Anonymous Market Product Classification Based on Deep Learning

Lina Yang
Qilu University of Technology (Shandong Academy of Sciences)
Shandong Computer Science Center (National Supercomputer Center in Jinan)
Shandong Provincial Key Laboratory of Computer Networks, China
898246244@qq.com

Ying Yang*
Qilu University of Technology (Shandong Academy of Sciences)
Shandong Computer Science Center (National Supercomputer Center in Jinan)
Shandong Provincial Key Laboratory of Computer Networks, China
yangy@sdas.org

Huanhuan Yu
Qilu University of Technology (Shandong Academy of Sciences)
Shandong Computer Science Center (National Supercomputer Center in Jinan)
Shandong Provincial Key Laboratory of Computer Networks, China
1414449224@qq.com

Guichun Zhu
Qilu University of Technology (Shandong Academy of Sciences)
Shandong Computer Science Center (National Supercomputer Center in Jinan)
Shandong Provincial Key Laboratory of Computer Networks, China
1569729147@qq.com

ABSTRACT

With the rapid development of Internet technology, the abuse of dark networks and anonymous technology has brought great challenges to network supervision. Therefore, it is important to study the anonymous market. In this paper, we propose a single-mode multivariate classification model for anonymous market product classification. Divide anonymous markets products into 5 categories. Our algorithm uses the word vector embedded in a convolutional neural network based on Word2vec training. Compared with the simple machine learning classification model, the accuracy of the single-mode multivariate classification model on the test set is 91.84%. By studying the classification of anonymous market products, law enforcement personnel can better supervise anonymous market of illegal products and maintain network security.

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Natural language processing • Information extraction

KEYWORDS

Anonymous market, Convolutional neural network, Word2vec

AIIPCC '19, December 19–21, 2019, Sanya, China

ACM ISBN 978-1-4503-7633-4/19/12…$15.00
https://doi.org/10.1145/3371425.3371467

1 Introduction

The whole Internet consists of three parts. A page that is accessed through traditional keyword searches or hyperlinks, we call it a surface net [1] [2]. Deep web refers to services and products that cannot be included in regular search engines such as Google. It mainly refers to web pages that can be viewed after login, such as forums and blogs. The dark network refers to the network that can only be accessed through specific software or configuration. The most commonly used specific software is TOR. Darknet is a fast-growing market for sales of illegal products. In order to evade the law enforcement personnel's investigation and research on the dark network, dark network product suppliers deliberately misclassify the product and increase the difficulty of investigating and analysing illegal products. For example, products suppliers that sell pirated books are indexed to child pornography [3]. Without manual intervention, it is difficult for law enforcement officials to extract the correct relevant product information. To improve this process, we introduce an improved deep learning classification algorithm that automatically classifies anonymous trading market products based on product titles.

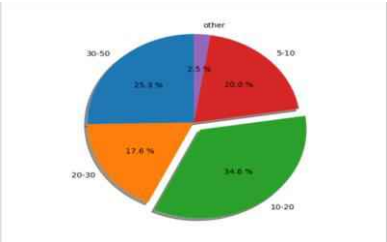

Figure 1: Anonymous Market Product Title Length Distribution.

According to statistics, the title of the anonymous trading market product is generally 10 to 50 characters long, as shown in Figure 1 [4] [5]. Therefore, we attribute the classification to a short text classification problem. This paper will propose a single-mode multivariate classification model, and try to incorporate deep learning into the short text classification architecture, which is compared with traditional machine learning classification.

2 Related Work

In the early days, most of the researches on dark nets focused on the data acquisition of the dark network [6] [7]. Hsinchu Chen designed and implemented a focused crawler in the literature [8] to capture and analyse the subject-related content. There are also some studies used to discover hidden services. Clement Guitton in the literature [9] based on the Hidden Wiki in the dark network, Snapp BBS to obtain the hidden service address, crawl and analyse the content of the hidden service.

There have also been several attempts to analyse the anonymous market of the dark network [10] [11]. In the paper，the author conducted an 8-month crawl on "silk road", analysed the product list and product distribution. However, their classification relies on the classification provided by the supplier. In order to avoid the investigation of law enforcement personnel, some suppliers deliberately misclassify the products, so such classification is not used in all anonymous trading markets.

Table 1: Product Title Classification Field.

Class	Description
Data	buy and sell personal infromation, account password,etc.
Porn	providing a variety of pornography. child pornography,etc.
Service class	providing various hacking services,password deciphering,etc.
Drug	auction of various drugs
Books	providing book transactions or pirated books

At present, commonly used machine learning algorithms include Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM) classification models. In machine learning, due to the different features selected, it may result in sparse features. Recently, Kim [12] achieved good results in short text sentiment classification by combining word vectors with convolutional neural networks. These works prove that convolutional neural networks are suitable for the field of natural language processing.

Classified according to the title of the product, the classification system needs to give each title sample a unique label for the domain category. In order to conduct research on the anonymous trading market, a one-month crawler was used, and anonymous market data was collected by onionscan. Products in the anonymous trading market were divided into five categories as shown in Table 1. In each field, we used 1,500 training samples, 200 verification samples, and 300 test samples. Precision, recall, and F1 are used for performance evaluation.

3 Traditional Classification Model

3.1 Data Preprocessing

There are a lot of noise characteristics in the original data corpus, especially for the dark network data, there are spelling errors, lack of content confusion, etc., which is very unfavorable to classify them directly. Because classifiers require a uniform format for classification, processing objects can improve classification efficiency. Therefore, in order to solve this problem, we need to carry out preparatory work before the classification, and pre-process the corpus required for all classification experiments. Data preprocessing is the process of segmenting the original data set, decommissioning words and noise words to prepare for the subsequent classification work.

Word segmentation is the basis for text categorization. Text segmentation is the separation of a series of strings into separate words. Chinese text segmentation is divided into dictionary-based and non-lexical-based word segmentation algorithms according to whether or not to introduce a dictionary. The three popular automatic word segmentation algorithms are based on string matching word segmentation algorithm, word segmentation algorithm based on semantic understanding, and word segmentation algorithm based on probability and statistics.

3.2 Feature Extraction

Feature extraction is the process of selecting a subset of the words from the pre-processed training set, which will be used as features in the subsequent text classification process.

We denote the category as C, and the word set of the training sample set is expressed by W= {W1, W2,,,,,,, WN}, By calculating the contribution degree V (Wi, C) of each word Wi (i∈[0,n]) in the set to the class C, and K words exceeding the critical value as the feature items of this kind, The feature set is expressed in the form of Sc={Cw1, Cw2, Cw3, ..., Cwn}.

3.3 Text Representation

Text representation is the transformation of actual text into another form that computers can recognize. Currently, the Vector Space Model (VSM) is a popular representation. The representation of the VSM is to represent each document as a vector form of $d=<t_1,w_1; t_2,w_2...;t_n,w_n>$. Where t_i is the entry and w_i is the weight of t_i in the document d. The calculation method commonly used for the weight w_i is the TD-IDF algorithm.

3.4 Feature Weight Calculation

When we extract feature words from text, each feature value is given a certain weight according to the certain rule TF-IDF (word frequency-inverse document frequency), which is a commonly used feature weight calculation method. Its size reflects the importance of the feature item of the text. If a word

has a larger TF-IDF in the document, then the word is generally more important in this article.

3.5 Traditional Classification Algorithm

The traditional machine learning text classification algorithm can be divided into supervised learning, semi-supervised learning and unsupervised learning according to the characteristics of the learning process. Currently, the most common is supervised learning, which uses three classification methods: NB, KNN and SVM to classify products in the anonymous trading market. Among them, SVM has the best effect and the accuracy rate reaches 89.8%.

Support vector machine is a machine learning method based on traditional learning theory. It uses a nonlinear mapping to map the training data to a higher dimension and find the best hyperplane at a higher latitude for classification purposes.

The SVM classification algorithm has a more accurate classification on the text corpus than most traditional classification algorithms. At the same time, it has more stable classification performance. Types of trading products in the dark market trading market are divided into four areas according to the product title, and 2000 samples are used in each field. The NB, KNN, and SVM algorithms are used to train them. The results are shown in Table 2.

Table 2: Precision of Various Algorithms.

Classification algorithm	Precision	Recall	F1
NB	0.861	0.85	0.83
KNN	0.854	0.86	0.90
SVM	0.898	0.90	0.89

In order to further study the trading products of the dark web trading market and improve the classification accuracy rate, we try to introduce the deep learning classification model. And the single-mode multivariate classification model is proposed. The word vector of word2vec training is embedded in the convolutional neural network (CNN). Compared with the traditional classification, the classification accuracy is improved.

4 Single Mode Multivariate Classification Model

4.1 Convolutional Neural Network (CNN)

Deep learning has been widely used in text categorization. Compare with traditional machine learning technology, it solves the problem of data dependency and feature processing better. Text categorization technology based on word vector and convolution neural networks in considered not only the accuracy of words, but also the relative position words in text. This will improve the accuracy of classification.

CNN consists of an input layer, a hidden layer and an output layer, and parameters are optimized by a back propagation algorithm.

(1) Input layer. As with machine learning, the model performs pre-processing of the input data. The common pre-processing methods in the input layer are de-equalization, normalization, and so on. The input layer is the word vector matrix of words in the sentence. If there are n words and the word vector latitude is k, then the size of this matrix is n×k.

(2) Hidden layer. The hidden layer consists of a convolutional layer and a pooled layer, usually alternately iteratively appearing through a set of convolution kernels of different sizes (filter) hi×k (hi is the number of words contained in the convolution kernel window), the convolution operation of the text from front to back results in several feature mappings. The pooling layer is responsible for compressing the features, simplifying the computational complexity of the network, and proposing the main features, including both average pooling and max pooling. Usually text categorization uses max pooling to select the most important information. The variable length sentence input problem can be solved by the pooling layer, and the maximum value in each feature mapping vector is output.

(3) Output layer. The output layer takes the pooling layer as input, calculates the probability of the data in each category field through the Softmax classifier, and input the final result.

When CNN model is used to classify products in anonymous trading market, we use 1500 data in each field to train classifiers, and 200 data in each field are verified. After 4 iterations, the accuracy of the verification set can reach 91.0%, as shown in Figure 2.

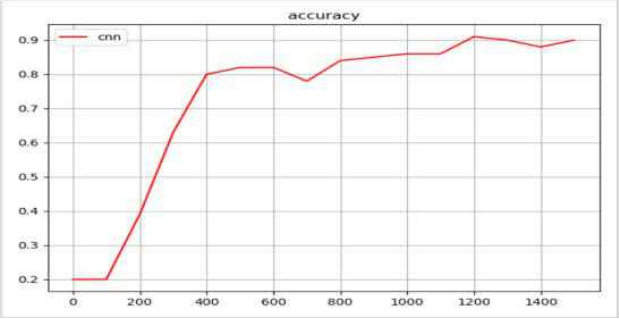

Figure 2: CNN Classification Accuracy Rate.

4.2 Word Vector Embedded in CNN Based on Word2vec Training

The word2vec is to vectorize all words. The Word2vec model extracts word vectors based on context information of words in the text, and the generated word vectors carry contextual semantic information. The Word2vec model has two main training models, CBOW and Skip-Gram. CBOW is suitable for small databases, while Skip-Gram performs better in large databases.

We use the Skip-Gram model in the word2vec model to train the sample data of the anonymous trading market, and use the trained data as the input of the CNN model to train the automatic classification of trading products. The word vector based on word2vec training is embedded in the single mode of CNN. The multivariate classification model is shown in Figure 3.

The parameter configuration of the single-mode multivariate classifier embedded in CNN based on word2vec training is as follows: ①. Convolution kernel window size is 5. Convolution kernel number is 256; ②.Dropout value is 0.5; ③.Fully connected layer neuron size is 128. The classification result of the single-mode multivariate classifier embedded in CNN based on word2vec training is compared with the general convolutional neural network CNN classifier. The word vector embedded in CNN based on word2vec training has lower error and higher accuracy. As shown in Figures 4 and 5. Vertical coordinates indicate accuracy, and transverse coordinates indicate the number of training sets per load.

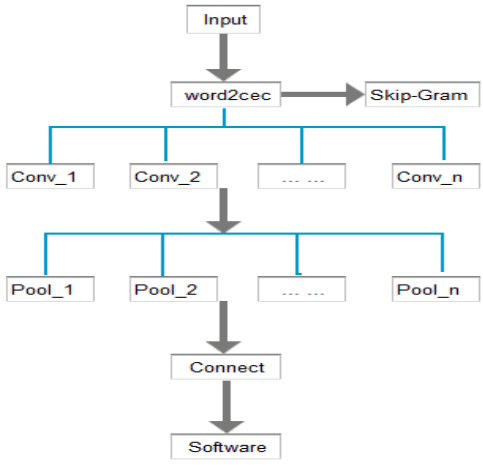

Figure 3: Single-mode Multivariate Classification Model.

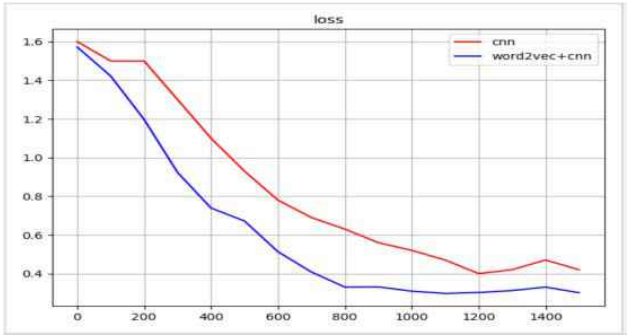

Figure 4: Loss.

After the sample data were trained using the word2vec training-based word vector embedded in CNN single-mode multivariate classifier, we used the test to evaluate the effect of the classifier, and finally achieved an average accuracy of 91.84%. The accuracy, recall rate and F1 value of each classification under the classifier are shown in Table 3.

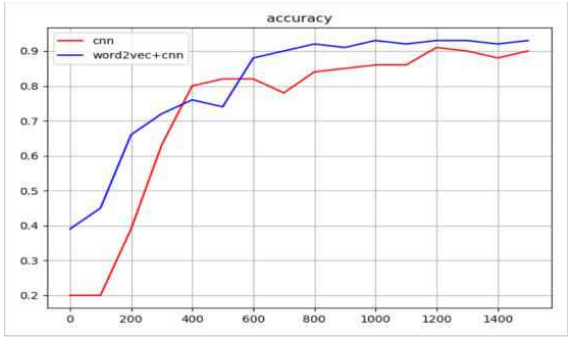

Figure 5: Accuracy.

Table 3: Accuracy, Recall and F1 Values for Each Category.

class	precision	recall	F1
Data	0.83	1.00	0.91
Pron	0.90	0.90	0.90
Service	1.00	0.80	0.89
Drugs	0.90	1.00	0.95
Books	1.00	0.90	0.95

5 Summary and Future Work

In this article, we use the single-mode multivariate classifier model to classify products in the anonymous market by product title. Our algorithm mainly uses the word vector based on word2vec training to embed CNN to train the classifier. Its accuracy rate is 91.84%, which is higher than the traditional machine learning classification. It can enable law enforcement officers to better investigate the trading products provided by the anonymous trading market, without the supplier's intention to provide the wrong category for avoiding tracking.

Our training sets and test sets only use data from a single dark-net market. In future research work, we will extract test data from a wider market, making the classification model more representative. In addition, our algorithms can take into account features other than the product list text when classifying products. For example, the price of the product, etc.

REFERENCES

[1] Xuan Zhang and K P Chow (2018). A Framework for Dark Web Threat Intelligence Analysis. International of Digital Crime and Forensics, 2018(10), 108-117.
[2] Bao Kai (2016). Study on vulnerability of dark network based on TOR. Computer application technology.
[3] http://deepmix5e3vptpr.onion.
[4] R D Kowalski and Malinowski A (2008). How to meet in anonymous network. Theoretical Computer Science, 399(1-2), 141-15.
[5] Yang Yi and Guo Han (2017). Darknet Resource Exploring based on Tor. Communications Technology, 50(10).
[6] Tong Yuen Yum (2017). Explore the Dark Web. The University of Hong Kong.
[7] Frank R, Monk B, et al. (2016). Surfacing collaborated network in dark web of find illicit and criminal content. Intelligence and Security Informatics, IEEE, 109-114.

[8] Chen H (2008). Discovery of improvised explosive device content in the Dark Web. IEEE International Conference on Intelligence and Security Informatics, 88-93.
[9] Guitton C (2013). A review of the available content on Tor hidden services: The case against further development. Computes in Human Behavior, 29(6), 2805-2815.
[10] G Branwen (2015). Silk Road: Theory and Practice. http://www.gwern.net/Silk%20Road, Aguest 2015, Accessed, 2015-12-09.
[11] N Christin (2013). Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In Proceeding of 22nd international conference Steering Committee, pp. 213-214.
[12] Kim Y (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.