U-BERTopic: An urgency-aware BERT-Topic modeling approach for detecting cyberSecurity issues via social media

Majed Albarrak and **Gabriele Pergola**
University of Warwick , England

Arshad Jhumka
University of Leeds, England

Abstract

For computer systems to remain secure, timely information about system vulnerabilities and security threats are vital. Such information can be garnered from various sources, most notably from social media platforms. However, such information may often lack context and structure and, more importantly, are often unlabelled. For such media to act as alert systems, it is important to be able to first distinguish among the topics being discussed. Subsequently, identifying the nature of the threat or vulnerability is of importance as this will influence the remedial actions to be taken, e.g., is the threat imminent?. In this paper, we propose U-BERTopic, an urgency-aware BERT-topic modelling approach for detecting cybersecurity issues through social media, by integrating sentiment analysis with contextualized topic modelling like BERTopic. We compare U-BERTopic against three other topic modelling techniques using four different evaluation metrics for topic modelling and cybersecurity classification by running on a 2018 cybersecurity-related Twitter dataset. Our results show that (i) for topic modelling and under certain settings (e.g., number of topics), U-BERTopic often outperforms all other topic modelling techniques and (ii) for attack classification, U-BERTopic performs better for some attacks such as vulnerability identification in some settings.

1 Introduction

There has been a noticeable increase in the number of cyberattacks as well as in the severity of their consequences. The UK Department for Science, Innovation and Technology's survey shows that one-tenth of companies and nonprofit organizations fell victim to cybercrime in one year (2023), marking a 29% increase from the previous year (Johns and Ell, 2023). The financial impact of cyberattacks has also increased dramatically according to Ponemon Institute and IBM Security's report (Institute, 2023), with the average cost of a data breach in 2023reaching USD 4.45 million, the highest level ever, representing a 2.3% increase compared to the previous year.

Tweet	Urgency Level
@MsftSecIntel: New blog post: Microsoft researchers analyzed Zerobot 1.1, the latest version of the Go-based DDoS botnet that spreads primarily through IoT and web application vulnerabilities. This version expands the malware's reach to different types of devices	Urgent (DDoS)
@troyhunt: He's back! But **unable** to choose a secure password That must be... **frustrating**	Normal (Negative)
@Unit42_Intel: We're seeing vulnerability scanning and active exploitation attempts for CVE-2022-1388. Within 10 hours, our Threat Prevention signature triggered 2,552 times. Read for more details and recommended mitigation actions.	Urgent (Zero-Day Attack)
@SCMagazine: Identity authentication **failure** can **cost** financial firms as much as $42 million	Normal (Negative)

Table 1: Examples of cybersecurity-related tweets, some conveying urgency while others are informational.

To protect IT infrastructure from cyberattacks, it is important for security engineers to obtain timely information about system vulnerabilities and threats. Social media is proving to be an important outlet where these issues are discussed. However, such information are often unstructured, may lack context and, very often, unlabelled. Table 1 shows some examples of tweets that are security-related. However, while the last tweet seems only informational, the third tweet, on the other hand, appears to carry more information about security incidents (e.g., active exploitation).

1.1 Urgency-aware modelling of cyberSecurity issues

For social media to act as a cybersecurity alert system, it is crucial that relevant security issues such as threats and vulnerabilities are accurately identi-

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 196–211
July 29–30, 2024.

fied. Further, security issues that are important are often captured on social media posts as those that carry some sense of urgency. For example, the third tweet in Table 1 captures urgency through *"within 10 hours ... 2,552 times"*. To further understand the urgent issue, topics need to be extracted accurately to enable identification of the relevant problem and also to enable adequate handling of these security problems. To this end, we propose U-BERTopic, an urgency-aware BERT-topic modelling technique. U-BERTopic extends BERTopic by adapting C-FT-IDF to include a notion of urgency.

Two main problems exist: (i) topic identification and (ii) cybersecurity issue detection. We evaluate the performance of U-BERTopic on a 2018 security-related Twitter dataset and also compare against three other topic modelling techniques using four different but complementary metrics. Our results show that (i) often, U-BERTopic ourperforms other topic models and sometimes is the only model that detects a given security issue and (ii) when classification is performed on tweets, U-BERTopic achieves best performance for certain attack classes under specific settings such as topic number.

The paper is structured as follows: Section 2 discusses related work. We introduce U-BERTopic in Section 3. Section 4 details the evaluation performed and Section 5 explains a case study. Limitations are discussed in Section 6 and we conclude the paper in Section 7.

2 Related work

2.1 Deep learning for attack detection

Behzadan et al. (2018) construct a dataset of recent vulnerabilities tweets and conduct binary and multiclass deep learning classification on that dataset. They collect the data using a customising stream listening tool of Tweepy (Roesslein, 2009), and then they manually label the tweets. Behzadan et al. (2018) use CNN layers to apply binary and multiclass classification at the same time. Using the same X dataset of the previous paper, The work of Dionísio et al. (2019) shows how multilayer classification architecture can improve the performance of the model. They build a CNN classification model with an LSTM extraction layer to achieve better results. The work has high F1 score results, and they restrict their dataset to have only a set of cybersecurity accounts rather than including keywords or hashtags.

LSTM and CNN are used in Fang et al. (2020)'s work to classify cyberthreat events on X (Twitter). They collect related tweets over a period of 18 months and then process the data with LDA and word embedding to make the data ready for the deep learning layer. The results are both Name Entity Recognition (NER) and a threat event classification. Simran et al. (2019)'s paper enhances the work of Behzadan et al. (2018) by adding the Gated recurrent unit (GRU) layer in the CNN model. They study and compare 20 models including classical, deep learning and NLP techniques, and conclude that GRU with CNN model shows the best performance. Tekin and Yilmaz (2021) propose a two-layer of BILSTM and train them on Behzadan et al. (2018) dataset. The proposed paper mitigates the overfitting issue by adding drop-out layers to the architecture. Pre-processing tweets in Tekin and Yilmaz (2021) includes converting the characters, removing HTML and URL links, and removing new lines.

Bayer et al. (2022)'s work proposes a multi-level classifier that focuses on only one incident with its related events. They collected tweets about the Microsoft Exchange Server incident that occurred in 2021 and then combined three techniques to build their classifier levels. They fine-tune the multilevel pre-training model, BERT, adding generated instances by data augmentation and applying prompt tuning learning in the last layer. The idea is to enhance the adaption of new cyber threats or cybersecurity content by dedicating a classifier for each case.

TI-Prompt, by You et al. (2022), is a threat intelligence few-shots classification on Twitter. They use prompt-tuning on a Bert-based pre-trained language model to construct prompt templates, and then perform binary and multiclass classification using verbalizer refinement and enrichment to better map the predicted words. The results of this recent research outperform the work of Behzadan et al. (2018) and Dionísio et al. (2019) which highlight the significance of prompt engineering in classification tasks. However, manual verbalizers and prompts need human intervention and may affect the performance when changing the dataset (Zhou et al., 2023). Furthermore, discussions related to certain attacks, such as Zero-Day Attacks, do not rely on fixed terms or keywords due to the nature of zero-day vulnerabilities, which are previously unseen. Therefore, the supervised learning models in existing works show that they still need to en-

hance their generalisation ability to perform well on new, unseen cybersecurity events in social media without human intervention and labelling.

2.2 Topic modelling and sentiment analysis for attacks clustering

Shu et al. (2018)'s work proposes temporal sentiment analysis on Twitter to cluster the events and predict future cybersecurity attacks. They use NLP techniques such as n-gram and TF-IDF to include the word sequences and the importance of terms in the clustering and classification tasks. Logistic regression is used for the machine learning-based sentiment analysis task, and then the k-means algorithm is applied to the unsupervised clustering task with regard to mean sentiment scores for each subject. Gupta et al. (2016) conducted a cybersecurity lexicon-based sentiment analysis on Twitter in two different periods to show the changes of the emotions and reactions in the cybersecurity events. They apply IBM Watson's Insights model in the research.
Furthermore, Deb et al. (2018) extract cybersecurity-related dark web content and use VADER, Linguistic Inquiry and Word Count and SentiStrength sentiment approaches (Hutto and Gilbert, 2014) to predict future cyberattack events.
Adams et al. (2018) conduct an unsupervised LDA topic modelling on CAPEC dataset to cluster the patterns. The model is used to extract the pattern topics from the cyberattack description to understand the nature of the attack and to better assess the risk.
Wang et al. (2023) propose TDM contextualized topic modelling to predict cyberattacks. They conduct a comparison study between some topic modelling approaches such as LDA, NMF, and Neural Topic modelling. They found that TDM outperformed the others, and showed better semantic clustering. Their TDM model's architecture contains the Combined Topic Model,CTM, of Bianchi et al. (2020) which uses an autoencoder and pre-trained representations. CTM uses the variational autoEncoder ProdLDA of Terragni et al. (2021) with SBERT embedding representations of Reimers and Gurevych (2019), but Wang et al. (2023) use CyBert pre-trained representations instead to have more cybersecurity focus. However, the review shows a gap in understanding criticality and urgent sentiments in cybersecurity context.These meanings are essential for the early prediction of Zero-Day Attacks. Table 2 shows the literature works and their algorithms and techniques.

3 U-BERTopic model

In the following paragraph, we introduce in detail U-BERTopic, which extends traditional topic modelling to focus tones of urgency and necessity characterising cybersecurity issues. First, proposing uC-TF-IDF which is cybersecurity focused of BerTopic (Grootendorst, 2022)'s c-TF-IDF to include the sentiment, urgent scores of the text. Furthermore, we apply Cybert (Ranade et al., 2021) which is a cybersecurity LLM model.

3.1 BERTopic topic model

Grootendorst (2022) introduced BERTopic, a topic modelling approach based on BERT embeddings and a class-based TF-IDF to create dense clusters allowing for interpretable topics.It consists of four main steps. First, it converts the documents (tweets or posts in this context) into embeddings, via Sentence BERT, a BERT-based optimised model for sentence-level embeddings (Reimers and Gurevych, 2019). Then, the high-dimensional sentence embeddings are reduced to lower dimensions via UMAP (McInnes et al., 2018), a techniques for dimensionality reduction. After the embeddings has been reduced, a clustering algorithm, like HDBSCAN (Campello et al., 2013), is applied to cluster similar documents together. For each cluster of documents, a class-based TF-IDF (c-TF-IDF) is then calculated to find representative words for each topics, whose most representative terms for each cluster constitute the final topics.

3.2 uC-TF-IDF algorithm

We propose the urgency-class-based TF-IDF (uC-TF-IDF , Algorithm 1), which is an advancement of the BERTopic's c-TF-IDF. While the traditional c-TF-IDF treats the terms uniformly across all contexts, the urgency-class-based TF-IDF is designed to incorporate sentiment analysis into the term weighting process. Unlike BERTopics's c-TF-IDF, which calculates term frequencies based solely on their occurrences within clusters, uC-TF-IDF adjust these frequencies based on the sentiment conveyed in the texts. This new design allows uC-TF-IDF to dynamically prioritise terms that are not only frequent but also relevant in expressing the urgency and significance of topics, particularly

Work	UL	TB	NN	TM	SA	Model/Algorithm
Gupta et al. (2016)					✓	lexicon-based
Adams et al. (2018)	✓			✓		LDA
Deb et al. (2018)	✓				✓	logistic regression and k-means
Behzadan et al. (2018)			✓			CNN
Dionísio et al. (2019)			✓			LSTM, BiLSTM, and NER
Simran et al. (2019)			✓			GRU, CNN-GRU
Liu et al. (2020)	✓			✓		NMF, Jaccard similarity
Fang et al. (2020)			✓	✓		LDA and BiLSTM, NER
Huang and Ban (2020)			✓	✓		LSTM, Random Forest, LDA
Tekin and Yilmaz (2021)			✓			BiLSTM
Bayer et al. (2023)		✓	✓			GPT-3, human-in-the-loop filtering
You et al. (2022)		✓	✓			BERT, few-shots
Wang et al. (2023)	✓	✓	✓	✓		CTM, CyBert
U-BERTopic (Our)	✓	✓	✓	✓	✓	Urgency Extraction, BertTopic

Table 2: Comparison of U-BERTopic with existing NLP-based cyberattack detection works in X (derived from (Wang et al., 2023)). Abbreviations: UL: Unsupervised Learning; TB: Transformer-based; NN: Neural Networks; TM: Topic Modelling; SA: Sentiment Analysis.

beneficial in the domain of cybersecurity, where sentiment and immediacy can influence the interpretation of topics and consequent actions.

We describe the structure of the Post-Term Matrix and explain our method for integrating updated sentiment scores into the sentiment lexicon. Subsequently, we delineate our approach for adjusting term frequencies based on sentiment, and conclude with a description of how these frequencies are aggregated into class-based term frequencies and adapted into the new uC-TF-IDF formula (Algorithm 1). These steps aim to refine the detection and representation of critical topics discussed in social media posts.

Post-Term Matrix Given a set of social media posts, we define the posts set $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, where each p_i represents an individual post. The set of unique terms extracted from all posts is denoted as $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$. We construct the Post-Term Matrix $\mathbf{X}$ of dimensions $n \times m$, where each element x_{ij} quantifies the occurrence of term t_j in post p_i.

Document Sentiment Score To compute the sentiment score of posts, we first update the sentiment lexicon to tailor it for highlighting cybersecurity urgencies and threats. Once obtained this document sentiment score $S(p_i)$, this is subsequently used to adjust the weight of the uC-TF-IDF matrix as shown in Algorithm 1.

Updating cyberattack terms sentiment score in the sentiment lexicon Let $K = \{k_1, k_2, \ldots, k_n\}$ be the set of cyberattack keywords, and $V = \{v_1, v_2, \ldots, v_n\}$ are the corresponding new sentiment score. The sentiment lexicon is updated by the given set of pairs and then used to compute the sentiment score $S(p)$ of a post p. Aligned with previous works (Satyapanich et al., 2020; Trong et al., 2020), the Keywords Set $\mathcal{K} = \{k_1, k_2, \ldots, k_n\}$ is defined to include terms such as "exploit," "attack," and "zero-day" etc, with a high negative polarity.

To effectively identify and prioritize urgent cybersecurity threats from social media content, we enhance our term frequency adjustments and document analysis processes within the uC-TF-IDF framework. Given the urgency with attacks, a predetermined score of -5 is assigned to security keywords in the lexicon.

Sentiment Analysis We utilize the VADER sentiment analysis tool (Hutto and Gilbert, 2014), particularly its compound score, to compute the sentiment $S(p_i)$ of a post, by leveraging the updated lexicon. Each post p_i is associated with a compound sentiment score $S(p_i)$ from VADER, which reflects the overall sentiment ranging from -1 (most negative) to 1 (most positive).

Let $S(p_i)$ be the sentiment polarity score of post(tweet) p_i, where $S(p_i) \in [-1, 1]$. A post is considered to have negative sentiment if $S(p_i)$(compound sentiment score) < 0.

Algorithm 1 uC-TF-IDF Algorithm

Require: Set of posts P, Set of unique terms *Term*, Set of cyberattack keywords K with scores C
Ensure: Adjusted c-TF-IDF matrix for cyberattack keywords

Construct the Post-Term Matrix X for P using *Term*
Update the sentiment lexicon L with cyberattack keywords K and scores C
for each post p_i in P **do**
    Compute sentiment score $S(p_i)$ for post p_i
    if $S(p_i) < 0$ **then**
        for each term t_j in *Term* **do**
            $u_{TF}(t_j, p_i) \leftarrow 2 \cdot x_{ij}$
        end for
    else
        $u_{TF}(t_j, p_i) \leftarrow x_{ij}$
    end if
end for
for each class C corresponding to a topic cluster **do**
    for each term t_j in *Term* **do**
        uC-TF$(t_j, C) \leftarrow \sum_{p_i \in C} u_{TF}(t_j, p_i)$
    end for
    Compute $IDF(t_j, P)$ for term t_j
    uC-TF-IDF$(t_j, C, P) \leftarrow$ uC-TF$(t_j, C) \times$ IDF(t_j, P)
end for
return the matrix of uC-TF-IDF values for each term and class

3.3 Term frequency and document analysis

Adjusted term frequency $u_{TF}(t_j, p_i)$ for term t_j in document p_i is thus calculated as follows:

$$u_{TF}(t_j, p_i) = \begin{cases} 2 \times x_{ij} & \text{if } S(p_i) < 0, \\ x_{ij} & \text{otherwise.} \end{cases} \quad (1)$$

where $S(p_i)$ is the sentiment score derived from VADER's compound score.

Subsequently, for each class C of posts, representing a cluster of thematically similar content, the class-based term frequency $uC_{TF}(t_j, C)$ sums the adjusted frequencies across all documents:

$$uC_{TF}(t_j, C) = \sum_{p_i \in C} u_{TF}(t_j, p_i), \quad (2)$$

thus, creating a robust metric that encapsulates both the frequency of terms and their urgency (Algorithm 1).

Dataset Labeling (Dionísio et al., 2019)
Cybersecurity-related
- True
- False
Cyberattack Type
- Leak (Selected)
- DDoS (Selected)
- General
- Vulnerability (Selected)
- Ransomware (Selected)
- Botnet (Selected)
- 0-day attack (Selected)

Table 3: Dataset labeling overview

We extend this concept to compute uC-TF-IDF, which enhances the identification of critical discussions by integrating the inverse document frequency $IDF(t_j, P)$ for term t_j across all posts P:

$$\text{uC-TF-IDF}(t_j, C, P) = uC_{TF}(t_j, C) \times IDF(t_j, P). \quad (3)$$

with t_j being the particular term considered, C the class of documents, and P the set of all posts. This calculation aims to balance term commonality against their significance within specific classes while considering the cybersecurity relevance.

4 Evaluation

4.1 Datasets

We conduct a thorough experimental assessment using two distinct datasets. The first is the publicly available[1] cybersecurity dataset introduced by Behzadan et al. (2018), which comprises tweets collected in 2018. It includes tweets categorized into two classes: one class indicating if the tweet is related to cybersecurity, and the second class identifying the specific type of cyberattack discussed, such as *zero-day attacks*, *ransomware*, *DDoS*, *leaks*, or *botnets*. Table 3 illustrates the original labels by Behzadan et al. (2018), and the selected labels for the classification task.

Data collection. Additionally, we compiled a dataset from several well-known cyberthreat intelligence sources, including Microsoft Cyberthreat Intelligence (*@MsftSecIntel*), Cybersecurity and Infrastructure Security Agency (*@CISAgov*), and The Hackers News (*@TheHackersNews*), spanning

[1] https://github.com/behzadanksu/cybertweets

NPMI				
Model	**K = 20**	**50**	**100**	**150**
LDA	0.06	0.01	0.02	-0.05
CTM	0.08	0.07	0.11	0.12
BERTopic	**0.23**	**0.21**	**0.22**	**0.22**
U-BERTopic	0.22	**0.21**	**0.22**	0.21
Topic Coherence (CV)				
Model	**K = 20**	**50**	**100**	**150**
LDA	0.49	0.47	0.47	0.42
CTM	0.58	0.58	0.61	0.60
BERTopic	**0.65**	**0.61**	0.62	**0.63**
U-BERTopic	0.62	**0.61**	**0.63**	0.62
Topic Diversity				
Model	**K = 20**	**50**	**100**	**150**
LDA	0.56	0.55	0.56	0.59
CTM	0.86	0.79	0.50	0.36
BERTopic	0.85	0.86	0.87	0.83
U-BERTopic	**0.87**	**0.87**	**0.88**	**0.84**
Topic Quality				
Model	**K = 20**	**50**	**100**	**150**
LDA	0.27	0.26	0.26	0.25
CTM	0.50	0.46	0.31	0.22
BERTopic	**0.56**	**0.53**	0.53	**0.52**
U-BERTopic	0.54	**0.53**	**0.55**	**0.52**

Table 4: NPMI, Topic Coherence, Topic Diversity, and Topic Quality scores for Cybersecurity Dataset 2018 for the four models: LDA, CTM, BERTopic and U-BERTopic, (Number of Topics: $k = 20$ to $k = 150$).

Zero-day Attack				
	k=20	k=50	k=100	k=150
LDA	**0.98**	0.98	0.98	0.98
CTM	**0.98**	0.98	**0.99**	0.98
BERTopic	0.97	0.98	0.97	0.98
U-BERTopic	0.97	0.98	0.97	0.98
Botnet Attack				
	k=20	k=50	k=100	k=150
LDA	**0.96**	0.95	0.96	0.96
CTM	0.95	**0.97**	**0.97**	**0.97**
BERTopic	0.95	0.96	**0.97**	0.96
U-BERTopic	0.95	0.96	**0.97**	0.96
DDoS Attack				
	k=20	k=50	k=100	k=150
LDA	0.86	0.88	0.89	0.89
CTM	0.89	0.92	**0.93**	0.9
BERTopic	**0.9**	0.92	0.92	0.92
U-BERTopic	0.89	0.92	**0.93**	**0.93**
Leak Attack				
	k=20	k=50	k=100	k=150
LDA	0.99	0.99	0.99	0.99
CTM	0.99	0.99	0.99	0.99
BERTopic	0.99	0.99	0.99	0.99
U-BERTopic	0.99	0.99	0.99	0.99
Ransomware Attack				
	k=20	k=50	k=100	k=150
LDA	0.82	0.85	0.86	0.87
CTM	**0.9**	**0.91**	**0.93**	**0.92**
BERTopic	0.87	0.88	0.88	0.89
U-BERTopic	0.88	0.88	0.86	0.88
Vulnerability Attack				
	k=20	k=50	k=100	k=150
LDA	0.69	0.72	0.79	0.79
CTM	0.84	**0.87**	**0.88**	**0.88**
BERTopic	0.84	**0.87**	0.85	**0.88**
U-BERTopic	**0.85**	**0.87**	0.87	0.86

Table 5: Accuracy Scores by CyberAttack, Model and Number of Topics on Cybersecurity 2018 Dataset (k=20 to k=150)

from Jan. 1, 2021, to Dec. 30, 2022 [2]. The collected dataset comprises 112332 tweets (documents), and was curated to exclude retweets and advertisements.

4.2 Topic quality

U-BERTopic is evaluated and compared against several baselines by assessing the (i) intrinsic quality of the generated topics, and the (ii) classification accuracy based on them. In the evaluation of the topic quality, the proposed solution, along with three other topic modeling algorithms, i.e, LDA (Blei et al., 2003), BERTopic (Grootendorst, 2022), and the Contextualized Topic Model (CTM) (Bianchi et al., 2021), are assessed using four different metrics widely used in the literature: the Normalized Pointwise Mutual Information (NPMI), topic coherence (CV), topic diversity, and topic quality. The coherence metrics measure the quality and interpretability of the output topics, based on their human interpretability.

NPMI. The NPMI evaluates models by measuring the frequency with which topic words co-occur in the same documents (Bouma, 2009). Its normalise results range from -1 to 1, where 1 indicates perfect coherence between the words in a topic.

Topic Coherence (CV). Topic coherence (CV) measures the interpretability of the topics by assessing the semantic similarity between high-scoring words in the topics, based on an external corpus, such as Wikipedia (Röder et al., 2015). Higher CV values indicate better coherence of topics.

Topic Diversity. Topic diversity measures the extent to which resulting topics are distinct from one another, which is crucial in as neural topic mod-

[2]The code is publicly available: https://github.com/AICybersecurity2/UBERTopic/

els tend to suffer a lack of regularisation over the topic diversity, which is crucial in a specialised domain, such as cybersecurity, where the desired topics must be able to differentiate among specific cyberattack discussions.

Topic Quality. Topic quality is a derived metric from the product of topic diversity and topic coherence, and offers insights into how well a model balances the diversity of topics and their interpretability (Dieng et al., 2020; Wang et al., 2023; Zhang et al., 2023)[3].

Table 4 presents the results of a comparison among U-BERTopic and three other topic modeling algorithms, namely LDA, CTM and BERTopic across four evaluation metrics, averaged over five runs. The experiments span a range of topic numbers K from 20 to 150. The results demonstrate that U-BERTopic and BERTopic outperform the other two models across all metrics. Notably, these models maintain significantly higher scores, particularly in terms of diversity, and exhibit consistent performance as K increases. In contrast, CTM shows a dramatic drop in topic diversity values after K=60. A higher diversity indicates that U-BERTopic can generate a wider range of cybersecurity topics without sacrificing coherence. More detailed data about the comparison can be found in Tables A1 to A4 in the Appendix, as well as in Figures A1 to A4.

4.3 Topic modeling classification

A topic modeling classification was conducted on the cybersecurity dataset to further evaluate the proposed approach by detecting six types of cybersecurity events. These events represent the most urgent cybersecurity-related tweets, describing ongoing cyberattacks or warnings of potential threats. The categories are Zero-day, Botnet, Leak, Ransomware, DDoS, and Vulnerabilities. The evaluation involved applying topic modeling and classification to each category separately, and it utilises the OCTIS package (Terragni et al., 2021) for the classification process. The models were tuned based on the number of topics (k=20 to k=150), and classification accuracy scores were recorded for each category and each topic modeling algorithm. While the impact of the number of topics seems limited on the classification task compared to the impact on their intrinsic quality, such as diversity and coherence, we notice that the accuracy scores for LDA improved when the number of topics increases, particularly in the DDoS and Vulnerabilities categories. Overall, the results, as shown in Table A5, demonstrate high accuracy for most categories, though the Ransomware and Vulnerabilities categories exhibited the lowest accuracy scores across all models. While some classes, such as the 'Leak' attack class, consistently achieve high accuracy across all topic modeling algorithms due to data quality and limited instances, the proposed U-BERTopic model notably performs better in the Vulnerabilities category, achieving the highest accuracy score of 0.89. This suggests that U-BERTopic has enhanced capabilities for understanding cybersecurity events that entail particular concerning sentiments, such as vulnerabilities. Table 10 and Figures A5-A11 provide more detailed results of the classification experiments.

5 Instance examination

To further examine U-BERTopic's ability to capture the urgency level of cybersecurity discussion and news, a significant series of cyberattacks with a high impact was selected for a case study, in this case the Microsoft Exchange server attacks in 2021 (CISA, 2021) and 2022 (CISA, 2022). Our study evaluates whether the generated topics contain terms uniquely associated with these cyberattacks that will suggest better model performance in detecting urgency. The timeline of these attacks is as shown in Figure 1 (they occurred between 2021 and 2022). The 2022 dataset comprises data collected from January 2021 to December 2022 from Cyberthreat Intelligence X (formerly Twitter) accounts to cover the case. After data cleansing, all tweets about the Microsoft Exchange server were aggregated using keywords such as "Microsoft Exchange," "Outlook," "ProxyShell," "ProxyNotShell," and "MS Exchange." After that, all four topic modelling algorithms are applied and the results are shown in Table A6 in the Appendix. In this table , U-BERTopic can extract more MS Exchange Server ZDAs terms (ProxyShell, ProxynotShell, dearcry, and ProxyLogon). These urgency-related keywords were unseen before the event and they are either vulnerability or malware names associated with the attack.

6 Limitations and discussion

Limitations The proposed U-BERTopic model combines contextualized topic modeling with sentiment analysis to improve the system's ability to

[3] *The Optimizing and Comparing Topic Models Is Simple* (OCTIS) package (Terragni et al., 2021) is employed for the topic modelling evaluation.

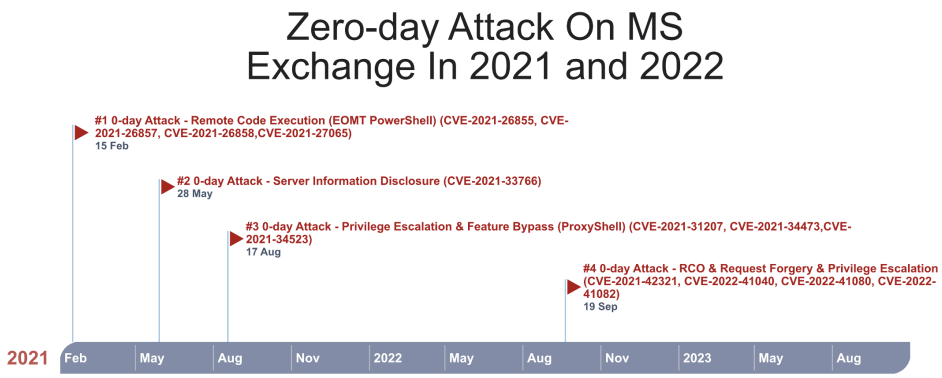

Figure 1: Cyberattacks timeline in Microsoft Exchange Server case study

learn the urgency level of cybersecurity issues. U-BERTopic employs lexicon-based sentiment analysis but, to accurately capture urgency within security-related content demands more sophisticated approaches, such as machine learning-based sentiment analysis. Existing datasets also lack urgency labels for better model training. Additionally, while the cybersecurity keywords and scores used to update the sentiment analysis lexicon are currently collected and estimated manually, cybersecurity events have a variety of terms and trends that need to be taken into account.

Discussion The evaluation and case study demonstrate significant potential for predicting ongoing cyberattacks. Utilizing domain-specific LLM-based topic modeling provides a more advanced tool for cybersecurity threat intelligence teams to improve their detection capabilities. While all topic modeling algorithms are capable of performing classification and event detection tasks, U-BERTopic investigates the sentiment nuances behind the content to enhance detection effectiveness. Furthermore, the positive results from the classification evaluation are promising, encouraging the development of more specialized datasets for urgency-aware cyberattack analysis.

The topic modeling metrics used in this study (NPMI, Diversity, Coherence, and Quality) assess the quality of the models' outputs from various perspectives. U-BERTopic yields more favorable results in topic diversity and topic quality. Although NPMI and Coherence (CV) results indicate that BERTopic has the highest scores, U-BERTopic still maintains high and competitive scores compared to BERTopic and significantly outperforms the other two models (CTM and LDA). This indicates that our enhancements to BERTopic do not compromise topic coherence while improving diversity. The topic modeling accuracy results show high scores for all models, including U-BERTopic. Evaluating the four models across various cyberattack categories reveals the degree to which each model understands discussions related to that category.

The selected case study (the cyberattack event: MS Exchange Server Zero-day attack) prompted extensive social discussions within cybersecurity communities, introducing many terms specific to this unfortunate event. U-BERTopic extracted more of these terms than others, which shows its superiority in capturing the nuances of urgency.

7 Conclusion and future work

In this paper, we have introduced U-BERTopic, an urgency-aware topic modelling designed to detect cyberattacks and enhance the CTI discovery process. U-BERTopic leverages probabilistic and neural NLP models, such as transformer-based word architectures and topic models for fine-grained detection of cybersecurity topics and sentiments. By integrating sentiment analysis with contextualized topic modelling like BERTopic, we spotlight the topics most representative of ongoing cyberattacks and urgent events. Our newly developed method, uC-TF-IDF, is tailored to extract requirements that are particularly relevant to urgent cybersecurity events. Comprehensive evaluations of topic modelling have been conducted, showing the improved ability of U-BERTopic in detecting sentiment-critical cybersecurity topics. Future work will expand upon this foundation by further integrating urgency in sentiment analysis into the topic modeling approach and comparing the performance with different large language models (LLMs).

References

Stephen Adams, Bryan Carter, Cody Fleming, and Peter A Beling. 2018. Selecting system specific cybersecurity attack patterns using topic modeling. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 490–497. IEEE.

Markus Bayer, Tobias Frey, and Christian Reuter. 2022. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *arXiv preprint arXiv:2207.11076*.

Markus Bayer, Tobias Frey, and Christian Reuter. 2023. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *Computers & Security*, 134:103430.

Vahid Behzadan, Carlos Aguirre, Avishek Bose, and William Hsu. 2018. Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5002–5007. IEEE.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

CISA. 2021. Urgent: Protect against active exploitation of proxyshell vulnerabilities. Technical report, the US Cybersecurity and Infrastructure Security Agency (CISA), USA, Washington DC.

CISA. 2022. Microsoft releases guidance on zero-day vulnerabilities in microsoft exchange server. Technical report, the US Cybersecurity and Infrastructure Security Agency (CISA), USA, Washington DC.

Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting cyber-events by leveraging hacker sentiment. *Information*, 9(11):280.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Nuno Dionísio, Fernando Alves, Pedro M Ferreira, and Alysson Bessani. 2019. Cyberthreat detection from twitter using deep neural networks. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Yong Fang, Jian Gao, Zhonglin Liu, and Cheng Huang. 2020. Detecting cyber threat event from twitter using idcnn and bilstm. *Applied Sciences*, 10(17):5922.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Babita Gupta, Shwadhin Sharma, and Anitha Chennamaneni. 2016. Twitter sentiment analysis: An examination of cybersecurity attitudes and behavior. *PROCEEDINGS OF THE 2016 PRE-ICIS SIGDSA/IFIP WG8.3 SYMPOSIUM: INNOVATIONS IN DATA ANALYTICS*.

Shin-Ying Huang and Tao Ban. 2020. Monitoring social media for vulnerability-threat prediction and topic analysis. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1771–1776. IEEE.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

IBM Security Ponemon Institute. 2023. Cost of a data breach report 2022.

Emma Johns and Maddy Ell. 2023. Cyber security breaches survey 2023.

Xiuwen Liu, Jianming Fu, and Yanjiao Chen. 2020. Event evolution model for cybersecurity event mining in tweet streams. *Information Sciences*, 524:254–276.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Priyanka Ranade, Aritran Piplai, Anupam Joshi, and Tim Finin. 2021. Cybert: Contextualized embeddings for the cybersecurity domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3334–3342. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Joshua Roesslein. 2009. tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5:724.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8749–8757.

Kai Shu, Amy Sliva, Justin Sampson, and Huan Liu. 2018. Understanding cyber attack behaviors with sentiment information on social media. In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, pages 377–388. Springer.

K Simran, Prathiksha Balakrishna, R Vinayakumar, and KP Soman. 2019. Deep learning approach for enhanced cyber threat indicators in twitter stream. In *International Symposium on Security in Computing and Communication*, pages 135–145. Springer.

Uğur Tekin and Ercan Nurcan Yilmaz. 2021. Obtaining cyber threat intelligence data from twitter with deep learning methods. In *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 82–86. IEEE.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

Hieu Man Duc Trong, Duc-Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5381–5390.

Yue Wang, Md Abul Bashar, Mahinthan Chandramohan, and Richi Nayak. 2023. Exploring topic models to discern cyber threats on twitter: A case study on log4shell. *Available at SSRN 4404537*.

Yizhe You, Zhengwei Jiang, Kai Zhang, Jun Jiang, Xuren Wang, Zheyu Zhang, Shirui Wang, and Huamin Feng. 2022. Ti-prompt: Towards a prompt tuning method for few-shot threat intelligence twitter classification. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 272–279. IEEE.

Duoyi Zhang, Yue Wang, Md Abul Bashar, and Richi Nayak. 2023. Enhanced topic modeling with multimodal representation learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 393–404. Springer.

Yuhang Zhou, Suraj Maharjan, and Beiye Liu. 2023. Scalable prompt generation for semi-supervised learning with language models. *arXiv preprint arXiv:2302.09236*.

A Appendix

This appendix serves as a supplementary section, including detailed figures and tables that provide a better insight into the experimental evaluations and additional analyses that demonstrate the extent of our findings. Specifically, the appendix presents data on topic modelling performance metrics and classification accuracy across different models and settings, as detailed in the main paper.

Model	20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.06	0.07	0.02	0.01	0.03	-0.02	0.01	-0.01	0.02	0	-0.03	-0.03	-0.02	-0.05
CTM	0.08	0.11	0.06	0.07	0.07	0.12	0.10	0.14	0.11	0.13	0.09	0.12	0.11	0.12
BERTopic	**0.23**	**0.24**	0.20	**0.21**	**0.21**	**0.22**	**0.24**	**0.22**	**0.22**	**0.22**	0.19	0.19	**0.22**	**0.22**
U-BERTopic	0.22	0.22	**0.21**	**0.21**	0.20	**0.22**	0.23	0.20	**0.22**	0.19	**0.21**	**0.20**	0.21	0.21

Table A1: **NPMI** Scores for cybersecurity dataset 2018 for the four models :LDA, CTM, BERTopic and U-BERTopic. (k=20 to k=150)

Model	20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.49	0.52	0.46	0.47	0.49	0.42	0.47	0.44	0.47	0.44	0.43	0.43	0.43	0.42
CTM	0.58	0.58	0.55	0.58	0.57	0.60	0.59	0.62	0.61	0.61	0.58	0.61	0.59	0.60
BERTopic	**0.65**	**0.65**	**0.60**	**0.61**	**0.60**	0.62	**0.66**	**0.63**	0.62	**0.62**	**0.61**	0.61	**0.62**	**0.63**
U-BERTopic	0.62	0.61	0.60	0.61	0.59	0.63	0.64	0.60	0.63	0.60	0.61	0.61	0.61	0.62

Table A2: **Coherence(CV)** Scores for cybersecurity dataset 2018 for the four models: LDA, CTM, BERTopic and U-BERTopic. (k=20 to k=150)

Model	20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.56	0.52	0.50	0.55	0.59	0.58	0.55	0.57	0.56	0.57	0.59	0.57	0.58	0.59
CTM	0.86	0.81	0.85	0.79	0.77	0.65	0.59	0.50	0.50	0.46	0.42	0.40	0.40	0.36
BERTopic	0.85	0.88	**0.90**	0.86	0.87	**0.88**	**0.87**	**0.87**	0.87	**0.86**	**0.85**	**0.85**	0.83	0.83
U-BERTopic	**0.87**	**0.9**	0.88	0.87	**0.88**	0.86	**0.87**	**0.87**	**0.88**	**0.86**	**0.85**	0.84	0.84	**0.84**

Table A3: **Topic Diversity** Scores for cybersecurity dataset 2018 for the four models :LDA, CTM, BERTopic and U-BERTopic. (k=20 to k=150)

Model	20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.27	0.27	0.23	0.26	0.29	0.24	0.26	0.25	0.26	0.25	0.25	0.24	0.25	0.25
CTM	0.50	0.47	0.47	0.46	0.43	0.39	0.35	0.31	0.31	0.28	0.24	0.25	0.24	0.22
BERTopic	**0.56**	**0.57**	**0.54**	**0.53**	**0.52**	**0.54**	**0.58**	**0.55**	0.53	**0.54**	0.51	**0.52**	**0.52**	**0.52**
U-BERTopic	0.54	0.55	0.53	**0.53**	**0.52**	**0.54**	0.55	0.52	0.55	0.51	**0.53**	0.51	**0.52**	**0.52**

Table A4: **Topic Quality** Scores for cybersecurity dataset 2018 for the four models :LDA, CTM, BERTopic and U-BERTopic. (k=20 to k=150)

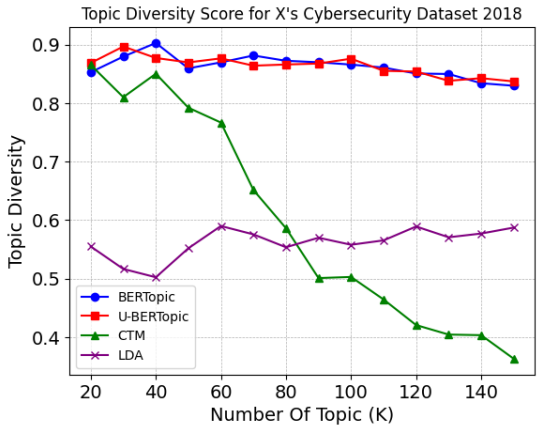

Figure A1: Topic Diversity

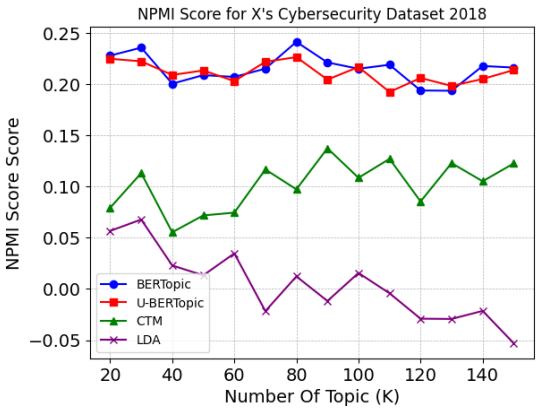

Figure A2: NPMI scores

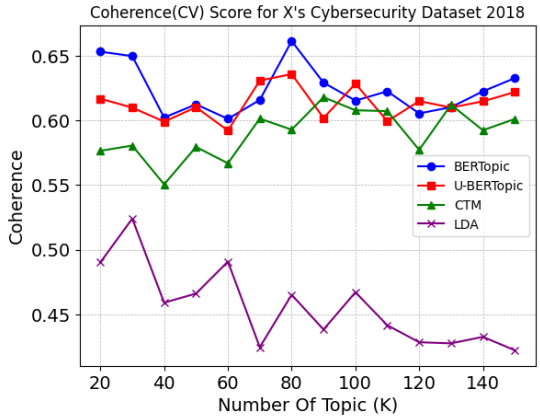

Figure A3: Coherence (CV) scores

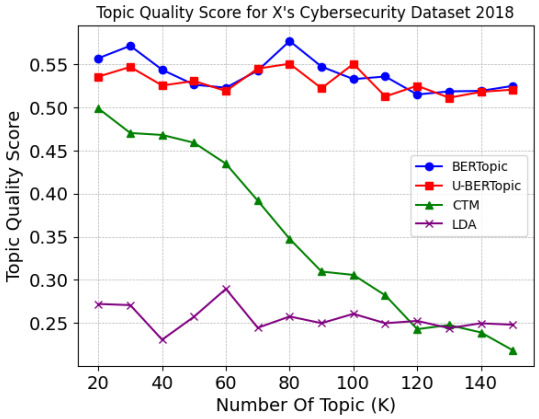

Figure A4: Topic Quality scores

Figure A5: Overview of the intrinsic evaluation metrics for topic modeling algorithms applied on the cybersecurity dataset 2018, showcasing measures of diversity, coherence, and quality across varying numbers of topics (k=20-150).

Zero-day Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.98
CTM	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.98
BERTopic	0.97	0.97	0.98	0.98	0.97	0.97	0.98	0.97	0.97	0.98	0.97	0.97	0.98	0.98
U-BERTopic	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.98	0.98	0.97	0.98	0.98
Botnet Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.96	0.95	0.96	0.95	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.96	0.96	0.96
CTM	0.95	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.97
BERTopic	0.95	0.96	0.96	0.96	0.97	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.96	0.96
U-BERTopic	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.96
DDoS Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.86	0.86	0.86	0.88	0.88	0.88	0.89	0.88	0.89	0.89	0.9	0.89	0.9	0.89
CTM	0.89	0.91	0.92	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.94	0.9
BERTopic	0.9	0.92	0.92	0.92	0.92	0.92	0.93	0.92	0.92	0.93	0.93	0.93	0.94	0.92
U-BERTopic	0.89	0.91	0.92	0.92	0.92	0.93	0.93	0.93	0.92	0.93	0.93	0.92	0.93	0.93
Leak Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
CTM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
BERTopic	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
U-BERTopic	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Ransomware Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.82	0.84	0.85	0.85	0.85	0.84	0.86	0.86	0.86	0.86	0.85	0.87	0.86	0.87
CTM	0.9	0.89	0.91	0.91	0.92	0.9	0.91	0.92	0.93	0.92	0.91	0.91	0.9	0.92
BERTopic	0.87	0.88	0.87	0.88	0.87	0.87	0.88	0.87	0.88	0.87	0.87	0.88	0.89	0.89
U-BERTopic	0.88	0.88	0.87	0.88	0.87	0.88	0.89	0.88	0.86	0.88	0.88	0.89	0.89	0.88
Vulnerability Attack														
	k=20	30	40	50	60	70	80	90	100	110	120	130	140	150
LDA	0.69	0.73	0.74	0.72	0.75	0.77	0.77	0.76	0.79	0.77	0.78	0.78	0.78	0.79
CTM	0.84	0.85	0.86	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.85	0.87	0.87	0.88
BERTopic	0.84	0.81	0.86	0.87	0.88	0.87	0.89	0.88	0.85	0.84	0.88	0.85	0.88	0.88
U-BERTopic	0.85	0.88	0.88	0.87	0.87	0.89	0.85	0.85	0.87	0.89	0.86	0.87	0.87	0.86

Table A5: Accuracy Scores by cyberAttack, model and number of topics on cybersecurity 2018 dataset

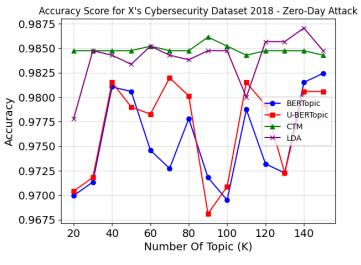

Figure A6: Accuracy - Zero Day Attack Label (k=20-150)

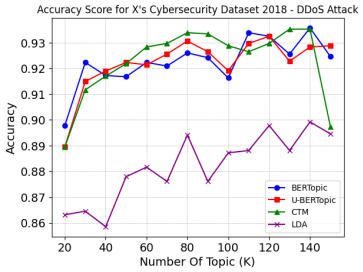

Figure A7: Accuracy - DDoS Attack Label (k=20-150)

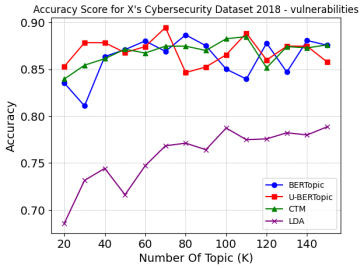

Figure A8: Accuracy - Vulnerabilities Label (k=20-150)

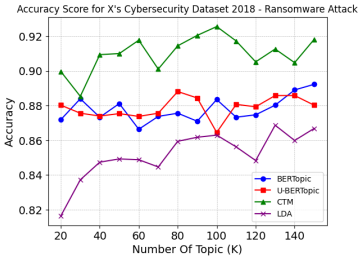

Figure A9: Accuracy - Ransomware Attack Label (k=20-150)

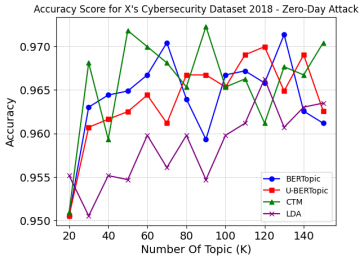

Figure A10: Accuracy - Botnet Attack Label (k=20-150)

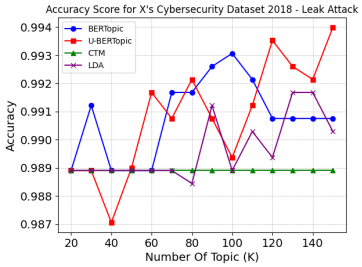

Figure A11: Accuracy - Leak Attack Label (k=20-150)

Figure A12: Comparative analysis of accuracy scores for various cyberattack labels in the Cybersecurity Dataset 2018.

Model	CTM	BERTopic	LDA	U-BERTopic
Topic1	reading , dark , infosec , security , credentials	exchange , microsoft, infosec, the, servers	microsoft, infosec, exchange, technology, software	techcrunch, technology, software, infosec, raises
Topic2	hacker , infosec , technology , news , ransomware	vulnerability, is, id, cve, unique	microsoft, exchange, infosec, vulnerability, software	proxyshell, proxylogon, exchange, servers proxynotshell
Topic3	cybersecurity , read , malware , details , proxyshell	techcrunch, technology, software, infosec, toward	infosec, microsoft, software, technology, exchange	owasp, knowage, xss, parameter, crosssite
Topic4	deal , bundle , outlook , mac , serghei	outlook, serghei, Microsoft, emails	exchange, infosec, microsoft, software	ransomware, servers, deploy, exchange, dearcry
Topic5	id , cve , unique , vulnerability , remote	thx, Pogowasright, continued, pcrisk, advintel	pogowasright, thx, exchange, microsoft, infosec	outlook, serghei, issues, emails, search
Topic6	server , vulnerability , user , files , attacker	office, deal, get, license, bundle	microsoft, infosec, software, technology, exchange	owasp, cyber, security, new, resources
Topic7	owasp , suite , knowage , parameter , xss	yanluowang, Ransomware, gang, decryptor, stolen	microsoft, exchange, vulnerability, server, code	thx, pogowasright, continued, pcrisk, advintel
Topic8	exchange , server , vulnerabilities , onpremises , exploited	Yahoo, gmail, iranian, hackers, tool	microsoft, infosec, exchange, software, technology	deal, office, bundle, mac, training
Topic9	toward , techcrunch , raises , technology , infosec	broward, breach, health, data, people	microsoft, exchange, infosec, technology, software	execution, remote, id, cve, unique
Topic10	surveillance, agents, data, breach, health	microsoft, exchange, ransomware, proxyshell, servers	spoofing, vulnerability, office, microsoft, feature	

Table A6: Comparison between topic modelling results on Microsoft Exchange Server case study dataset