ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200299

CIDetector: Semi-Supervised Method for Multi-Topic Confidential Information Detection

Jianguo Jiang [1,2] and **Yue Lu** [1,2] and **Min Yu** [1,2,*] and **Gang Li** [3,4,*] and **Yantao Jia** [5] and **Jiafeng Guo** [6] and **Chao Liu** [1] and **Weiqing Huang** [1]
* corresponding author

Abstract. Confidential information firewalling with text classifier is to identify the text containing confidential information whose publication might be harmful to national security, business trade, or personal life. Traditional methods, e.g., listing a set of suspicious keywords together with regular-expression based filter, fail to solve the multi-topic phenomenon, i.e., one text containing the confidential information with different topics. In this paper, we propose a semi-supervised method, CIDetector, for multi-topic confidential information detection. We introduce coarse confidential polarity as prior knowledge into word embeddings, which can regularize the distribution of words to have a clear task classification boundary. Then we introduce a multi-attention network classifier to extract task-related features and model dependencies between features for multi-topic classification. Experiments are conducted by real-world data from WikiLeaks and demonstrated the superiority of our proposed method.

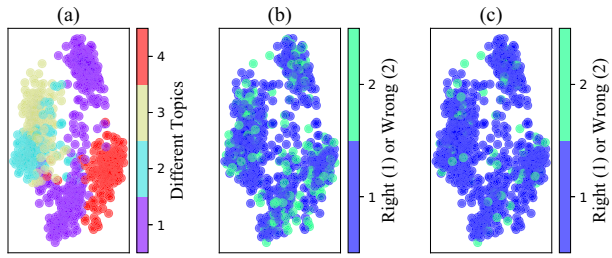

Figure 1: These diagrams show the word embedding visualization on labeled data using PCA. (a) shows the vectors of existing word embedding on the same topic coming together. (b) shows words with wrong (green) and right (blue) polarity in existing word embeddings. (c) shows wrong (green) and right (blue) polarity of words in our word embedding after introducing coarse confidential polarity.

1 INTRODUCTION

To prevent inadvertent disclosure of confidential information, many companies have been forced to institute strong policies on the email being sent from company servers [23]. In some cases, these policies require every outgoing email from an employee to be reviewed by that employees-manager before the email is released from the internal server to the Internet. In other cases, enforcement is purely reactive, but punishing a "leaking" employee doesn't un-leak the information that was released; the damage is still done. The optimal defense is to automatically detect confidential information and enforce the appropriate protection mechanism without degrading services or daily tasks.

The research community has proposed many methods for this task [3][7][10]. Most of the previous works are artificially specify a topic as the confidential class [2] or evaluated on a single topic or a few similar topics [6]. However, in reality, millions of text data generated daily often focus on disparate topics. The data could be collected from the real-world and the topics of the data could be diverse. Therefore, it is important and urgent to explore how to build a detector for multi-topic confidential information detection.

Existing methods cannot be used for multi-topic confidential information detection. Because the number of topics is large, it is unrealistic to train detectors on each topic and apply them together to detect multi-topic confidential information. Furthermore, the combination of single-topic detectors is not valid on unseen topics without introducing knowledge [12]. Existing word embeddings, such as Word2Vec [17] and GloVe [19], result in words with similar vectors but with opposite confidential polarities (as shown in Figure 1(a)(b)). Contexts alone are insufficient to achieve success in the absence of polarity-related features. Existing language models, such as BERT [8][15], achieve success on most of NLP tasks compared to word embeddings. But the available pre-trained BERT performs poorly (as shown in experiments) on confidential information detection because of the unlabeled corpus. Furthermore, there is not enough unlabeled confidential corpus to drive a new BERT.

Different topics have different levels of confidentiality. Introducing such coarse confidential polarity into word embeddings can regularize the distribution of words to have a relatively clear task classification boundary (as shown in Figure 1(c)), hence improving the confidential information detection performance. Driven by this motivation, we propose a semi-supervised method, CIDetector, for multi-topic confidential information detection. We introduce coarse confidential polarity as prior knowledge into a word embedding by topic-bridges. To take advantage of our confidentiality-oriented word embedding, we also introduce a multi-attention network classifier. We conduct the comparison experiments of our proposed method and the baseline

[1] Institute of Information Engineering, Chinese Academy of Sciences, China. Email: *yumin@iie.ac.cn*
[2] School of Cyber Security, University of Chinese Academy of Sciences, China
[3] School of Information Technology, Deakin University, VIC, Australia. Email: *gang.li@deakin.edu.au* ORCID: *0000-0003-1583-641X*
[4] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China
[5] Huawei Technologies Co., Ltd, China
[6] Institute of Computing Technology, Chinese Academy of Sciences, China

methods, using real-world data from WikiLeaks. The results show that our proposed method has a better performance in detecting confidential information. Moreover, our proposed method is still superior to the existing methods in the cases of a small training set. The main contributions of this work include:

1) To the best of knowledge, our work is the first to detect multi-topic confidential information, especially in the case of a few labeled data.
2) We propose a confidentiality-oriented word embedding by introducing coarse confidential polarity as prior knowledge and introduce a multi-attention classifier for multi-topic confidential information detection.
3) Extensive experiments using real-world data from WikiLeaks show the superiority of our proposed method. Compared with the baseline, our model improve F_1 by 7%.

The remainder of this paper is organized as follows: Section 2 gives a brief overview on the related work. Section 3 presents the details of our proposed method. Section 4 demonstrates the effectiveness of our proposed method with experiments. Finally, concluding remarks are offered in Section 5.

2 RELATED WORK

We review the existing researches from two perspectives: development of confidential information detection and methods for confidential information detection.

2.1 Development of Detection

From the development of confidential information detection, there are three main phases so far: symbol features, semantic features, and knowledge features.

Building detectors through traditional machine learning based on symbol features are first used. In 2011, Hart et al. used SVM and Naive Bayes (NB) based on TF-IDF to build a detector [9]. The significant drawback of symbol features is that it is unable to capture the relation between the words. In previous years, researchers used deep learning based on semantic features. Semantic features such as Word2Vec [18], BERT [8] have received a great deal of attention for their ability to model semantic similarity. Alzhrani et al. proposed a CNN based on pre-trained Word2Vec [4]. Unfortunately, large language models like BERT do not perform well on confidential information detection as discussed in the Introduction.

Recently researchers have begun to introduce prior knowledge into methods. In 2017, Yu et al. utilized sentiment dictionaries to refined word embedding to avoid generating similar word embedding vectors for sentimentally opposite words [24]. In 2018, Khosla et al. proposed Aff2Vec, an enriched word embeddings by a corpus of psychology and emotion for sentiment analysis [13]. In 2019, Akhtar et al. proposed an approach that learns a joint-representation on the labeled text and video through multi-task Learning [1]. However, confidential information detection does not have exact prior knowledge available. How to improve detectors in the case of limited data and inexact prior knowledge becomes critical.

2.2 Methods for Detection

There are two main types of methods for confidential information detection: supervised learning and semi-supervised learning.

Supervised learning constructs a predictive model by learning from a large number of training instances, where each training instance has a label indicating its ground-truth output. Wulczyn et al. developed a method that combines crowdsourcing and supervised models to analyze person attacks [22]. They annotated toxic information via crowdsourcing.

Though supervised learning has achieved great success, it is noteworthy that in the confidential information detection task it is difficult to get strong supervision information like fully ground-truth labels due to the high cost of the data-labeling process. Semi-supervised learning concerns the situation in which we are given a small amount of labeled data, which is insufficient to train a good classifier, while abundant unlabeled data are available. Karisani et al. proposed a new method, Word Embedding Space Partitioning and Distortion, to detect personal health mentions in social data [11]. Lee et al. explored semi-supervision for automatic classification of Adverse Drug Events tweets [14]. After referring to similar works, we also build our detector based on semi-supervised learning. But we are introducing coarse confidential polarity, which is a challenge and a difference from previous works.

3 METHODOLOGY

In this section, we present the technical details of our proposed method. First, we present the problem definition and notations. Then we cover the overview of CIDetector. Finally, we detail each component of CIDetector.

3.1 Problem Definition and Notations

Confidential information detection aims to classify text sequence into various confidential classes. The confidential class of information can be assessed based on the impact that might result from its leakage. The classes in this paper are Confidential and Non-Confidential. The "Confidential" class is assigned to information that would lead to damage to the national security. While the "Non-Confidential" class can be restricted to specifically authorized officials or released to the public without any damage to national security.

A text sequence may contain multiple topics but only belong to one confidential class. Confidential classes are related to topic information, but it is not possible to predict the confidential class directly based on topic information. We dig the relationship between topic information and confidential classes on the labeled corpus and introduce it as prior knowledge into word embeddings. So we need a labeled corpus with confidential classes and topics. To learn our confidentiality-oriented word embedding we also need a relatively large unlabeled corpus. We collect the unlabeled corpus through topic information in PlusD. To sum up, given the unlabeled corpus P and the labeled corpus Q with confidential classes K and topics T, our aim is to learn our Topic-Bridged Confidential Word Embedding (**TBC2Vec**) vectors $\mathbf{E} \in \mathbb{R}^{v \times d}$, where v is the size of vocabulary and d is the dimension of the word. On the other hand, to take advantage of our confidentiality-oriented word embedding we introduce **Transformer** [21] for multi-topic confidential information detection. Since the text sequence contains multiple topics, we utilize an adaptive attention as the Local Attention (**L-Attn**) to extract features. Then we utilize another attention as the Global Attention (**G-Attn**) to model dependencies between features. Hereafter we use the notation given in Table 1.

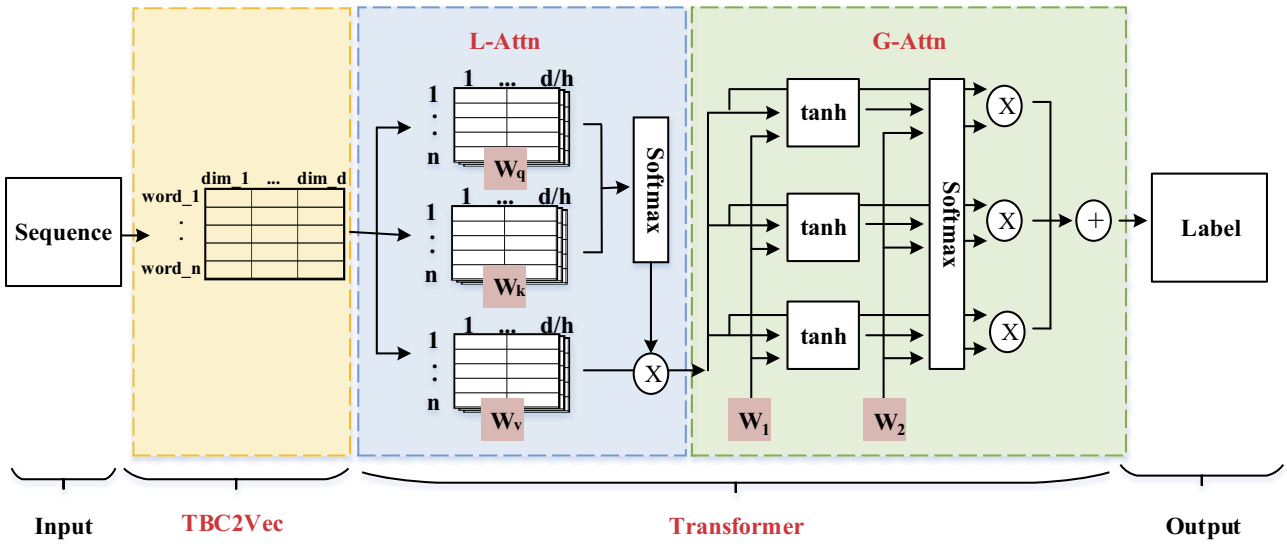

Figure 2: Illustration of the CIDetector.

Table 1: This table lists notation declarations.

Symbol	Meaning
$P = \{p_1, p_2, \ldots, p_t\}$	unlabeled corpus with topics
$Q = \{q_1, q_2, \ldots, q_t\}$	labeled corpus with topics
$Q = \{q_1, q_2, \ldots, q_k\}$	labeled corpus with confidential classes
$T = \{t_1, t_2, \ldots, t_i\}$	all topics
$K = \{k_1, k_2, \ldots, k_j\}$	all confidential classes
$\widetilde{T} = \{\widetilde{t}_1, \ldots, \widetilde{t}_m\} \subset T$	all topic-bridges
$W = \{w_1, w_2, \ldots, w_v\}$	all words
$\widetilde{W} = \{\widetilde{w}_1, \ldots, \widetilde{w}_n\} \subset W$	words having the ability to distinguish each topic-bridge

3.2 An Overview of CIDetector

As shown in Figure 2, the CIDetector consists of two components: TBC2Vec and Transformer. We learn TBC2Vec separately and obtain the word embedding vectors to initialize the word embedding layer. The TBC2Vec provides a solid foundation for Transformer — a well pre-trained word embedding with a relatively clear task classification boundary. The Transformer is based on TBC2Vec and used to extract confidential features and model dependencies between features.

The input of CIDetector is the sequence to be detected, and the output is its confidential label. Then we get the sequence vectors through look-up operation on pre-trained TBC2Vec vectors. To make use of the order of the sequence, we must inject information about the relative or absolute position of the words in the sequence. The positional embedding vectors have the same dimension as the pre-trained TBC2Vec vectors so that the two can be summed. We get the final word embedding vectors by adding the pre-trained TBC2Vec vectors and the positional embedding vectors together. The final word embedding vectors are then sent to the Transformer. We introduce the Transformer to use inputs and outputs of the same shape to make it possible to stack multiple layers. Multiple layers will have a good performance on a long sequence. A Transformer consists of two sublayers: L-Attn Sublayer and G-Attn Sublayer. By L-Attn Sublayer, the detector can capture which words are more informative in a local window of words. By G-Attn Sublayer, dependencies between the fine confidential features can be modeled. Finally, we utilize a fully-connected layer to output the label of the sequence.

3.3 Learning TBC2Vec

3.3.1 *Confidentiality-Aware Part*

Different topics have different levels of confidentiality, which is the coarse confidential polarity for multi-topic confidential information detection and the motivation of our work. The topics we named as topic-bridges are like bridges to connect words with confidential polarity and have the ability to guide confidential polarity prediction. These topic-bridges include: ① on such topics there is a lot of data, which accounts for a large portion of total data. ② on such topics the data is imbalanced at different confidential classes, where the number of data is different at each of confidential classes. Formally, we select topic-bridges according to the maximal statistical dependency criterion based on Information Gain (*IG*). Given two random variables topic t and confidential class k, the *IG* is defined in terms of their probability functions $\Pr(t)$, $\Pr(k)$ and $\Pr(k|t)$:

$$\begin{aligned} IG(t, K) = &- \sum_{a=1}^{j} \Pr(k_a) \log_2 \Pr(k_a) \\ &+ \Pr(t) \sum_{a=1}^{j} \Pr(k_a|t) \log_2 \Pr(k_a|t) \\ &+ \Pr(\bar{t}) \sum_{a=1}^{j} \Pr(k_a|\bar{t}) \log_2 \Pr(k_a|\bar{t}) \,, \end{aligned} \tag{1}$$

where j is the size of topics.

The selected topic t is required, individually, to have the largest $IG(t, K)$ with the confidential classes K, reflecting the largest dependency on the confidential classification. According to the value $IG(t, K) > \alpha$ and which confidential class k the topic t trend to fall into, we generate a salience topic-bridge set for each confidential classes $\widetilde{T}$.

This also requires to determine which word to receive the coarse confidential polarity. These words $\widetilde{W}$ should have the ability to distinguish each topic-bridge. Each word $\widetilde{w} \in \widetilde{W}$ we designed is offline extracted according to the following the two principles: ① The term frequency of the word in one topic-bridge is much higher than that in the other topic-bridges; ② The word is common in other topic-bridges, expressed as a small variance of term frequencies in other topic-bridges. Formally, we design the following formula to measure the importance of the word to the topic-bridge:

$$Score(\widetilde{w}, \widetilde{t}) = \frac{tf_{\widetilde{t}}(\widetilde{w}) - avg(TF(\widetilde{w}))}{var(TF_{\widetilde{-t}}(\widetilde{w}))}\,, \tag{2}$$

where $tf_{\widetilde{t}}(\widetilde{w})$ is the term frequency of the word $\widetilde{w}$ in the topic-bridge $\widetilde{t}$. $TF(\widetilde{w})$ is the collective of term frequencies, and $avg(\,)$ is the average. $TF_{\widetilde{-t}}(\widetilde{w})$ is the collective of term frequencies except the topic-bridge $\widetilde{t}$, and $var(\,)$ is the variance. According to this importance score, we generate a word set by selecting the top-N words for each topic-bridge.

For the task, the words $\widetilde{W}$ have the ability to distinguish different confidential classes. In the learning framework, if the predicted word is in the word set $\widetilde{W}$, the confidentiality-aware part will be activated. As to model the relationship between words and confidential classes, we expect to constrain words to be close to the words in the same confidential class and far away from the words in a different confidential class. According to this idea, we construct a set with word-pairs for each word $\widetilde{w}$. Each word-pair contains a positive word and a negative word. The positive words are randomly selected from the words which belong to the same confidential class with $\widetilde{w}$, and the negative words are randomly sampled from other confidential classes. We maximize a margin-based ranking criterion:

$$\mathcal{L}_{conf} = \sum_{(\widetilde{w}_{pos}, \widetilde{w}_{neg} \in \widetilde{W})} \left[\log \sigma(\widetilde{w} \cdot \widetilde{w}_{pos}) + \log \sigma(-\widetilde{w} \cdot \widetilde{w}_{neg})\right]\,, \tag{3}$$

where $\widetilde{W}$ is the word set, $\tilde{w}_{pos}$ is a positive word, $\tilde{w}_{neg}$ is a negative word $\sigma()$ is the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$. The objective function favors higher values of the similarity for positive word-pairs than for negative word-pairs.

3.3.2 Context-Aware Part

We utilize the context-aware part to capture semantic. Mikolov et al./ proposed Skip-Gram and Negative Sampling to learn word embedding vectors from a large-scale text corpus [18]. To simplify, we represent the objective:

$$\mathcal{L}_{cont} = \log \sigma(w \cdot w_c) + \sum_{(w_{neg} \in W)} \left[\log \sigma(-w \cdot w_{neg})\right]\,, \tag{4}$$

where w is the target word, w_c and w_{neg} are respectively the context word and the negative word through Negative Sampling, W is the set of all words.

3.3.3 Joint Learning

In our proposed method, the words' contextual information and coarse confidential polarity are inherently jointed to construct the TBC2Vec. We then obtain the following object function:

$$\mathcal{L}_{TBC2Vec} = \lambda\, \mathcal{L}_{cont} + (1-\lambda)\, \mathcal{L}_{conf}\,, \tag{5}$$

where λ is the combination parameter which balances the contribution of each part in the training process.

3.4 Integration into Transformer

3.4.1 L-Attn Sublayer

The L-Attn can focus on the coarse confidential features from TBC2Vec and extract them based on contextual information. It provides a set of weight vectors between words in a sequence. The L-Attn consists of multi-attention heads h working in parallel. It can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Given a sequence $S = \{x_1, \ldots, x_n\}$ of length n (padded where necessary), we represent the similarity between words on each head:

$$sim(t, c) = \mathbf{x}_t^{\mathbf{T}} \mathbf{W}_q^{\mathbf{T}} (\mathbf{W}_k \mathbf{x}_c + \mathbf{p}_{\mathbf{t-c}})\,, \tag{6}$$

where t and c are respectively the target word and the context word, $\mathbf{W}_k$ and $\mathbf{W}_q$ are the "key" and "query" matrices, and $\mathbf{p}_{\mathbf{t-c}}$ is the relative positional embedding vectors between the target word and the context word. The L-Attn attention weights $\mathbf{A}_{L\text{–}Attn}$ are then obtained by applying a softmax function on these similarities:

$$\mathbf{A}_{L\text{–}Attn}(t, c) = \frac{\exp(sim(t, c))}{\sum_{q=t-S}^{t-1} \exp(sim(t, q))}\,, \tag{7}$$

Finally, the head outputs a vector $\mathbf{h}_t$ by taking the average of the past representations weighted by their attention weights:

$$\mathbf{o}_t = \sum_{r=t-S}^{t-1} \mathbf{A}_{L\text{–}Attn}(t, c) \mathbf{W}_v \mathbf{x}_r\,, \tag{8}$$

where $\mathbf{W}_v$ is called the "value" matrix. Outputs from different heads are then concatenated together and multiplied by an output matrix $\mathbf{O}$ before feeding to the next layer.

For each head, we add a masking function to control for the span of the attention [20]. The attention weights from Equation (7) are then computed on the masked span:

$$\mathbf{A}_{L\text{–}Attn}(t, c) = \frac{m_z(t-r)\exp(sim(t, c))}{\sum_{q=t-S}^{t-1} m_z(t-q)\exp(sim(t, q))}\,, \tag{9}$$

We add a ℓ_1 penalization on the parameters z_i for each attention head i of the model to the loss function:

$$\mathcal{L}_{L\text{–}Attn} = -\log \Pr(w_1, \ldots, w_T) + \frac{\theta}{M} \sum_i z_i\,, \tag{10}$$

where $\theta > 0$ is the regularization hyper-parameter, T is the set of target words, and M is the number of heads. Our formulation is differentiable in the parameters z_i and we learn them jointly with the rest of the Transformer.

3.4.2 G-Attn Sublayer

The G-Attn provides a set of summation weight vectors for the different confidential features of the input sequence. It takes the output of L-Attn as input, and learns the matrix of G-Attn attention $\mathbf{A}_{G\text{–}Attn}$ as:

$$\mathbf{A}_{G\text{–}Attn}(\mathbf{O}_{L\text{–}Attn}) = softmax(\mathbf{W_2} \tanh(\mathbf{W_1} \mathbf{O}_{L\text{–}Attn}^{\mathbf{T}}))\,. \tag{11}$$

Here $\mathbf{O}_{L\text{–}Attn}$ is the output of L-Attn, and $\mathbf{W_1}$ is a weight matrix with a shape of d_{ff}-by-d and $\mathbf{W_2}$ is a weight matrix with a shape of n-by-d_{ff}. We can deem Equation (11) as a 2-layer MLP without bias, whose hidden unit numbers is d_{ff}, and parameters are $\{\mathbf{W_1}, \mathbf{W_2}\}$. The $softmax(\,)$ ensures all the computed weights sum up to 1.

A single output vector of G-Attn usually focuses on a specific dependency of the confidential features. To represent the overall confidential dependencies of the sequence, we need multiple that focus on different dependencies of the sequence. Thus we need to perform multiple hops of attention. We computer the weighted sums by multiplying the annotation matrix $\mathbf{A}_{G\text{–}Attn}$ and confidential features $\mathbf{O}_{L\text{–}Attn}$:

$$\mathbf{O}_{G\text{–}Attn} = \mathbf{A}_{G\text{–}Attn}\mathbf{O}_{L\text{–}Attn}\,. \tag{12}$$

4 EXPERIMENTAL DESIGN AND ANALYSIS

In this section, we evaluate the various models on WikiLeaks Cable Dataset. The purpose of these experiments is to clarify the influence of CIDetector for multi-topic confidential information detection.

4.1 Experiment Design

4.1.1 Dataset

There are few published datasets available for confidential information detection. We make public and provide an overview of our new dataset. Our new dataset consists of paragraphs extracted from WikiLeaks Public Library of US Diplomacy (PlusD). We refer to the dataset as WikiLeaks Cable Dataset. The WikiLeaks Cable Dataset contains 89,681 paragraph-level instances across 24 topics.[7] We collect the data in the PlusD but not in our dataset as unlabeled corpus. We use white space as a delimiter, normalize punctuations, remove special characters and convert the remaining characters to lowercase. After pre-processing, the details on our dataset are provided in Table 2.

Table 2: Statistics of WikiLeaks Cable Dataset.

WikiLeaks Cable Dataset	
type	Paragraph
# classes	2
# Confidential instances	48,302
# Non-Confidential instances	41,379
average paragraph length	145
vocabulary size	125,534

4.1.2 Models

We compare our proposed model with the following models.

- **TF-IDF + NB, TF-IDF + LR, TF-IDF + SVM** [5][9]: Hart et al. proposed a supervised Naive Bayes and Support Vector Machine based on TF-IDF for data loss prevention. Alzhrani et al. proposed a Logistic Regression based on TF-IDF for data leak prevention. We use the sklearn to implement and keep the default parameters.
- **BERT-Large-Cased** [8]: Devilin et al. introduced a new language representation model named BERT. The pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. We fine-tuned the pre-trained BERT for confidential information detection.
- **GloVe + CNN, Word2Vec + CNN** [4]: Alzhrani proposed a CNN based on Word2Vec for confidential information detection. We reproduce his model and use the same hyper-parameter setting. GloVe + CNN is the CNN with pre-trained word embedding vectors learned through GloVe. Word2Vec + CNN is the CNN with pre-trained word embedding vectors learned through Word2Vec. The pre-trained word embedding vectors are fine-tuning during the training of classifiers.
- **GloVe + LSTM, Word2Vec + LSTM** [16]: Ma et al. explored LSTM for rumor analysis. We reproduce their model and use the same hyper-parameter setting. These models are similar to the above models, only replacing CNN with LSTM. The pre-trained word embedding vectors are fine-tuning during the training of classifiers.
- **GloVe + Transformer, Word2Vec + Transformer (Baseline)** [21]: Vaswani et al. proposed a multi-attention model for text classification. It is the most advanced model for text classification. We reproduce their model and use the same hyper-parameter setting. These models are similar to the above models, only replacing GRU with Transformer. The pre-trained word embedding vectors are fine-tuning during the training of classifiers.

4.1.3 Hyper-Parameters

We tune the hyper-parameters of our proposed model on the validation set. And we do not otherwise perform any dataset-specific tuning other than early stopping on the validation set.

- **Hyper-Parameters in TBC2Vec**: We set the the combination parameter $\lambda = 0.8$ to balance the confidentiality-aware part and the context-aware part. We set $\alpha = 0.002$ to select topic-bridges and we select top-1000 to select words. We set the number of positive words as 2 and the number of negative words as 2 in the confidentiality-aware part. We set the number of negative words as 4 in the context-aware part.
- **Hyper-Parameters in Transformer**: We set the size of the word embedding as $d = 512$. We set $h = 8$ for multi-projection in L-Attn and set G-Attn with $d_{ff} = 2048$ units. We train the network with a mini-batch size of 64 and a learning rate of 0.001 by back-propagation and the gradient-based optimization is performed using the Adam update rule. We use a 0.5 dropout rate on the fully-connected layer during training.

4.1.4 Evaluation Setup

We use F_1, and $AUPRC$ to measure these models because there is a slight imbalance in the dataset. The F_1 reflect the ability to identify the confidential information. The higher of F_1, the better of the model for confidential information detection. The $AUPRC$ is short for Area Under Precision-Recall Curve. It can help us analyze the models when F_1 is the same value.

4.2 Experiment Analysis

4.2.1 Main Comparisions

On multi-topic confidential information detection, there are three types of topics: seen topics, unseen topics that are strongly associated with seen topics, and unseen topics that are weakly associated with seen topics.

Seen topics are the data trained and tested on the same topics. We use 10-fold cross-validation and re-partition the dataset into training,

[7] For all acronyms in the dataset, we have looked for an official source as a reference to the meaning of that acronym — *https://wikileaks.org/plusd/about-ta/*

Table 3: Results of seen topics.

Type	Model	Unlabeled Corpus	$F_1(\%)$	$AUPRC(\%)$
Supervision	TF-IDF + NB	-	67.27	69.08
	TF-IDF + LR	-	66.45	68.69
	TF-IDF + SVM	-	67.62	69.46
Semi-Supervision	BERT-Large-Cased	-	65.14	-
	GloVe + CNN	Twitter	70.56	74.12
	Word2Vec + CNN	GoogleNews	70.85	74.82
	GloVe + CNN	PlusD	71.70	75.71
	Word2Vec + CNN	PlusD	72.17	76.19
	GloVe + LSTM	Twitter	69.98	73.38
	Word2Vec + LSTM	GoogleNews	70.61	74.08
	GloVe + LSTM	PlusD	71.03	75.05
	Word2Vec + LSTM	PlusD	71.25	75.43
	GloVe + Transformer	Twitter	73.51	77.91
	Word2Vec + Transformer	GoogleNews	73.79	78.30
	GloVe + Transformer	PlusD	74.86	79.37
	Word2Vec + Transformer (Baseline)	**PlusD**	**75.31**	**79.84**
	TBC2Vec + CNN	PlusD	78.13	85.53
	TBC2Vec + LSTM	PlusD	77.85	85.13
	TBC2Vec + Transformer (Ours)	**PlusD**	**82.49**	**90.07**

validation, and test set. These partitions correspond to 80%, 10%, and 10% of the original dataset. The dataset is separated into 71,747 training instances, 8,967 validation instances, and 8,967 test instances. From Table 3, we have the following observations: (1) As expected, our proposed model performs significantly better than the other models in F_1 and $AUPRC$ at the same time. Compared with the baseline model, our proposed model improve F_1 by 7% and $AUPRC$ by 10%. Our proposed model outperforms the best previously reported model, yielding a new state-of-the-art. It demonstrates introducing prior confidential polarity into semantic-oriented word embeddings can help improve detection capability. (2) The semi-supervised models based on Transformer are better than the semi-supervised models based on CNN and LSTM. For instance, TBC2Vec + Transformer is better than TBC2Vec + CNN and TBC2Vec + LSTM. It indicates the classification ability of Transformer is better than that of CNN and LSTM. (3) The semi-supervised models with PlusD are better than the semi-supervised models with GoogleNews and Twitter. For instance, Word2Vec(PlusD) + CNN is better than Word2Vec(GoogleNews) + CNN. It shows the importance of the corpus domain as discussed in Section 1. (4) The semi-supervised deep learning models are better than the traditional machine learning models. For instance, semi-supervised LSTMs are better than TF-IDF + SVM. It shows the effectiveness of semi-supervised models based on neural networks.

Table 4: Results of unseen topics that are strongly associated with seen topics.

Topic	Model	$F_1(\%)$
MASS	Word2Vec + Transformer (Baseline)	74.91
	TBC2Vec + CNN	76.68
	TBC2Vec + LSTM	76.12
	TBC2Vec + Transformer (Ours)	**81.36**
PINS	Word2Vec + Transformer (Baseline)	74.82
	TBC2Vec + CNN	76.60
	TBC2Vec + LSTM	76.08
	TBC2Vec + Transformer (Ours)	**81.35**

Unseen topics that are strongly associated with seen topics are the unseen topics in the test set but having a strong association with the seen topics in the training set. We select top-2 such topics according to information gain. We use these topics — MASS and PINS — respectively as the evaluation set. We use the rest topics as the training set. Table 4 shows the performance of the models on such unseen topics. Our proposed model performs better than the baseline model. Besides, the results of all models are as good as their results in the seen topics. This is because these topics are similar to the topics of the training set. It shows that the capabilities of detectors are not greatly affected if the detectors encounter unseen topics that are strongly associated with seen topics.

Table 5: Results of unseen topics that are weakly associated with seen topics.

Topic	Model	$F_1(\%)$
SNAR	Word2Vec + Transformer (Baseline)	67.85
	TBC2Vec + CNN	72.60
	TBC2Vec + LSTM	70.03
	TBC2Vec + Transformer (Ours)	**78.55**
EAGR	Word2Vec + Transformer (Baseline)	67.79
	TBC2Vec + CNN	72.73
	TBC2Vec + LSTM	70.15
	TBC2Vec + Transformer (Ours)	**78.69**

Unseen topics that are weakly associated with seen topics are the unseen topics in test set but having weakly association with the seen topics in training set. We select bottom-2 such topics according to information gain. We use these topics — SNAR and EAGR — respectively as the evaluation set. We use the rest topics as the training set. Table 5 shows the performance of the neural network models on such unseen topics. The results of all models are not as good as their results on the seen topics. This is because these topics are different from the topics of training set. However, our proposed model is robust, least affected by these unseen topics, and better than the baseline model. For example, there is a Non-Confidential instance on

EAGR — *"In Ararat one farm had been affected and approximately 300 swine culled..."*. *Ararat* and *swine* are in the non-confidential word set of our word embedding. These two terms are selected by SENV (a topic in training set) to enter the non-confidential word set.

4.2.2 Ablation Study and Hyper-Parameter Sensitivity

We do an ablation study on our proposed model through removing each component or replacing our components with simple components. In addition, we explore the impact of two key hyper-parameters.

Table 6: Results of ablation study.

Model	$F_1(\%)$	$\Delta(\%)$
CIDetector (Ours)	82.49	0
CIDetector - TBC2Vec + Word2Vec	75.31	-7.18
CIDetector - L-Attn	80.57	-1.92
CIDetector - G-Attn	79.54	-2.95

From Table 6, we have the following observations: (1) Replacing the TBC2Vec with Word2Vec, the model performs worse, indicating the importance of word embedding having a clear task classification boundary. In other words, introducing prior confidential polarity into semantic-oriented word embeddings can help improve detection capability. (2) The model without L-Attn performs worse, indicating the effectiveness of fine-tuning the coarse confidential features and focus on them. (3) The model without G-Attn performs worse, indicating the effectiveness of modeling dependencies between features.

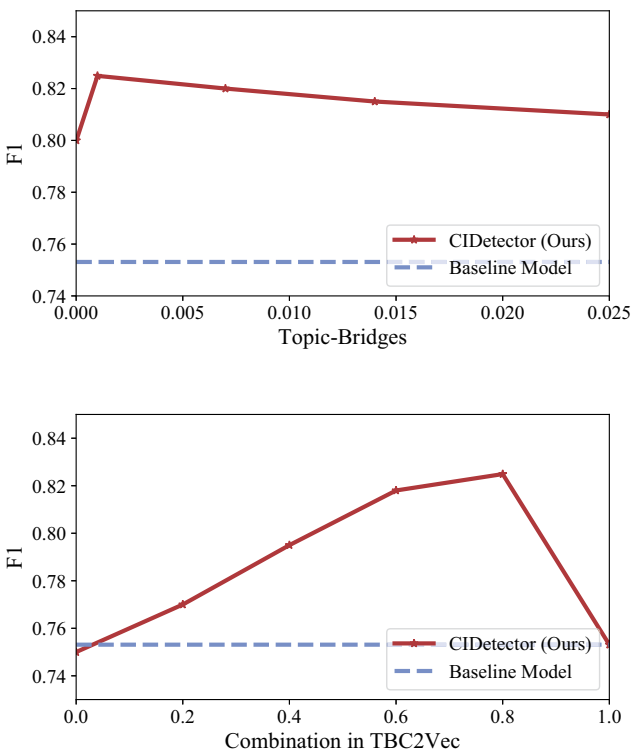

Figure 3: Results of the topic-bridges (α) and the TBC2Vec combination (λ).

From Figure 3, (1) we find the topic-bridges α affects the performance of our proposed model. A small value of α represents more coarse confidential polarity we introduced on topics. As α decreases, F_1 value increases first, then decreases. This result tells us to keep purity as high as possible while increasing the quantity of prior knowledge. (2) We find the combination in TBC2Vec λ affects the performance of our proposed model. A small value of λ represents the context-aware part is more important than the confidentiality-aware part. This result tells us not to destroy the original semantic information too much while learning the coarse confidential polarity. Replacing all the semantic information with coarse confidential polarity, our proposed model is a little worse than the baseline model.

4.3 Analysis on Various Training Set Sizes

In the real world, data is hard to be collected, especially confidential data. We set up an experiment to evaluate the quality of models on a small training set.

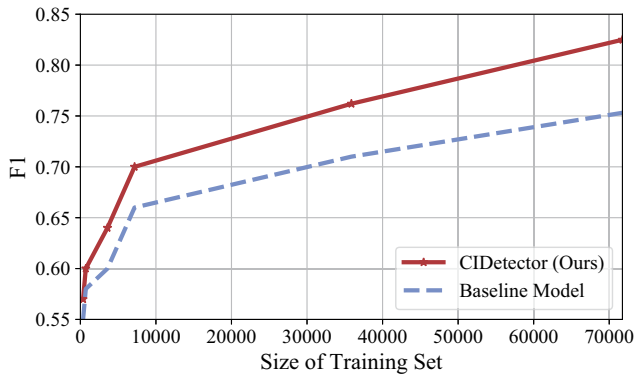

Figure 4: Results of Training Set Size.

Figure 4 shows the performance of our proposed model and baseline model on various sizes of the training set. As expected, both of them are highly dependent on the size of the training set. We observed that reducing the training set size hurts the performance of models. What is even more remarkable is that our proposed method performed better than the baseline model. To achieve the same capability of the baseline model, our proposed method requires only half of the training set according to the F_1 results. This instability of the baseline model is caused by the following two aspects: On the one hand, the baseline model lacks prior confidential knowledge when the training data is not enough. On the other hand, existing word embeddings cannot give a basic task-related support when the training data is enough.

5 CONCLUSION

In this paper, we proposed a semi-supervised method, CIDetector, for improving the performance of multi-topic confidential information detection. On the one hand, we designed the TBC2Vec by introducing coarse confidential polarity into word embedding in order to regularize the distribution of words to have a relatively clear task classification boundary; On the other hand, we introduced the Transformer by using attention mechanisms instead of CNNs/RNNs in order to extract confidential features and model dependencies between features. The results of the experiments on the WikiLeaks Cable Dataset show that our proposed method outperforms the existing state-of-the-art methods. In the future, we are going to explore what fine-grained confidential features or exact prior knowledge is.

ACKNOWLEDGEMENTS

This work is supported by National Key R&D Program of China (No.2018YFB0803402), Deakin Univesity ASL 2019, Xinjiang Science&Technical Research Fund 2018, and the work was completed when Gang Li was on ASL in Chinese Academy of Sciences.

REFERENCES

[1] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya, 'Multi-task learning for multi-modal emotion recognition and sentiment analysis', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*. Association for Computational Linguistics, (2019).

[2] Sultan Alneyadi, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy, 'A semantics-aware classification approach for data leakage prevention', in *Australasian Conference on Information Security and Privacy (ACISP'14)*, pp. 413–421. Springer, (2014).

[3] Erdem Alparslan and Hayretdin Bahsi, 'Security level classification of confidential documents written in turkish', in *International Conference on User Centric Media*, pp. 329–334. Springer, (2009).

[4] Khudran Alzhrani, Fahad Saud Alrasheedi, Faris Anwar Kateb, and Terrance E Boult, 'Cnn with paragraph to multi-sequence learning for sensitive text detection', in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS'19)*, pp. 1–6. IEEE, (2019).

[5] Khudran Alzhrani, Ethan M Rudd, C Edward Chow, and Terrance E Boult, 'Automated big security text pruning and classification', in *2016 IEEE International Conference on Big Data (BigData'16)*, pp. 3629–3637. IEEE, (2016).

[6] Khudran Alzhrani, Ethan M Rudd, C Edward Chow, and Terrance E Boult, 'Automated us diplomatic cables security classification: Topic model pruning vs. classification based on clusters', in *2017 IEEE International Symposium on Technologies for Homeland Security (HST'17)*, pp. 1–6. IEEE, (2017).

[7] Richard Chow, Philippe Golle, and Jessica Staddon, 'Detecting privacy leaks using corpus-based association rules', in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08)*, pp. 893–901. ACM, (2008).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*. Association for Computational Linguistics, (2019).

[9] Michael Hart, Pratyusa Manadhata, and Rob Johnson, 'Text classification for data loss prevention', in *International Symposium on Privacy Enhancing Technologies Symposium (PETS'11)*, pp. 18–37. Springer, (2011).

[10] Jianguo Jiang, Yue Lu, Min Yu, Gang Li, Chao Liu, Weiqing Huang, and Fangtao Zhang, 'Sentiment embedded semantic space for more accurate sentiment analysis', in *International Conference on Knowledge Science, Engineering and Management (KSEM'18)*, pp. 221–231. Springer, (2018).

[11] Payam Karisani and Eugene Agichtein, 'Did you really just have a heart attack? towards robust detection of personal health mentions in social media', in *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW'18)*, pp. 137–146. International World Wide Web Conferences Steering Committee, (2018).

[12] Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda, 'Maximal margin labeling for multi-topic text categorization', in *Advances in neural information processing systems*, pp. 649–656, (2005).

[13] Sopan Khosla, Niyati Chhaya, and Kushal Chawla, 'Aff2vec: Affect–enriched distributional word representations', *arXiv preprint arXiv:1805.07966*, (2018).

[14] Kathy Lee, Ashequl Qadir, Sadid A Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri, 'Adverse drug event detection in tweets with semi-supervised convolutional neural networks', in *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*, pp. 705–714. International World Wide Web Conferences Steering Committee, (2017).

[15] Chao Liu, Xinghua Wu, Min Yu, Gang Li, Jianguo Jiang, Weiqing Huang, and Xiang Lu, 'A two-stage model based on bert for short fake news detection', in *International Conference on Knowledge Science, Engineering and Management (KSEM'19)*, pp. 172–183. Springer, (2019).

[16] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha, 'Detecting rumors from microblogs with recurrent neural networks.', in *IJCAI'16*, pp. 3818–3824, (2016).

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', in *1st International Conference on Learning Representations (ICLR'13)*, (2013).

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances in neural information processing systems (NIPS'13)*, pp. 3111–3119, (2013).

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning, 'Glove: Global vectors for word representation', in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP'14)*, pp. 1532–1543, (2014).

[20] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin, 'Adaptive attention span in transformers', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, (2019).

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in neural information processing systems (NIPS'17)*, pp. 5998–6008, (2017).

[22] Ellery Wulczyn, Nithum Thain, and Lucas Dixon, 'Ex machina: Personal attacks seen at scale', in *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*, pp. 1391–1399. International World Wide Web Conferences Steering Committee, (2017).

[23] William Yerazunis, Mamoru Kato, Mitsunori Kori, Hideya Shibata, and Kurt Hackenberg, 'Keeping the good stuff in: Confidential information firewalling with the crm114 spam filter & text classifier', *White Paper for Black Hat USA*, (2010).

[24] Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang, 'Refining word embeddings for sentiment analysis', in *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP'17)*, pp. 534–539, (2017).