Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.035741
Article

Deep Learning Based Cyber Event Detection from Open-Source Re-Emerging Social Data

Farah Mohammad[1,*], Saad Al-Ahmadi[2] and Jalal Al-Muhtadi[1,2]

[1]Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh, 11543, Saudi Arabia
[2]College of Computer & Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia
*Corresponding Author: Farah Mohammad. Email: fsheikh@ksu.edu.sa
Received: 01 September 2022; Accepted: 12 November 2022; Published: 30 August 2023

Abstract: Social media forums have emerged as the most popular form of communication in the modern technology era, allowing people to discuss and express their opinions. This increases the amount of material being shared on social media sites. There is a wealth of information about the threat that may be found in such open data sources. The security of already-deployed software and systems relies heavily on the timely detection of newly-emerging threats to their safety that can be gleaned from such information. Despite the fact that several models for detecting cybersecurity events have been presented, it remains challenging to extract security events from the vast amounts of unstructured text present in public data sources. The majority of the currently available methods concentrate on detecting events that have a high number of dimensions. This is because the unstructured text in open data sources typically contains a large number of dimensions. However, to react to attacks quicker than they can be launched, security analysts and information technology operators need to be aware of critical security events as soon as possible, regardless of how often they are reported. This research provides a unique event detection method that can swiftly identify significant security events from open forums such as Twitter. The proposed work identified new threats and the revival of an attack or related event, independent of the volume of mentions relating to those events on Twitter. In this research work, deep learning has been used to extract predictive features from open-source text. The proposed model is composed of data collection, data transformation, feature extraction using deep learning, Latent Dirichlet Allocation (LDA) based medium-level cyber-event detection and final Google Trends-based high-level cyber-event detection. The proposed technique has been evaluated on numerous datasets. Experiment results show that the proposed method outperforms existing methods in detecting cyber events by giving 95.96% accuracy.

Keywords: Social media; twitter; cyber; events; deep learning

1 Introduction

The number of cyberattacks and data theft events is rapidly rising. The threats to society and the economy have significantly increased as a result of technological innovation and internet-based platforms [1]. Currently, there is a lack of sophisticated security measures and threat detection mechanisms for software industries and organizations. According to estimates, cybercrimes and cyberattacks are becoming more serious and frequent, and businesses must deal with several difficulties as a result. Most of the time, machine learning models, tools, and applications have been developed to address this problem. It is necessary to define a model that can capture this sort of threat [2].

Another factor in the inability to identify cyberattacks is the absence of information that attackers communicate and plot on various forums before carrying out an assault. A crucial tool for dealing with such circumstances is seen to be the messages made on public forums. Not just this, but also numerous other ongoing studies have found that such media choices have a significant impact. A remarkable study by Khatoon et al. [3] examined the possibility of using a Twitter post to alert the Japanese populace to an impending earthquake. The findings indicate that tweets were able to communicate more effectively and fastly than official announcements made by the Japan Meteorological Agency (JMA) [4]. This study illustrates a change in human behaviour caused by the use of such instruments to gather specific information.

On social media, it has been common practice to share hacker services like disseminating harmful software and software vulnerabilities. Moreover, hackers utilize these exploits for system flaws to compromise the organization's security network to carry out undesirable actions. These include stealing confidential data, spying and espionage as well as launching distributed denial-of-service assaults [5]. A prime example of this occurred on October 14, 2014, when 254 unique software flaws belonging to multiple vendors, including Adobe, Oracle, and Microsoft, were made public on discussion boards [6]. Due to the hackers' inability to access their prior communications, this catastrophe occurred.

There is numerous research that has been covered by Open Source Intelligence (OSINT). Their analysis of OSINT indicates that the cyber security industry offers a variety of scenarios based on offensive and defensive methods that may be sufficient for a business to become secure. On the other hand, OSINT also depends on the conversation of hackers on social media sites. They also address Twitter's role in significant cyber events, such as the publication of multiple zero-day Denial-of-service (DDoS) vulnerabilities in Microsoft Windows, user reports on various DDoS attacks, the publication of sensitive data, and the origins of ransomware operations [7,8].

Shin et al. [9] discussed that open data sources are a great place to find information about threats. Their work highlight the crucial aspect of the security of installed software and systems is the early detection of developing security threats from such information. In their research, they offer a novel event detection method that, regardless of the volume of mentions, can instantly identify significant security events from Twitter, such as new threats and the revival of an assault or related event. In contrast to the current methods, their suggested method identifies candidate events from among hundreds of occurrences by keeping track of new and re-emerging words. Then, by grouping tweets associated with the trigger words, it creates events. With this method, they were able to identify emerging and resurgent dangers as soon as feasible.

Another work [10] introduced a system that mines text for data on cybersecurity-related events and uses that data to fill a semantic model in preparation for inclusion into a knowledge network of cybersecurity data. It was trained using a fresh corpus of 1,000 English news items from 2017 to 2019 that are richly annotated using event-based labels and cover both cyberattack and vulnerability-related

incidents. Their proposed model defines 20 argument types that are appropriate for events, along with five event subtypes and their semantic functions (e.g., file, device, software, money). Rich linguistic features and word embeddings can be incorporated using the proposed system, which employs various deep neural network techniques. The results of their testing on each part of the event detection pipeline demonstrated that each subsystem functions effectively.

The work of another research [11] presented that a common perception is that social media serves as a sensor for many societal events, such as epidemics, demonstrations and elections. Social media is used as a crowdsourced sensor to gather information on ongoing cyberattacks, according to their description. Their method requires no training or labeled samples and detects a wide range of cyber-attacks. A novel query expansion strategy based on convolution kernels and dependency parses was used to model semantic structure and identify crucial event features. They also showed that their methodology reliably recognizes and encodes events, exceeding previous methods, through a large-scale investigation across Twitter.

Many researchers have attempted to extract detailed semantic information about cyber security events, however, they have only extracted event arguments that fall within the span of sentences [12–14]. When the event arguments that need to be recognized are dispersed across numerous phrases, these investigations still have limitations. In this study, they presented a methodology for efficiently extracting cyber security events from cyber security news, blogs, and announcements at the document level. Most of the work formulates the job of extracting document-level events as a sequence tagging issue. The objective is to extract from documents the cybersecurity-related arguments. The first step is to embed the characters and add the word information to the character representations. Then, to obtain the cross-sentence context information, they construct a sliding window technique. Finally, their approach forecast what each character's label will be. The experimental findings show the efficiency of the proposed model, which they test using three approaches and a Chinese cyber security dataset.

The processing of natural language processing (NLP) is an important field of artificial intelligence. The field of natural language processing (NLP) makes the potential interaction between people through social media forums. NLP makes use of Artificial Intelligence (AI) algorithms to take social media reviews as a dataset, process it, and then provide them in a format that is processable for analysis and prediction [15]. To successfully analyze the data, sentiment analysis, is one of the most significant approaches that determine the polarity of specific entities and events, which may be either positive, neutral, or negative. To identify cyber incidents in advance of their occurrence, this research work provides a hybrid sentiment analytic approach with a deep learning model. In this article, we developed a neural network-based end-to-end threat intelligence architecture without the need for additional feature engineering or processing pipeline techniques. The proposed technique is composed of data collection, data transformation, feature extraction using deep learning and final event detection. The proposed technique is beneficial for an organization where high security at an early stage is desirable. The following is the primary goal that this study aims to achieve:

- The key contribution of this research is to define a new cyber event detection mechanism that is user-friendly and easily adaptable to any organizational setup.
- A novel approach of using social page count and Google trending mechanism with LDA produces better event detection at both levels (medium as well as high).
- From the experimental evaluation, it has been observed that the proposed technique produces better accuracy with a value of almost 96% due to the usage of new similarity measures.

The remainder of the paper is structured as follows: Section 2 discusses the proposed methodology, Section 3 describes experimental results and discussions and Section 4 concludes the proposed research.

2 Methods and Materials

This section describes the main phase and includes the corresponding diagrams. The suggested framework, which recognizes cyber-events from social media information is shown in Fig. 1. The detail of data collection, data transformation, extraction of features and cyber-event prediction has been discussed in the following subsections.

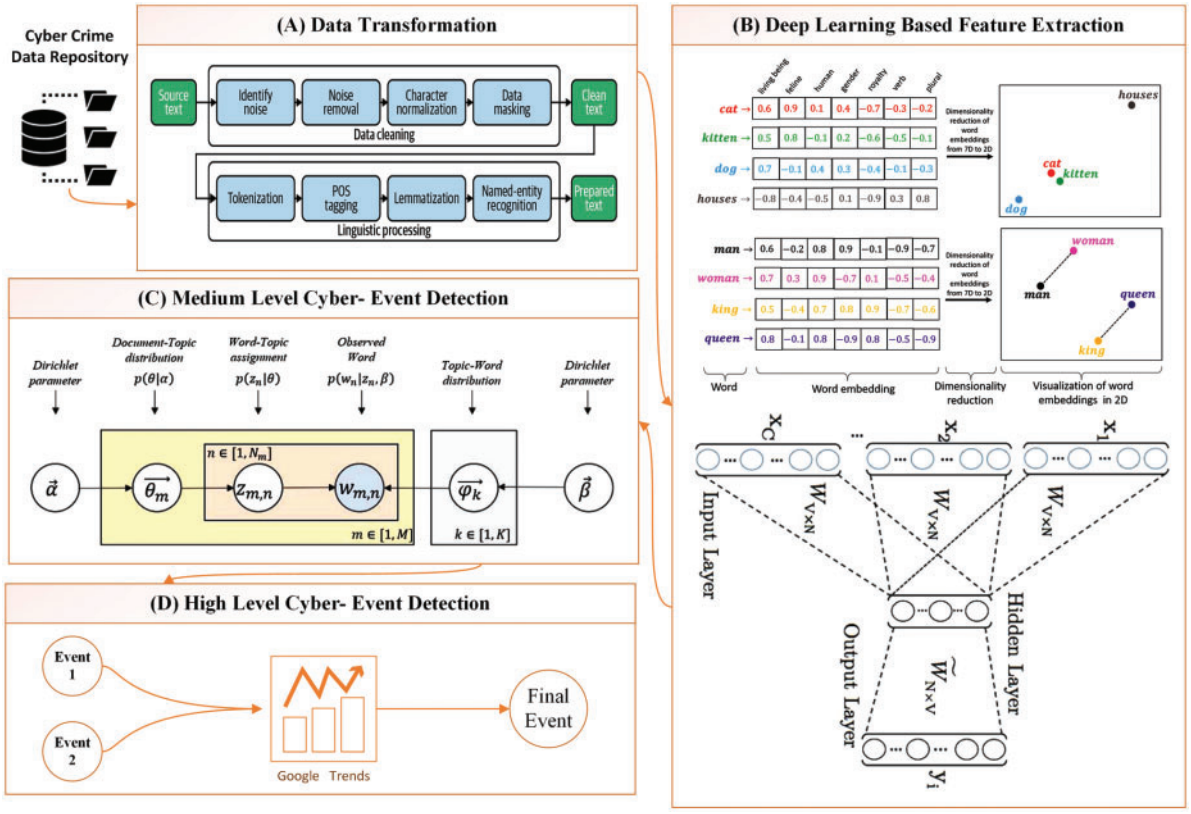


Figure 1: Proposed framework for cyber event detection

2.1 Data Collection

We have collected tweets from several websites to identify cyber events. These tweets were gathered straight from social media using the Twitter Application Programming Interface (API), sometimes known as the streaming API [16] because it enables us to obtain tweets in real-time. Twitter [17] has two unique applications available. The first one is a search API that retrieves prior tweets that adhere to requirements that a user has specified. Another dataset was collected using streaming API, which is completely different from the search API. To collect tweets over an extended period, this application

keeps an open Hypertext Transfer Protocol (HTTP) connection. Different cyber-threat and cyber-event-related data have been gathered based on these APIs. Table 1 provides a detailed description of the collected data.

Table 1: Dataset description

#.	Dataset description	URL	Number of reviews
DS-I	Cyber dataset	https://shorturl.at/ajrTX	2225
DS-II	Dataset containing cyber threats	https://shorturl.at/jkoLW	31281
DS-III	Tweets containing cyber threats	https://shorturl.at/celv1	1578
	Review description		
#.	No. of document	Vocabulary size	Review contents
DS-I	45	7,679	Cybercrime and hacking
DS-II	32	6,323	Herrasment, Pishing
DS-III	35	4,231	Hacking, Privacy, spam event

Two different streams have been adopted for data collection keywords-based and Twitter accounts of well-known security professionals [17], well-known security news sources, security firms and their research groups and vulnerability feeds. In addition, we also collected the Twitter accounts that different organizations' security experts follow. The first dataset [DS-I] from the aforementioned table is about cyber-threat and was taken from the provided website. Nearly 31281 tweets about cyber threats are included in this dataset. The tweets' vocabulary size was almost 48019 words. The following dataset is DS-II, which is gathered from a BBC news link and has about 2225 tweets on various cyber-related offenses that are posted by various communities. Last but not the least, the Twitter cyber threat dataset i.e., DS-III includes a total of 1578 tweets that were collected from a variety of social media sites, including Kaggle, Wikipedia, and Twitter.

2.2 Data Transformation

The classification of tweets enables the separation of several events connected to a term by event type. For instance, if the word "linux" is found to be re-emerging, several events, such as the discovery of new Linux vulnerabilities and the debut of new Linux malware may happen on the same day. As part of our strategy, we include broad phrases such as "attack," "hack," and "leak" in the "others" keyword list to guarantee that it does not overlook critical security-related incidents. After the data has been collected, the tweets are preprocessed and converted so that a list of terms by using the following predefined rules:

- Each tweet has undergone through named entity recognition (NER) process to compile a list of names of individuals which subsequently exclude from tweets.
- To find the proper nouns, including virus names, vulnerabilities, company names, and product names, we apply a part of speech (POS) tag to every tweet.

- Symbols, Hypertext Markup Language (HTML) tags, Uniform Resource Locator (URLs) and Twitter handles are all stripped from each tweet. Moreover, the most frequent terms that appear in the majority of texts are stop-words such as “the”, “a”, “of”, “or”, “to” etc are also eliminated.
- Twitter accounts generate a lot of noise when it comes to monitoring because many Twitter users abuse them for self-promotion that is also removed.
- Each word is lemmatized so that all of its possible inflected forms can be represented by a single lemma.

After the successful deployment of the proposed data transformation process, some of the obtained topics from different tweets are shown in Table 2.

Table 2: Resulting terms from transforming data

Data lemmatization	[‘program’, ‘security’, ‘bug’], [“Hotel,” “malware,” “program,” “disclose,” “customer,” “individual,” “data”], [‘London,’ ‘information,’ ‘passport,’ ‘grab,’ ‘violation’], [‘person’, ‘exposure,’ ‘collude’] [‘internet’, ‘mafia’, ‘organization’, ‘freedom’], [‘coder,’ ‘criticise,’ ‘stealing’] [‘washing’], [“worker,” “cost,” “fraud”], [leak, record], [‘crime’, ‘brand’, ‘stolen’]
Preprocessing	[‘program’, ‘security’, ‘bug’], [“hotel,” “malware,” “program” “disclose,” “customer,” “individual,” “data”], [‘london,’ ‘information,’ ‘passport,’ ‘grab,’ ‘violation’], [‘person’, ‘exposure,’ ‘collude’] [‘internet’, ‘mafia’, ‘organization’, ‘freedom’], [‘coder,’ ‘criticise,’ ‘stealing’] [‘washing’], [“worker,” “cost,” “fraud”], [leak, record], [‘crime’, ‘brand’, ‘stolen’]

2.3 Deep Learning-Based Feature Extraction

Following the data transformation process, the feature vector has been extracted using a deep learning-based feature extraction approach. Algorithm 1 shows the working of the deep learning-based feature extraction process. The performance of deep learning-based prediction models like word2vec is superior to that of conventional machine learning models like Term Frequency-Inverse

Document Frequency (TF-IDF) and frequency-based models. The Continuous Bag of Words (CBOW) mechanism has been used in this work where the input layer assigned a weight to each of the transformed words. These weighted words have several neurons in a hidden layer that is fully coupled to them. This layer's size is modified by the word vector dimensions obtained.

Algorithm 1: Feature Set Generation

```
Input: A Set of reviews
Output: A set of obtained features
Initialization:
Tweet: privacy authorities receive 65000 mail about cyber attacks on govt. websites.
Feature Extraction ()
    E ← Φ
    F← Feature Seed
        For Tweet, ti in review r
            Learn template form seed features
                Search pos ← r[j].pos, pos_text← r[j].text, pos_arc ← r[j].dep for each ti
                IF r[j].tag in ('NN', 'NNP', 'NNS', 'NNPS')
                    Extract selected feature and add it to E
                    Update F
        Terminate
 fetures ← ['privacy', 'authorities', 'cybr attacks','websites']
```

Assume that V is the word-vocabulary vector and N is the word vectors dimension. The anticipated weight W1 of size $V.N$, where each row represents vocabulary, will be calculated by the hidden layer. On the other hand, the output layer, which has been represented by a predetermined matrix with the name WO and the size of $N.V$, is completely connected to the hidden layer. The columns of this matrix, like the last one, each represent a word from the dictionary. Consider sending a training tweet that includes the phrases "the security leakage" "Hacker attack due to weak security" and "data lost" to have a better knowledge of this procedure.

This tweet has a word count of 12, with each word being represented by its index. Assume that the proposed neural network uses 12 input neurons and 12 output neurons to represent this. The WI and WO will therefore be configured as 12×3 and 3×12 matrices for this example.

According to the characteristics of neural networks, each neuron will first be given a random weight as indicated below.

$$
\mathrm{WI} = \begin{matrix}
-0.094491 & -0.443977 & 0.313917 \\
-0.490796 & -0.229903 & 0.065460 \\
0.072921 & 0.172246 & -0.357751 \\
0.104514 & -0.463000 & 0.079367 \\
-0.226080 & -0.154679 & -0.038422 \\
0.406115 & -0.192794 & -0.441992 \\
0.181755 & 0.088268 & 0.277574 \\
-0.055334 & 0.491792 & 0.263102
\end{matrix}
$$

$$
\mathrm{WO} = \begin{matrix}
0.023074 & 0.479901 & 0.432148 & 0.375480 & -0.364732 & -0.119840 & 0.266070 & -0.351000 \\
-0.368008 & 0.424778 & -0.257104 & -0.148817 & 0.033922 & 0.353874 & -0.144942 & 0.130904 \\
0.422434 & 0.364503 & 0.467865 & -0.020302 & -0.423890 & -0.438777 & 0.268529 & -0.446787
\end{matrix}
$$

The result from the hidden layer can be determined using the data input vector as follows: Ht = IWI = [−0.490796 −0.229903 0.065460]

H stands for the hidden layer, I for the input vector, and ***WI*** for the weight matrix. After that, an activation vector is used to connect the input layer to the output layer. Additionally, each output shows the word embedding of the input vocabulary items, which each represent one feature. The Feature that is produced by the word2vec technique described above is vectorized and has many dimensions. Each vector is transformed into a comprehensible 2D presentation to reveal the secret word. The final features in real space have been represented using t-distributed stochastic neighbour embedding (t-SNE) for these objectives.

2.4 Medium Level Event Detection

LDA has been used for medium-level event detection which is used as a filter to obtain words that accurately describe a cyber-event. The Bayesian theorem is a widely used statistical metric that forms the foundation of the LDA's construction. With the use of LDA, we identify collections of various cyber events based on their semantic similarity to a certain document. With this study, LDA extracts many subjects from each tweet document [18]. In cyber-event modeling, we assume that each event in our data collection is easily representable as a composite of numerous other events and that each event represents the representation of numerous other words.

LDA creates a set of topics from words identified in a given document by comparing a given set of documents to each term (ti). ***N*** topics have been produced as a result, and each subject represents a ***Nt*** keyword. The two variables ***N*** and ***Nt*** are used to adjust how specialized. LDA Obtain a list of subjects P(|d) for each document di, where each topic is a concatenation of the terms P(t|) described in Eq. (1). P(ti|d), where ***I*** is a specific topic and P(ti|i = j) is the probability of a term ***ti*** in a topic ***j***, is the probability of a term ti in document ***d***. The likelihood of choosing a phrase from subject ***j*** is P(i = j|d).

$$P\left(t_i|d\right) = \sum_{j=1}^{N_\theta} P\left(t_i|\theta_i = j\right) P\left(\theta_i = j|d\right) \tag{1}$$

LDA is used to estimate the document topic distribution P(θ|d) and topic term distribution P(t|θ) using an unlabeled corpus of documents. Gibbs sampler [19] executes many times for each word ti in a document di and then samples a new subject j depending on them. Ctθ represents the number of topics, CDθ represents the number of documents containing topic assignments, T represents all subject assignments, θ-i represents all topic terms. Based on these total counts, posterior probabilities are calculated as stated in the below equations.

$$P\left(\theta_i = j|t_i, d_i, \theta_{-i}\right) = \frac{C_{t_ij}^{t\theta} + \beta}{\sum_w C_{tj}^{t\theta} + T\beta} \times \frac{C_{D_ij}^{D\theta} + \alpha}{\sum_\theta C_{d_i\theta}^{D\theta} + \theta\alpha} \tag{2}$$

$$P\left(t_i|\theta_i = j\right) = \frac{C_{t_ij}^{t\theta} + \beta}{\sum_w C_{tj}^{t\theta} + T\beta} \tag{3}$$

$$P\left(\theta_i = j|d_i\right) = \frac{C_{D_ij}^{D\theta} + \alpha}{\sum_\theta C_{d_i\theta}^{D\theta} + \theta\alpha} \tag{4}$$

Because LDA's recursive functionality improves topic modeling's clarity, we can characterize LDA as part of NLP. Because of its statistical basis, a mix of subjects are produced and each topic can be

inferred from the others. LDA model is merely a collection of words and its description is outlined in below:

- In the initial stage, the modal displays the number of subjects you wish to extract from the input.
- Each word in a document is given a temporary topic in the second step of the algorithm. If there are any terms that repeat, distinct themes may be allocated to each one. This assignment is temporary and will change when an algorithm discovers a word that fits the topic.
- The last step of the algorithm updated the subjects that were provided. This update is effective based on the following two criteria: This work demonstrates, with regard to a single phrase, the frequency with which a particular word is used across themes.

The second question is, how often do the subjects come up in the provided document?

The organization of LDA is presented in Fig. 2, here we can see the many components of the LDA algorithm, which comprise the following:

- Parameter of dirichlet (α)
- Regarding the topical distribution of the paper (θ_d)
- Word count assignment for each subject ($Z_{d,n}$)
- Observed word ($W_{d,n}$)
- Topics (β_k)
- Topic hyperparameters (η)

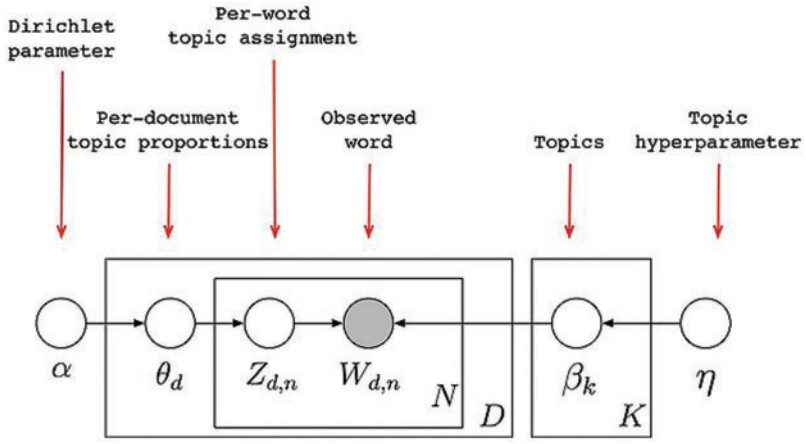


Figure 2: Complete framework of LDA

In order to successfully deployed the LDA, the Gensim library was used to develop LDA techniques in Python (version 3.4). To implement LDA within Gensim it provides a wrapper. This programming-based work involved the structuring of the input text and pattern-finding. Table 3a displays the calculations and weights for a few topics that were produced using LDA. After the successful deployment of the proposed LDA procedure, the topic obtained are considered to be as medium-level events. Some of the obtained medium-level events are depicted in Table 3b.

Table 3: (a) Scoring values of the topic terms, (b) Medium level event detection using LDA

(a)	
Score of LDA	[(0, '0.0.88∗"program" + 0.026∗"security" + 0.031∗"bug" + 0.020∗" hotel" + 0.020∗"program" + 0.020∗"error" + 0.020∗"bug" + 0.020∗"information" + 0.020∗"coder" + 0.020∗"organization"), (1, '0.041∗"function" + 0.041∗"break" + 0.036∗"take" + 0.031∗"stolen" + 0.021∗"user" + 0.021∗" program" + 0.021∗"task" + 0.021∗"hard" + 0.021∗"internet" + 0.021∗"card"), (2, '0.033∗"worker" + 0.42∗"bug" + 0.42∗"cost" + 0.30∗"farad" + 0.020∗"stolen" + 0.016∗"crime" + 0.021∗"brand" + 0.021∗"stolen" + 0.021∗"site" + 0.021∗"penalty"), (3, '0.031∗"penalty" + 0.031∗"breach" + 0.020∗"bug" + 0.020∗"cheater" + 0.030∗"brand" + 0.014∗"farad" + 0.021∗"take" + 0.031∗"verify" + 0.016∗"away" + 0.021∗"stolen"), (4, '0.031∗"user" + 0.031∗"criticise" + 0.031∗"stealing" + 0.031∗"error" + 0.032∗"farud" + 0.022∗"program" + 0.022∗"stolen" + 0.021∗"bug" + 0.021∗"worker" + 0.021∗"cost")]
Coherence Score	Topic N1 = 4 and Value of Coherence is 0.6543 Topic N2 = 8 and Value of Coherence is 0.6942 Topic N3 = 12 and Value of Coherence is 0.6971 Topic N4 = 24 and Value of Coherence is 0.621 Topic N5 = 28 and Value of Coherence is 0.6872 Topic N6 = 32 and Value of Coherence is 0.6323
(b)	
Tweet ID	Identified Medium-Level Event
T1	Organization, London, Freedom
T2	Insider Attack, Phishing
T3	Cryptojacking, Malware
T4	Zero-day, Expolite
T5	Cyber attack, michaelfassbender, colmmeaney, mark halloran, hacking attack

2.5 Trend Analysis (Medium Level Event Detection)

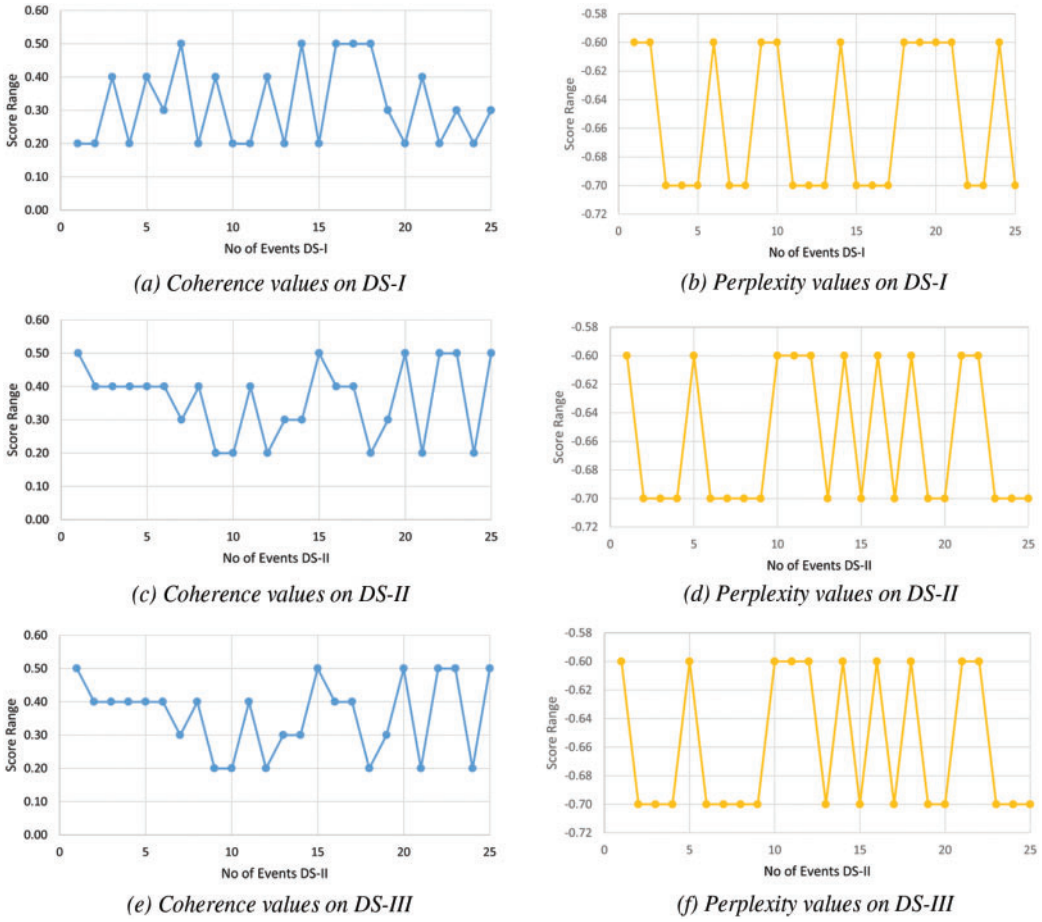
Due to their complexity and ongoing improvement, the cyber events produced by LDA are regarded as medium-level events. Google Trends [20,21] has been used to obtain a specific trending event, also known as high-level cyber-events, using the social page count method (SPC) that Google Trends has introduced. We calculate weights for all of the events that were collected from LDA in SPC by calculating the trend of the cyber-event throughout Google based on the cyber-event with the biggest trending weight that was picked as the final cyber-event. SPC counts the total number of pages utilized to gauge a specific site because it is natural for more important sites to receive more connections when it comes to how it operates.

$$Ti\ score = log\left\{\frac{\sum_{i=1}^{n} Wi}{\sum_{i=1}^{n} Wi(Ti)}\right\} \tag{5}$$

The Internet search phrase is Ti and Wi is the total number of webpages that contain Ti, in Eq. (5).

3 Experimental Results

This section discusses the experimental evaluation and findings of the proposed method. Numerous experiments have been performed to test the viability of the proposed technique. In the very first experiment, the coherence and perplexity of the proposed model on various datasets have been computed. Fig. 3 shows the obtained result on a different dataset.



(a) Coherence values on DS-I (b) Perplexity values on DS-I

(c) Coherence values on DS-II (d) Perplexity values on DS-II

(e) Coherence values on DS-III (f) Perplexity values on DS-III

Figure 3: Performance of the proposed model on different datasets

Another experiment shows the uncalculated, fluctuating relationships between vocabulary words and the dataset document's occurrences of those words. A variety of subjects are seen in the reviews. The collection's previously unidentified themes are found in this effort, and those themes are then annotated onto the documents. Finally, it reveals the latent topical patterns that are presented in Fig. 4.

Figure 4: Identification of cyber-events at the medium level using latent dirichlet allocation

Through the use of Google Trends, the findings of Google Trends to carry out high-level event detection experiments have demonstrated some of the important event subjects. The results of four distinct sorts of themes have been shown in Fig. 5 when it was trending on Google. Fig. 6 shows the Google Trends that determine word count and key topic keywords.

Figure 5: High-level final event

In addition, we compared the outcomes of the experiments to those obtained through the use of the IDCNN & BiLSTM & CRF [22] and SentiStrenght & VADER & ARIMAX [23] methodologies, as presented in Table 4. The IDCNN & BiLSTM & CRF approach may identify cyber incidents using tweets taken from Twitter [24–28]. They use both machine learning methods (Bidirectional Long

Short-Term Memory, or BiLSTM) and natural language processing techniques to accomplish the multi-task learning methodology (IDCNN, or Iterated Dilated Convolutional Neural Network).

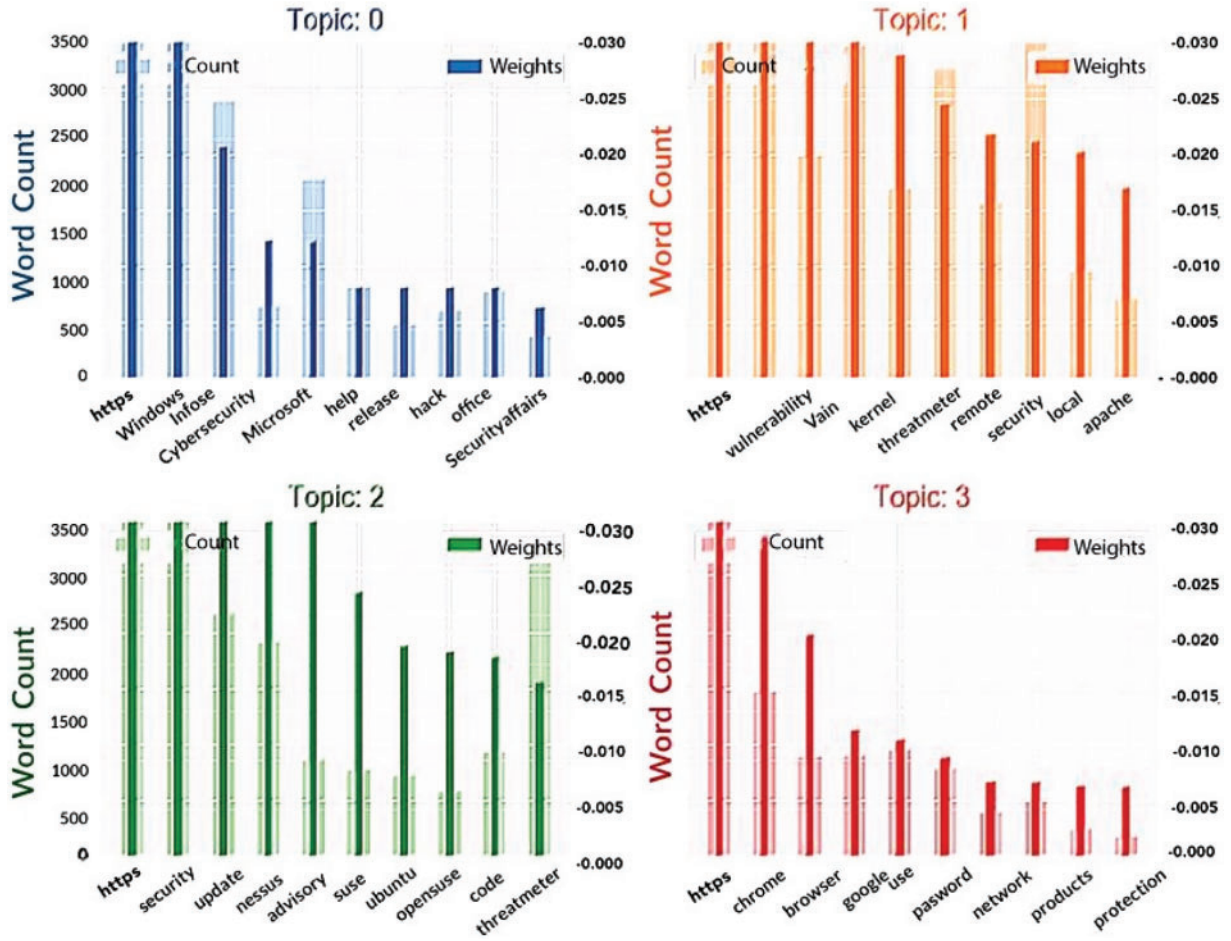


Figure 6: Using google trends to determine word count and key topic keywords

The events are identified by applying sentiment analysis to hacker forum reviews, which is done using VADER+SentiStrength+ARIMAX. They also identify the patterns of behaviour that are associated with cyber occurrences by analyzing over 400,000 posts over two years, beginning in January 2020 and concluding in January 2022. These posts were culled from more than one hundred different hacker forums.

Table 4: Comparison with the state of art techniques

#	Techniques	Recall	Accuracy	Precision	F1 Score
1	IDCNN & BiLSTM & CRF	96.98	95.02	94.99	95.97
2	SentiStrenght & VADER & ARIMAX	96.31	95.56	95.29	95.80

(Continued)

Table 4 (continued)

#	Techniques	Recall	Accuracy	Precision	F1 Score
3	Proposed Technique	97.62	95.96	96.13	96.86

It has been determined that the cyber event detection methodology that we have suggested performs far better than any existing baselines. The most pertinent cyber events from a huge set of observed events cannot be extracted using the approaches that are now available. According to the findings, the utilization of LDA has the potential to increase the performance of cyber event detection. Fig. 7 presents the findings in their entirety.

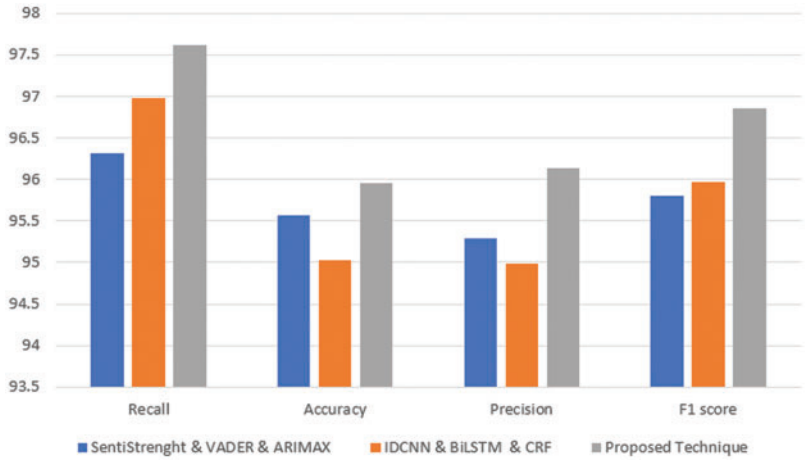


Figure 7: Comparison with existing techniques

4 Conclusion

The most widely used medium for communication in the present technological era is social media forums that enable people to converse and express their thoughts. As a result, there is an increase in the volume of content published on social media platforms. These open data sources have a plethora of information regarding the threat. The prompt detection of newly-emerging threats to the security of software and systems may be inferred from such information. This research offers a distinct event detection technique that can quickly recognize cyber events from public forums such as Twitter. Data collection, data processing, feature extraction using deep learning, medium-level cyber-event detection based on LDA, and high-level cyber-event detection based on Google Trends are the key phases of the proposed model. The proposed approach has been evaluated on several datasets. According to the results of the experimental evaluation, the suggested technique produced effective cyber event detection. In future work, the proposed work will be updated with new similarity measures and topic modeling techniques.

Funding Statement: This research work is funded by a grant from the Center of Excellence in Information Assurance (CoEIA), KSU.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Gu, "Sharing economy, technological innovation and carbon emissions: Evidence from Chinese cities," *Journal of Innovation & Knowledge*, vol. 7, no. 3, pp. 100228, 2022.

[2] F. F. Alruwaili, "Artificial intelligence based threat detection in industrial internet of things environment," *Computers, Materials & Continua*, vol. 73, no. 3, pp. 5809–5824, 2022.

[3] S. Khatoon, M. A. Alshamari, A. Asif, M. M. Hasan, S. Abdou *et al.,* "Development of social media analytics system for emergency event detection and crisis management," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 3079–3100, 2021.

[4] H. Higuchi, H. Kobori, S. Lee and R. B. Primack, "Declining phenology observations by the Japan meteorological agency," *Nature Ecology & Evolution*, vol. 5, no. 7, pp. 886–887, 2021.

[5] A. Bhardwaj and V. Mangat, "Distributed denial of service attacks in cloud: State-of-the-art of scientific and commercial solutions," *Computer Science Review*, vol. 39, no. 1, pp. 100332, 2021.

[6] R. Choi, A. Nagappan, D. Kopyto and A. Wexler, "Pregnant at the start of the pandemic: A content analysis of COVID-19-related posts on online pregnancy discussion boards," *BMC Pregnancy and Childbirth*, vol. 22, no. 1, pp. 1–11, 2022.

[7] N. Chowdhury and V. Gkioulos, "Cyber security training for critical infrastructure protection: A literature review," *Computer Science Review*, vol. 40, no. 1, pp. 100361, 2021.

[8] E. Bout, V. Loscri and A. Gallais, "How machine learning changes the nature of cyberattacks on IoT networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 248–279, 2021.

[9] H. Shin, W. Shim and J. Moon, "Cyber security event detection with new and re-emerging words," in *Proc. of the 15th ACM Asia Conf. on Computer and Communications Security*, Taipei, Taiwan, pp. 665–678, 2020.

[10] T. Satyapanich, F. Ferraro and T. Finin, "Casie: Extracting cyber security event information from text," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, pp. 8749–8757, 2020.

[11] R. P. Khandpur, T. Ji, S. Jan and G. Wang, "Crowdsourcing cyber security: Cyber attack detection using social media," in *Proc. of the 2017 ACM on Conf. on Information and Knowledge Management*, Singapur, pp. 1049–1057, 2017.

[12] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam *et al.,* "A review of content-based and context-based recommendation systems," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 3, pp. 274–306, 2021.

[13] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat and M. A. Moni, "A comparative analysis of active learning for biomedical text mining," *Applied System Innovation*, vol. 4, no. 1, pp. 23–33, 2021.

[14] K. Shaukat, F. Iqbal, T. M. Alam, G. K. Aujla, L. Devnath *et al.,* "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends in Computer Science and Information Technology*, vol. 5, no. 1, pp. 50–54, 2020.

[15] Y. Kang, Z. Cai, C. W. Tan, Q. Huang and H. Liu, "Natural language processing (NLP) in management research: A literature review," *Journal of Management Analytics*, vol. 7, no. 2, pp. 139–172, 2020.

[16] C. Barrie and J. C. T. Ho, "academictwitteR: An R package to access the twitter academic research product track v2 API endpoint," *Journal of Open Source Software*, vol. 6, no. 62, pp. 3272, 2021.

[17] A. Nawaz, T. Ali, Y. Hafeez and M. R. Rashid, "Mining public opinion: A sentiment based forecasting for democratic elections of Pakistan," *Spatial Information Research*, vol. 30, no. 1, pp. 169–181, 2022.

[18] B. Ozyurt and M. A. Akcayol, "A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA," *Expert Systems with Applications*, vol. 168, no. 1, pp. 114231, 2021.

[19] T. Park and S. Lee, "Improving the Gibbs sampler," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 2, pp. e1546, 2022.

[20] D. Aaronson, S. A. Brave, R. A. Butters, M. Fogarty, D. W. Sacks *et al.,* "Forecasting unemployment insurance claims in realtime with Google Trends," *International Journal of Forecasting*, vol. 38, no. 2, pp. 567–581, 2022.
[21] S. Pullan and M. Dey, "Vaccine hesitancy and anti-vaccination in the time of COVID-19: A Google trends analysis," *Vaccine*, vol. 39, no. 14, pp. 1877–1881, 2021.
[22] R. Askarizad, H. Jinliao and S. Jafari, "The influence of COVID-19 on the societal mobility of urban spaces," *Cities*, vol. 119, no. 1, pp. 103388, 2021.
[23] X. Wang, R. Liu, J. Yang, R. Chen, Z. Ling *et al.,* "Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering," in *2022 IEEE 25th Int. Conf. on Computer Supported Cooperative Work in Design (CSCWD)*, Singapur, pp. 406–413, 2022.
[24] C. Borchers, J. Rosenberg, B. Gibbons, M. A. Burchfield and C. Fischer, "To scale or not to scale: Comparing popular sentiment analysis dictionaries on educational twitter data," in *Int. Conf. on Educational Data Mining*, Durham, UK, pp. 1–7, 2021.
[25] S. R. Baker, N. Bloom, S. J. Davis and T. Renault, "Twitter-derived measures of economic uncertainty," [Online]. Available: Policy Uncertainty.com
[26] M. Birjali, M. Kasri and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, no. 1, pp. 107134, 2021.
[27] M. R. R. Rana, S. U. Rehman, A. Nawaz, T. Ali and M. Ahmed, "A conceptual model for decision support systems using aspect based sentiment analysis," *Proceedings of the Romanian Academy Series A-Mathematics Physics Technical Sciences Information Science*, vol. 22, no. 4, pp. 381–390, 2021.
[28] J. R. Saura, D. Ribeiro-Soriano and P. Z. Saldaña, "Exploring the challenges of remote work on twitter users' sentiments: From digital technology development to a post-pandemic era," *Journal of Business Research*, vol. 142, no. 1, pp. 242–254, 2022.