneously. Since the 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

[1] MITRE. Common vulnerabilities and exposures `https://cve.mitre.org/`

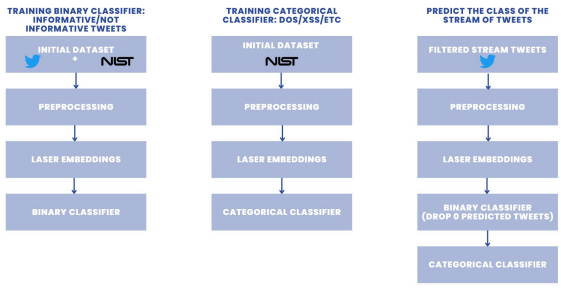

Figure 1: Tweet classification

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.

The overall approach is illustrated in *Figure 1 (on the right)* and consists of the following key steps: 1) Data collection, 2) Pre-processing, 3) Feature extraction using sentence embeddings[2] and representations to capture the user context [5], 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, loss, F1-score, 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 first 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.

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 is the CVE by type dataset available on the CVE Details website[3]. This dataset provides MITRE CVE data categorized into 12 different types and sorted by year. 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 the live data stream, we also collected posts published by users from a selected list containing 36 users 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).

4.3. Datasets

In order to build, train and validate the different proposed functionalities, three tailored datasets have been crafted (available on GitHub[4]). 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.

[2] https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/

[3] https://www.cvedetails.com/vulnerabilities-by-types.php

[4] https://github.com/AnastasiaCotov/Improving-Cybersecurity-Awareness-Tweet-Classification.git

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: 13,919. 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.

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 that belong 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 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[5], where the keywords represent the three classes to be added to the dataset.

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.

4.4. Data quality, pre-processing and content representation

Basic preprocessing is applied to the collected data datasets. Tweets with extremely short text and duplicated ones are not considered. Text pre-processing is performed taking into account the presence of multiple languages. Thus, to delete stopwords, the 50 languages *stopwordsiso* Python library has been utilized. The entire corpus is converted to lowercase and the Unicode function is applied. Then emojis, transport and map symbols, URLs, special characters, diacritics are deleted. Text is normalized with multilingual stemming using SnowballStemmer class provided in the NLTK, library[6]. It was noted that stemming the words results in higher accuracy of binary classifiers.

Selected and cleaned posts are then converted to sentence embeddings by Language-Agnostic SEntence Representations (LASER) using the official Python library developed by Meta. In this way, each post is represented by a 1024-dimensional array. 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 [5], 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 testing User2Vec's ability to capture the author's context sensibly. To validate the hypothesis that accounts that usually post cybersecurity-related content will tend to share similar information in the future, 20 tweets chosen from each user are selected to represent their most recent activity.

4.5. Keywords for data ingestion

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,

[5] https://github.com/DocNow/twarc

[6] https://www.nltk.org/

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. 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: 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. 1, 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 (LR) classifier built using the *sklearn* module.

The binary classifier has been trained with a combination of inputs, 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 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.

For both models, the evaluation of the methodology is performed using a random k-fold cross-validation setup.

5. Results

5.1. Binary classifiers evaluation

The results of the various configurations of FFNN and LR are illustrated in *Table 1.*

The experiments were conducted changing also network's layers, so that 'laser_bnl2' corresponds to the model comprising only LASER sentence representation vectors, batch normalization, and L2 regularization method, while the model 'laser_bnl1' has L1 regularization. The 'laser_bnl1l2' uses both L1 and L2 regularizations. The next models contain the LASER and User2Vec concatenated vectors, where 'laser_bnl2u2v' has in its structure batch normalization layers and L2 regularization. The next model 'laser_bnl2u2v100' is obtained by randomly dropping 100 dimensions from the input vector. Finally, 'tm_laser' and 'tm_laseru2v' models represent the parameter-tuned models with LASER embeddings only, and next the combined embeddings with User2Vec.

Model	Acc	Loss	F1	TT
laser_l2	0.76	0.72	0.66	53.12
laser_bnl1	0.83	5.87	0.82	83.49
laser_bnl2	0.85	2.89	0.84	82.78
laser_bnl1l2	0.83	5.69	0.82	87.27
laser_bnl2u2v	0.85	2.81	0.83	84.67
laser_bnl2u2v100	**0.87**	**3.04**	**0.86**	**84.69**
tm_laser	0.86	0.80	0.85	10.03
tm_laseru2v	0.87	0.81	0.85	3.71
LR_laser	0.83	5.48	0.83	0.66
LR_laser_u2v	0.84	5.49	0.83	0.65

Table 1: Model Evaluation Metrics

In general, the models based on LASER achieve higher test accuracy scores. Moreover, it is clear that providing the user context improves the model's performance. The model 'laser_bnl2u2v100' achieves the best performance metrics overall, with an accuracy of 87%. The model 'tm_laseru2v', also performs well with competitive metrics. 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 The baseline LR classifier 'LR_laser_u2v' exhibits a slightly lower accuracy score of 84%. However, it just requires 0.65 seconds for training, whereas the FFNN training times are roughly two orders of mag-

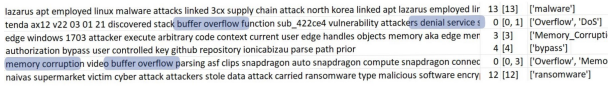

Figure 2: A categorization example

nitude higher.
The outcomes come out in favour 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 categories are 'spam' and 'file inclusion' with a detection accuracy of 98% and 95% respectively. However, it is worth noticing that certain categories display lower accuracy, such as 'buffer overflow' with 79% and 'bypass' with 78%.
Some examples are illustrated in *Figure 2* where the second row is categorized as 'buffer overflow' in the original dataset. However, the classifier jointly decided on membership to the denial of service (DoS) class. If the text is carefully analysed, it can be concluded that the model's multiple categorizations fit 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 research, a multi-stage classifier was introduced. It 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 to detailed vulnerability classes, defined using both official and custom sources. The proposed approach is independent of 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[7]. 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.

References

[1] Carlo Bono, Mehmet Oğuz Mülâyim, Cinzia Cappiello, Mark James Carman, Jesus Cerquides, Jose Luis Fernandez-Marquez, Maria Rosa Mondardini, Edoardo Ramalli, and Barbara Pernici. A citizen science approach for analyzing social media with crowdsourcing. *IEEE Access*, 11:15329–15347, 2023.

[2] Jerry Andriessen, Thomas Schaberreiter, Alexandros Papanikolaou, and Juha Röning. *Cybersecurity Awareness*. Springer, 1 edition, 2022. ISBN 978-3-031-04226-3. Advances in Information Security.

[3] Nuno Dionísio, Fernando Alves, Pedro Ferreira, and Alysson Bessani. Cyberthreat detection from twitter using deep neural networks. 07 2019. doi: 10.1109/IJCNN.2019.8852475.

[4] Hyuk-Yoon Kwon Jeong-Ha Park. Cyberattack detection model using community detection and text analysis on social media. *ICT Express*, 8(4):499–506, 2022. ISSN 2405-9595. doi: https://doi.org/10.1016/j.icte.2021.12.003.

[5] Ibrahim Hallac, Semiha Makinist, Betul Ay, and Galip Aydin. user2vec: Social media user representation based on distributed document embeddings. pages 1–5, 09 2019. doi: 10.1109/IDAP.2019.8875952.

[7] `https://twitter.com/TwitterDev/status/1621026986784337922`