IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2019, accepted April 2, 2019, date of publication April 11, 2019, date of current version April 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910229

Analyzing and Identifying Data Breaches in Underground Forums

YONG FANG, YUSONG GUO, CHENG HUANG, AND LIANG LIU
College of Cybersecurity, Sichuan University, Chengdu 610065, China

Corresponding author: Cheng Huang (opcodesec@gmail.com)

This work was supported by the Sichuan University Postdoctoral Research Foundation under Grant 19XJ0002.

ABSTRACT Recently, underground forums play a crucial role in trading and exchanging leaked personal information. Meanwhile, the forums have been gradually used as data breaches' information sources. Therefore, it shows an upward trend in announcing the results of data theft by posting in the forums. Identifying these threads can make the compromised third-party respond quickly to the data breach incident. For this purpose, we presented a system to identify the threads which are related to data breaches automatically. The system can monitor and discover data breaches in underground forums in real-time. In addition, the study further revealed the wording characteristics of the threads by applying the feature extraction method based on LDA topic model. In this paper, the data set was collected from the surface web and the dark web. Besides, to improve the performance of the system, we compared various supervised classification algorithms in this application scenario and selected the best method for the classifier. Through the system, we identified more than 92% of data breach threads on the experimental data set.

INDEX TERMS Data breach, underground forum, text classification.

I. INTRODUCTION

Nowadays, the rapid growth of the Internet provides convenience for both individuals and organizations. People submit their personal information, such as name, email address, mobile phone number, while using the Internet to work, shop, and make friends. In recent years, frequent data breaches have drawn people's attention. This kind of leaked personal data is used for collision attack, credit card fraud, etc. by hacker groups. Therefore, more and more researchers have focused on users' data leakage. And we notice that lots of researchers concentrate on Data Leakage Detection (DLD) [1], [2] and Data Leakage Prevention (DLP) [3], [4]. The purpose of their research is to prevent the data from being leaked and detect whether the data has been compromised or leaked during the transmission. These kinds of methods have a good effect on the prevention and detection of data breaches within organizations. However, the capability to detect data breaches from information sources in the first place is of great significance to the organization's information security construction. Therefore, we are inclined to identify whether people's discussions in underground forums are related to data breaches.

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Qian.

As we know, underground forums are a kind of essential platforms which facilitate hackers' communications. These forums' contents are not only about technologies, but also security incidents. The main activities of the forums are aimed at underground commercial and malicious activities. Furthermore, some forums' main purpose is helping hackers to exchange databases which have been leaked. Talking about the latest data breaches and exchanging already leaked information are common in almost every forum. Many hackers even post announcements in underground forums as soon as they get data from compromised organizations or individuals. Therefore, most of the time, data breaches are disclosed firstly in underground forums. For example, on September 4th, 2016, the paid-to-click site ClixSense suffered a data breach. The data was posted online by attackers. And on August 28th, 2018, a user named "helen250" posted a thread in Chinese underground forum to sell 130 million China Lodging Group's user data, and the total number of leaked data reached 500 million.

Much of the previous research is about cybercrime, hacking, social network analysis in underground forums. Here we focus on discovering threads about data breaches. These threads are a kind of source of data breach threat intelligence. By focusing on these threads, network security analysts can discover and understand the current network data

leakage situation. Automated identification of these threads can quickly provide victims with a loss estimate. In this paper, our data set is collected from underground forums and communities on the surface web and the dark web. The research questions we would like to answer in this study include:

a) What are the commonalities of threads related to data breaches? **b)** How to automate the extraction of the characteristics of these threads? **c)** How to identify these threads automatically?

In general, the specific contributions of our research are the following:

- The paper proposes a system, the fully-automated technique for identifying data breach threads in underground forums, which addresses the challenge in quick discovery of data breaches. Our approach compares the performance of five popular supervised classification algorithms and chooses the optimal algorithm for the system.
- The study presents a method of feature selection based on Latent Dirichlet Allocation (LDA) topic model, which is a way of providing efficient feature vectors. We also present the wording characteristics by applying this approach.
- This system has scalability and can work for every English forum. More than 92% of data breach threads are identified in the experiment.

The following paper structure is mentioned below. The related work will be described in Section II. Section III will introduce the method we used. Section IV is about our data set. The experiments and discussions on results are presented in Section V. The last part concludes the research and summarizes the future job.

II. RELATED WORKS

In order to provide the information about our research, this section introduces prior works from three perspectives: 1) data breach related, 2) underground forum analysis and 3) text classification based on LDA topic model.

A. DATA BREACH RELATED

Some previous research focused on identifying and evaluating data breaches during transmissions [1]–[5]. On the other hand, personal information protection awareness and the risks of data breach were elaborated in many articles [6]–[9]. Butler *et al.* presented a method called REAPER which demonstrates how to leverage unique data points within credential dump to identify its distribution [10]. Li *et al.* identified data breach services in an underground supply chain [11].

B. UNDERGROUND FORUM ANALYSIS

In the early stages, much research of underground forums showed that cybercriminals cooperate and trade different products and services such as 0-day, rat [12]–[15]. Thomas *et al.* analyzed the way of cybercriminals' communications and what they exchange in forums [16]. Pastrana *et al.* focused on finding cybercrime actors in a large underground forum [17]. For evaluating private interactions, Overdorf *et al.* developed a method for automatically labelling threads that are likely to trigger private messages [18]. These studies were used to explore the market of underground forums and the social relationships of members.

As for analyzing the subject of threads, Zhang *et al.* developed a system iDetector for detecting cybercrime-suspected threads [19]. This study was used to identify threads which are benign or cybercrime-suspected. In the field of cybercriminal market research, Portnoff *et al.* built tools for automated analysis of cybercriminal markets. The tool was designed to detect the product and its price in the thread [20].

C. TEXT CLASSIFICATION BASED ON LDA TOPIC MODEL

Latent Dirichlet Allocation (LDA) plays a crucial role in feature selection of text categorization. It was proposed by Blei *et al.* [21]. Ramage *et al.* applied Labeled LDA to make a supervised topic model on credit attribution in multi-labeled corpora [22]. Tasci and Gungor compared the efficiency of LDA and traditional models in feature selection [23]. Many scholars combined LDA with supervised classification algorithms for text classification and achieved good results [24]–[26].

Different from these above works, we consider all threads in underground forums which are related to data breaches. Further, we propose a system to identify these threads.

III. METHODOLOGY OVERVIEW

As Figure 1 shown, the system structure is divided into three parts: collecter and pre-processer, feature selector, and classifier. Collecter and pre-processer are used for the collection and processing of raw data. The feature vector of the thread is generated by the feature selector. And the classifier is used to identify whether a thread is related to a data breach. They will be described in detail below.

A. COLLECTOR AND PRE-PROCESSER

We develop web crawlers for collecting underground forum discussions on the surface web and the dark web. Since each forum has a different structure, it is necessary to design different crawlers for all forums. Threads, replies, and members (including their profiles) in forums are our web crawler's targets (See Section IV. A for details). In crawlers, we add mechanisms for solving inaccessible, anti-crawling and other issues.

Each thread applies a natural language processor to remove punctuation, stop words and lower case all the words. We find data breach threads have special strings that can be used to describe the breach. This kind of representative strings in threads for maximizing the characteristics of the data breach will be illustrated by two examples and labeled by DBID recognizer. In the following paper, these labeled strings are collectively referred to as DBID. (See Section IV. B for details).

B. FEATURE SELECTOR

The feature selector is used to obtain the latent topic distribution and the statistical nature, and provides the classifier with feature vectors. We choose the LDA topic model as a

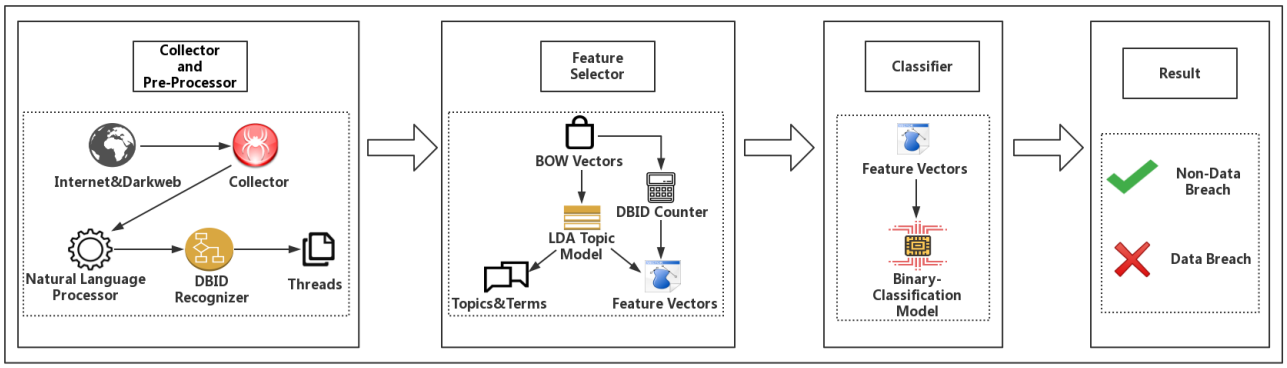

FIGURE 1. System architecture of identifying data breach threads.

topic distribution feature selector. In this model, each thread is represented as a bag-of-words (BoW) vector. LDA needs a document-term matrix to be the input. Here we define threads as T_i, $T = \{T_1, T_2, \ldots, T_n\}(|T| = n)$, the content of threads as d_i, $d = \{d_1, d_2, \ldots, d_n\}(|d| = n)$. Applying a BoW model to each d_i, we get s_i, $s_i = \{w_1, w_2, \ldots, w_n\}$, $w_i \in N$. N is a dictionary of all the words. A thread-BoW matrix as a document-term matrix is to be the input of LDA.

For training our LDA topic model, the number of topics K is indispensable. It is an effective method to determine the number of topics through perplexity and coherence. The perplexity is a statistic measure of how well a probability model predicts a sample. And it is calculated as [21]:

$$perplexity = exp\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\} \quad (1)$$

M is the total number of threads. N_d is the number of words in a thread and $p(w_d)$ represents the probability of text. Perplexity indicates uncertainty when predicting. A low perplexity indicates the probability distribution is good at predicting the sample [21].

However, LDA gives no guarantee on the interpretability of their output [27]. Therefore, coherence measures were presented by the previous work [28] to rate topics. The following formula is used to calculate the coherence:

$$coherence = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + 1}{D(w_j)} \quad (2)$$

$D(w_j)$ is the document frequency of word w_j (i.e., the number of documents with least one token of type w_j) and $D(w_i, w_j)$ is the co-document frequency of word types w_i and w_j (i.e., the number of documents containing one or more tokens of type w_i and at least one token of type w_j). [29], [30] By evaluating the values of perplexity and coherence, we choose the best K to train the LDA topic model. (See Section IV. C for details).

In addition to the topic distribution, a feature vector also contains statistical nature. DBID counter calculates the number of each DBID which is extracted and replaced by DBID recognizer. The quantitative characteristics of these different DBIDs are used to compose statistical nature vectors. Finally, we combine the topic distribution and the statistical nature to form the feature vector for the classifier. (See Section IV. D for details).

C. CLASSIFIER

The classifier is a key factor of identifying whether a thread is related to a data breach. We expect to select the classification algorithm which makes the system identify more data breaches. Therefore, we compare five supervised learning algorithms and choose the best performing algorithm for the classifier. These five classification algorithms are Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbor (K-NN). During the experiment, we perform a grid search to find each algorithm's optimal parameter. (See Section IV. D for details).

D. IDENTIFICATION PROCESS

In order to identify a thread whether it is talking about a data breach, given a thread, the thread's title and content will be separated to form a document. The feature vector which consists of the topic distribution and the statistical nature is extracted from the document based on the feature selector. Through inputting the feature vector into our classifier, the thread will be labeled as the non-data breach or data breach.

TABLE 1. An overview of the forums collected.

Name	Language	Threads	Members	Source
Nulled	English	139,969	599,085	Crawled&Leaked
Breach Forums	English	2,018	2,065	Crawled
Hack This Site	English	8,297	80,060	Crawled
Hellbound Hackers	English	1,286	43,830	Crawled
Hidden Answers	English	19,950	20,008	Crawled
Brotherhood	English	166	2192	Crawled

IV. EXPERIMENT

A. DATA SET

For the data set, we consider six forums: Nulled, BreachForums, HackThisSite, Hellboundhackers, HiddenAnswers, Brotherhood. Two ways we choose to collect the data are

crawling and using the public leaked database dump. Table 1 shows the specific information.

1) NULLED

Nulled is an infamous forum. It is a cracking community which is specializing in leaks, services, and tools for data breaches. In 2016, this forum was hacked. The hacker made the full database of the forum public. We use this database in our research. Meanwhile, we also crawled partial threads which are the latest.

2) BREACH FORUMS

Breach Forums claims it is the biggest community-based data trading forum on the Internet. This forum started in 2016. We crawled all the threads and members as of Oct 2018.

3) HACK THIS SITE

Hack This Site is a training ground for hackers to test and expand their hacking skills. This website provides a large forum for users to discuss cybersecurity. We crawled a partial of threads and members as of Oct 2018.

4) HELLBOUND HACKERS

Hellbound Hackers provides tools, courses, and discussion platforms which are related to hacking technologies. It started in 2003 and is still active now. We crawled part of threads as of Sept 2018.

5) HIDDEN ANSWERS

Hidden Answers is a big onion service community. Although it's not a community for hackers, lots of members in it are talking about hacking technologies and security news. We crawled all the threads and members as of Sept 2018.

6) BROTHERHOOD

Brotherhood is a small onion service forum. The purpose of the forum is to share hacking tools and services. We crawled all the threads and members as of Sept 2018.

Supervised learning algorithms require labeled data. Fortunately, each forum is divided into different categories. And our collected threads are marked by the forum administrator. We divide these threads into two categories (one is related to data breach, and the other is not). However, this method of labeling is not in line with the actual situation in fact. Firstly, some threads in the non-data breach board also talk about a data breach. Secondly, some of the classified threads are misclassified. Thirdly, forum administrators will re-edit threads which do not conform to the forum rules, such as changing threads' title and content to ''closed'' or ''deleted''. To guarantee the accuracy of labels, we check whether the label and subject are consistent.

Our experimental data set contains a total of 10,000 threads. We set the overall ratio of positive and negative samples to 1 : 1 and split the data 7 : 3 to form training and test subset. Table 3 shows the number of data breach and non-data breach samples in the training set and test set.

TABLE 2. The proportion of each forum in the experimental data set.

Name	Amount(Training/Test)	Proportion
Nulled	5000(3500/1500)	50.0%
Breach Forums	450(315/125)	4.5%
Hack This Site	1,800(1,260/540)	18.0%
Hellbound Hackers	300(210/90)	3.0%
Hidden Answers	2,410(1,687/723)	24.1%
Brotherhood	40(28/12)	0.4%

TABLE 3. Total number of positive and negative samples in training and test sets.

Data set type	Non-data breach	Data breach	Total
Training set	3,500(50%)	3,500(50%)	7,000
Test set	1,500(50%)	1,500(50%)	3,000

As we said before, each forum focuses on different fields. Therefore, our experimental data is composed of random sampling according to the distribution of Table 2.

B. PRE-PROCESSING

The title and content of a thread are the most representative attributes. Given a thread, we extract and combine them into a document. These original documents have a lot of extraneous characters that will serve as noise to the model training. For this reason, we remove all non-alphabetic characters and stop-words from each document in order to reduce the effect of noise.

Usually, when hackers publish the relevant announcement, they will mention the compromised third-party and the leaked documents. It is obvious that lots of threads which are related to data breaches have commonalities. Using segmentation directly will ignore this key factor. To minimize the loss of segmentation, the DBID recognizer tokenizes these commonalities. We choose two examples (detailed in Table 4) from our data set to show the commonalities and how the DBID recognizer works.

TABLE 4. Two example threads about data breaches, where commonalities are in bold.

Thread	Content
Example 1	**Twitter.com** Mirror list **twitter.txt.gz** (**338.4 MiB**) Note: i did not upload it, i found it somewhere in clearnet and sharing here. Download at your own risk, i am not responsible for anything.
Example 2	**X2** Premium Spotify Accounts Here! Enjoy these accounts: **username:password**, **aa@bb.com:xxxxx**

As shown in table 4, both example 1 and example 2 are related to breaches. The highlighted strings in table 4 are our concerns. In example 1, ''Twitter.com'' and ''twitter.txt.gz'' show that this thread is talking about the data breach of Twitter. We match such strings and replace them with special identifiers (**dbdomaindb** and **dbfiledb**). They are used to replace domains and file names respectively. Here ''338.4 MiB'' indicates the file size in example 1. In most cases, the author mentions the size of the leaked file storage

TABLE 5. After processing by the DBID recognizer, the replaced DBIDs are in bold.

Thread	Content
Example 1	**dbdomaindb** Mirror list **dbfiledb** (**dbsizedb**) Note: i did not upload it, i found it somewhere in clearnet and sharing here. Download at your own risk, i am not responsible for anything.
Example 2	**dbcountdb** Premium Spotify Accounts Here! Enjoy these accounts: **dbaccpwddb**, **dbemailpwddb**

in the thread. To preserve this characteristic, **dbsizedb** is used to replace the file size. "X2", "username:password" and "aa@bb.com:xxxx" represent quantity identification words, account-password pair and email-password pair in example 2. These strings are identified as **dbcountdb**, **dbaccpwddb**, **dbemailpwddb**. Table 5 presents the processed threads after replacement.

Based on these rules, the DBID recognizer is built to automatically extract and replace the above string. However, the large amount of DBIDs in a document make a negative impact on the performance of LDA model. For this reason, we set the maximum number of each identifier in a thread to 2. After pre-processing, we generate document vectors by the BoW model.

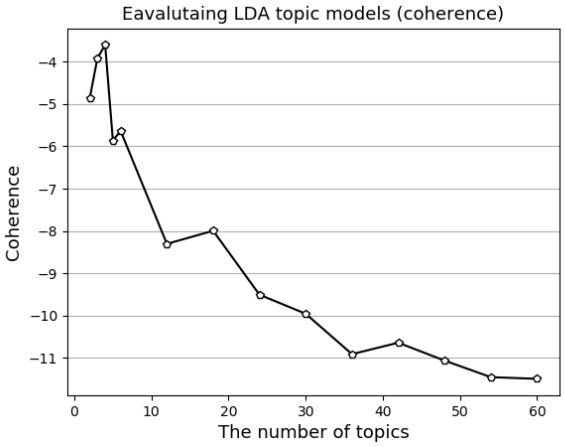

FIGURE 2. The LDA model's coherence of different number of topics.

C. LDA TOPIC MODEL

In Section III. B, we choose the LDA topic model as the topic feature selector. Whether the correct number of topics K is selected has a great influence on the performance of the model during the training. Here the coherence and the perplexity are indicators for evaluating the performance of the model. By comparing fourteen LDA models with different K, we select the K that is the most suitable for this experiment. In the experiment, K is set to 2 - 5 and 6 - 60 (interval 6). Figure 2 and Figure 3 show the relevant changing curve of coherence and perplexity under different topic numbers.

In Figure 2, coherence is the highest when the number is 4. And from 5 to 60, the curve shows a downward trend. It reaches the highest value in this interval at the point where

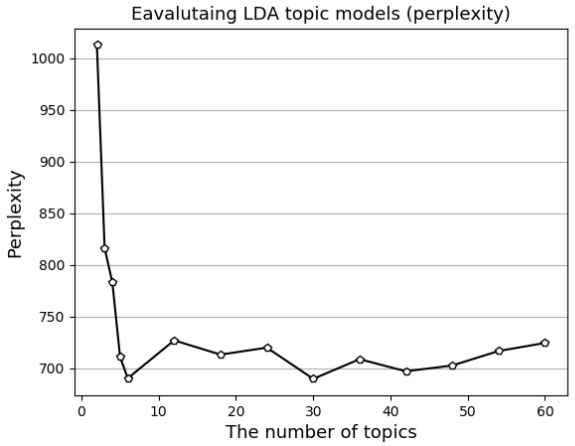

FIGURE 3. The LDA model's perplexity of different number of topics.

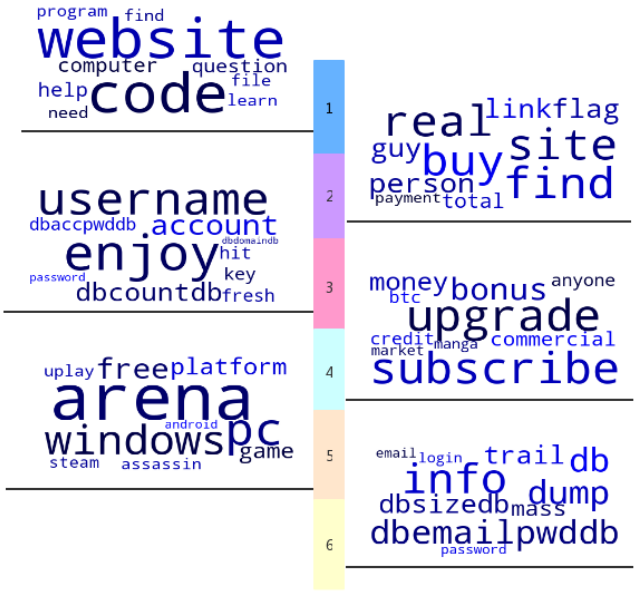

FIGURE 4. Six topics generated by the LDA model.

the number of topics is 6. Figure 3 presents that perplexity is decreasing clearly from 2 to 6. The value of perplexity is between 690 and 724 from 5 to 60. Although the coherence is not the highest value in the case of $K = 6$, it is the highest in the range of $5 - 60$. Meanwhile, the perplexity reaches the lowest value where $K = 6$. On the other side, the more the number of topics, the more parameters are estimated by the LDA model and the more computational cost. Combining the results of two figures, we select $K = 6$.

Through the LDA topic model, we extract the ten most representative terms for each topic. As shown in Figure 4, topic 1 is about programming. Topic 2 is not clear, and we guess it may be related to CTF (Capture The Flag). Topic 4 and topic 5 are related to trading and entertainment. Obviously, the most frequent terms in topic 3 and 6 are directly related to data breaches.

The topic model divides the data breach into two categories. One is talking about personal account leakage, and the

other is about the leakage of databases. In most cases, when the author publishes data breach information, it will mention what kind of data is included, such as emails, accounts, and passwords. Meanwhile, as we expected, DBIDs appear in Figure 4. However, the appearance of ''enjoy'' is not in our expectation. We investigate some threads and find that ''enjoy'' is often used to be the end of content by the authors.

TABLE 6. The 12-dimensional feature vector contains a topic distribution probability vector and a DBID statistical nature vector.

No.	Type
Feature 1	The probability of Topic 1
Feature 2	The probability of Topic 2
Feature 3	The probability of Topic 3
Feature 4	The probability of Topic 4
Feature 5	The probability of Topic 5
Feature 6	The probability of Topic 6
Feature 7	The number of domains (dbdomaindb) in the thread
Feature 8	The number of filenames (dbfiledb) in the thread
Feature 9	The number of file sizes (dbsizedb) in the thread
Feature 10	The number of quantity identification words (dbcountdb) in the thread
Feature 11	The number of account-password pairs (dbaccpwddb) in the thread
Feature 12	The number of email-password pairs (dbemailpwddb) in the thread

D. TEXT CLASSIFICATION

In general, each thread is a sparse vector. And it contains redundant information and noise in the original high-dimensional space. The topic model generates the topic distribution and preserves the semantic features of the original thread. This achieves the goal of reducing the thread's dimension. Besides, as we described in section IV. B, the data breach threads have distinctive features (DBID). Although we consider DBIDs when training the topic model, the number of each DBID is a crucial feature for the classification. By applying the DBID counter to calculate the number of DBIDs in each thread, we generate a statistical nature vector. And we combine the topic distribution probability vector and the statistical nature vector to form the feature vector of a thread. Table 6 shows the detailed component of a feature vector.

To find the most suitable classification algorithm for our research, we choose five typical classification algorithms, i.e., Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbor(K-NN) for comparison experiments. We calculate three metrics (precision, recall, and f1-score) to evaluate the performance. Identifying more threads about data breaches is what we expect. Thus, as for this experiment, recall is the most effective measure of this model. To maximize the classification performance of each algorithm, we adopt the method of cross-validation based on grid search.

Figure 5 compares the precision, recall and F1 score which are from the analysis of 5 algorithms. From the graph above we can see that the Naive Bayes model has the highest precision. However, it is bad at retrieving over the total amount of relevant instances. If there is a large number of missing

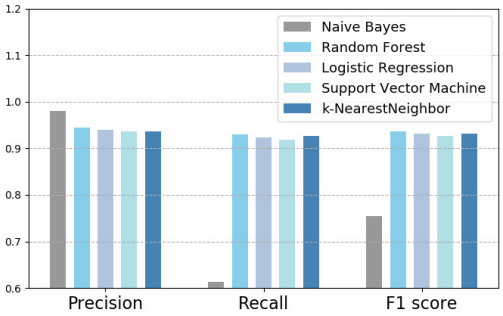

FIGURE 5. Performance comparisons of five algorithms for the classifier.

reports, the compromised party will not be able to deal with the breach in time. For this purpose, we need a model with high recall. The Random Forest model performs well in terms of precision and recall. And it has the highest F1 score. Therefore, we choose the Random Forest as the classification algorithm for the classifier.

E. COMPARISION

In this part, we compare our method with other different methods by 10-fold cross-validations. Based on the same data set, we choose the LDA-SVM and Word2vec [31]-SVM, which are widely used in research on underground forums and have a good effect on the classification of threads, for comparison. In this set of experiments, the precision, recall, and the F1 score are used as standard evaluation metrics.

TABLE 7. Evaluation of different methods.

Method	Prec	Rec	F1
LDA-SVM	91.9%	91.2%	91.6%
Word2vec-SVM	89.8%	75.1%	81.8%
Our method	**94.4%**	**92.9%**	**93.6%**

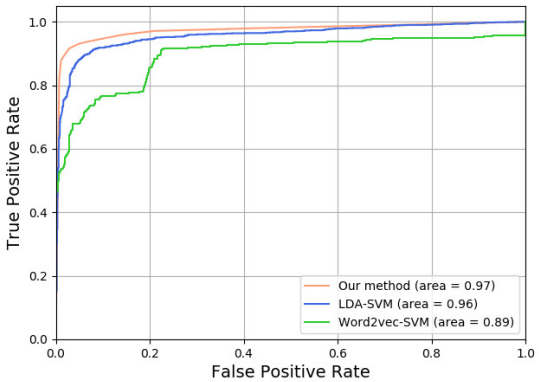

FIGURE 6. ROC curves of different methods.

The experimental results are presented in Table 7. As we can see, our method achieves 94.4% precision, 92.9% recall and 93.6% F1 score. The three indicators of the other two methods are lower than ours. And for evaluating the stability of the three models, the receiver operating characteristic (ROC) curves which are based on the 10-fold cross-validations are shown in Figure 6. Clearly, our method's

ROC curve has the highest AUC value. According to these two figures, our method outperforms the other two methods in data breach identification. The reason for this result is that, in our method, we build more expressive feature vectors for threads. We not only focus on subjects of threads but also expressions related to data breaches.

V. CONCLUSION AND FUTURE WORK

Underground forums have gradually become the first choice for hackers to announce data breaches. At the same time, more and more people choose to share and trade the leaked data they have in the underground forum. Analyzing and identifying these threads makes compromised third parties understand their damage early and respond to incidents in time. Similarly, personal information disclosure can also be noticed in this way.

This study proposes a system to identify data breach threads in underground forums. To ensure the authenticity and diversity of the experimental data, both the forums on the surface web and the dark web are the data collections. And the wording characteristics of data breach threads are found by the LDA model. Besides, we combine LDA topic probability distribution and statistical nature into more expressive feature vectors. To get better results, we compare five supervised learning algorithms and select Random Forest for classifier training. Finally, our system can identify more than 92% of data breach threads.

Our system aims to identify whether a thread in an underground forum is talking about data breaches. It cannot identify the authenticity of the thread. In other words, our system is difficult to detect whether a thread is a rumor. This may increase the time cost for responding to data breaches. In the future work, our concern will focus on the authenticity of data breach threads, and propose a method for quantitatively measuring the severity of those breaches. At the same time, identifying threads in different languages is also a problem that we will consider in the future.

REFERENCES

[1] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 51–63, Jan. 2011.

[2] S. A. Kale and S. Kulkarni, "Data leakage detection," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 1, no. 9, pp. 668–678, 2012.

[3] M. Lu, P. Chang, J. Li, T. Fan, and W. Zhu, "Data leakage prevention for resource limited device," U.S. Patent 8 286 253 B1, Oct. 9, 2012.

[4] T. G. Brown and B. S. Mann, "System and method for data leakage prevention," U.S. Patent 8 578 504 B2, Nov. 5, 2013.

[5] G. Katz, Y. Elovici, and B. Shapira, "CoBan: A context based model for data leakage prevention," *Inf. Sci.*, vol. 262, pp. 137–158, Mar. 2014.

[6] J. Onaolapo, E. Mariconti, and G. Stringhini, "What happens after you are PWND: Understanding the use of leaked Webmail credentials in the wild," in *Proc. Internet Meas. Conf.*, 2016, pp. 65–79.

[7] D. Jaeger, H. Graupner, A. Sapegin, F. Cheng, and C. Meinel, "Gathering and analyzing identity leaks for security awareness," in *Proc. Int. Conf. Passwords*. Cham, Switzerland: Springer, 2014, pp. 102–115.

[8] K. Thomas *et al.*, "Data breaches, phishing, or malware?: Understanding the risks of stolen credentials," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 1421–1434.

[9] X. Shu, K. Tian, A. Ciambrone, and D. Yao. (2017). "Breaking the target: An analysis of target data breach and lessons learned." [Online]. Available: https://arxiv.org/abs/1701.04940

[10] B. Butler, B. Wardman, and N. Pratt, "REAPER: An automated, scalable solution for mass credential harvesting and OSINT," in *Proc. APWG Symp. Electron. Crime Res. (eCrime)*, Jun. 2016, pp. 1–10.

[11] W. Li, J. Yin, and H. Chen, "Targeting key data breach services in underground supply chain," in *Proc. IEEE Conf. Intell. Secur. Inform. (ISI)*, Sep. 2016, pp. 322–324.

[12] H. Fallmann, G. Wondracek, and C. Platzer, "Covertly probing underground economy marketplaces," in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment*. Berlin, Germany: Springer, 2010, pp. 101–110.

[13] J. Franklin, A. Perrig, V. Paxson, and S. Savage, "An inquiry into the nature and causes of the wealth of Internet miscreants," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2007, pp. 375–388.

[14] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 71–80.

[15] R. Thomas and J. Martin, "The underground economy: Priceless," in *Proc. Mag. USENIX SAGE*, vol. 31, no. 6, 2006, pp. 7–16.

[16] K. Thomas *et al.*, "Framing dependencies introduced by underground commoditization," in *Proc. Workshop Econ. Inf. Secur.*, 2015.

[17] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing eve: Analysing cybercrime actors in a large underground forum," in *Proc. Int. Symp. Res. Attacks, Intrusions, Defenses*. Cham, Switzerland: Springer, 2018, pp. 207–227.

[18] R. Overdorf, C. Troncoso, R. Greenstadt, and D. McCoy. (2018). "Under the underground: Predicting private interactions in underground forums." [Online]. Available: https://arxiv.org/abs/1805.04494

[19] Y. Zhang, Y. Fan, S. Hou, J. Liu, Y. Ye, and T. Bourlai, "iDetector: Automate underground forum analysis based on heterogeneous information network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 1071–1078.

[20] R. S. Portnoff *et al.*, "Tools for automated analysis of cybercriminal markets," in *Proc. 26th Int. Conf. World Wide Web Steering Committee*, 2017, pp. 657–666.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[22] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Association for Computational Linguistics, vol. 1, 2009, pp. 248–256.

[23] S. Tasci and T. Gungor, "LDA-based keyword selection in text categorization," in *Proc. 24th Int. Symp. Comput. Inf. Sci. (ISCIS)*, Sep. 2009, pp. 230–235.

[24] L. Cui, F. Meng, Y. Shi, M. Li, and A. Liu, "A hierarchy method based on LDA and SVM for news classification," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Dec. 2014, pp. 60–64.

[25] Y. Wei, W. Wang, B. Wang, B. Yang, and Y. Liu, "A method for topic classification of Web pages using LDA-SVM model," in *Proc. Chin. Intell. Automat. Conf.* Singapore: Springer, 2017, pp. 589–596.

[26] D. Quercia, H. Askham, and J. Crowcroft, "TweetLDA: Supervised topic classification and link prediction in twitter," in *Proc. 4th Annu. ACM Web Sci. Conf.*, 2012, pp. 247–250.

[27] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 399–408.

[28] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2010, pp. 100–108.

[29] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Association for Computational Linguistics, 2011, pp. 262–272.

[30] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, Association for Computational Linguistics, 2012, pp. 952–961.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

YONG FANG received the Ph.D. degree from Sichuan University, Chengdu, China, in 2010, where he is currently a Professor with the College of Cybersecurity. His research interests include network security, Web security, the Internet of Things, big data, and artificial intelligence.

YUSONG GUO received the B.Eng. degree in information security from Sichuan University, Chengdu, China, in 2017, where he is currently pursuing the M.A. degree with the College of Cybersecurity. His current research interests include Web security and artificial intelligence.

CHENG HUANG received the Ph.D. degree from Sichuan University, Chengdu, China, in 2017. From 2014 to 2015, he was a Visiting Student with the School of Computer Science, University of California, CA, USA. He is currently an Assistant Research Professor with the College of Cybersecurity, Sichuan University, Chengdu, China. His current research interests include Web security, span networks, social privacy, system security, and artificial intelligence.

LIANG LIU received the M.A. degree from Sichuan University, Chengdu, China, in 2010, where he is currently an Assistant Professor with the College of Cybersecurity. His current research interests include malicious detection, network security, system security, and artificial intelligence.