JES
Journal of Engineering Sciences

Vol 13, Issue 05, May /2022
ISSN NO:0377-9254

TEXT CLASSIFICATION FOR NEWSGROUP USING-MACHINE LEARNING

K. Nagendra Gopal [1], Dr.P.Vamsi Krishnam Raja[2], Dr.Pamidi Srinivasulu [3]

[123] Department of computer science and engineering, Swarnandhra College of Engineering and Technology, Narsapur, West Godavari, Andhra Pradesh, India

Abstract: Model can get best text classification accuracy. With the developments of internet technologies, dealing with a mass of law cases urgently and assigning classification cases automatically are the most basic and critical steps. Convolutional Neural Networks (CNNs), has been shown to be effective for text classification. To better apply CNNs into law text classification, this paper presents a new semi-supervised Convolutional Neural Networks (SSC) framework. Our method combines unlabeled data with a small labeled training set to train better models, and then integrates into a supervised CNN. More specifically, for effective use of word order for text categorization, we use the feature of not low-dimensional word vectors but high-dimensional text data, that is, a small text regions is learned based on sequences of one-hot vectors. To better improve the prediction accuracy of the scheme, we seek effective use of unlabeled data for text categorization for integration into a supervised CNN.We compare the proposed scheme to state-of-the-art methods by the real datasets. The results demonstrate that the semi-supervised learning

1. INTRODUCTION

The ML and DL methods covered in this paper are applicable to intrusion detection in wired and wireless networks. Readers who wish to focus on wireless network protection can refer to essays such as Soni et al, which focuses more on architectures for intrusion detection systems that have been introduced for MANETs.Security breaches include external intrusions and internal intrusions. There are three main types of network analysis for IDSs: misuse-based, also known as signature-based, anomaly-based, and hybrid. Misuse-based detection techniques aim to detect known attacks by using the signatures of these attacks . They are used for known types of attacks without generating a large number of false alarms. However, administrators often must manually update the database rules and signatures. New (zero-day) attacks cannot be detected based on misused technologies. Anomaly-based techniques study the normal network and system behavior and identify anomalies as deviations from normal behavior. They are appealing because of their capacity to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, therefore making it difficult for attackers to know which activities they can perform undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviors can be categorized as anomalies. Hybrid detection combines misuse and anomaly detection. It is used to increase the detection rate of known intrusions and to reduce the false positive rate of unknown attacks. Most ML / DL methods are hybrids.

Vol 13, Issue 05, May /2022
ISSN NO:0377-9254

2. RELATED WORK

Automatic text classification has always been an important application and research topic since the inception of digital documents. Today, text classification is a necessity due to the very large amount of text documents that we have to deal with daily. In general, text classification includes topic based text classification and text genre-based classification. Topic-based text categorization classifies documents according to their topics [1]. Texts can also be written in many genres, for instance: scientific articles, news reports, movie reviews, and advertisements. Genre is defined on the way a text was created, the way it was edited, the register of language it uses, and the kind of audience to whom it is addressed. Previous work on genre classification recognized that this task differs from topic-based categorization [2]. Sebastiani gave an excellent review of text classification domain [3]. Thus, in this work apart from the brief description of the text classification we refer to some more recent works than those in Sebastiani's article as well as few articles that were not referred by Sebastiani. In Figure 1 is given the graphical representation of the Text Classification process. The task of constructing a classifier for documents does not differ a lot from other tasks of Machine Learning. The main issue is the representation of a document [4]. In Section 2 the document representation is presented. Thus dimension reduction methods are called for. Two possibilities exist, either selecting a subset of the original features [5], or transforming the features into new ones, that is, computing new features as some functions of the old ones [6]. Although stemming is considered by the Text Classification community to amplify the classifiers performance, there are some doubts on the actual importance of aggressive stemming, such as performed by the Porter Stemmer [7]. An ancillary feature engineering choice is the representation of the feature value [8]. Often a Boolean indicator of whether the word occurred in the document is sufficient. Most of the text categorization algorithms in the literature represent documents as collections of words. An alternative which has not been sufficiently explored is the use of word meanings, also known as senses. Kehagias et al. using several algorithms, they compared the categorization accuracy of classifiers based on words to that of classifiers based on sense[9]. Methods for feature subset selection for text document classification task use an evaluation function that is applied to a single word [10]. Scoring of individual words (Best Individual Features) can be performed using some of the measures, for instance, document frequency, term frequency, mutual information, information gain, odds ratio, χ 2 statistic and term strength [11], [12]. What is common to all of these feature-scoring methods is that they conclude by ranking the features by their independently determined scores, and then select the top scoring features.

3. EXISTING MODEL

CNNs is a neural network that can make use of the internal structure of data. Some famous architectures have been proposed, such as, AlexNet, LeNet, GoogLeNet, VGG- 16, NiN. It is equipped with convolution layers interleaved with subsampling layers and then pass fully connection layer. Finally, output layer exports classification results, where the top layer makes use of the features generated by the lower layer to make classification. CNNs is marked by the locally-connection, weight share, and sub sampling. At first, locally-connection reduces the number of the neural parameters of each layer, and makes error with smaller breadth divergence from the output layer start. And then, the concept of weight share learns from the optic nerve

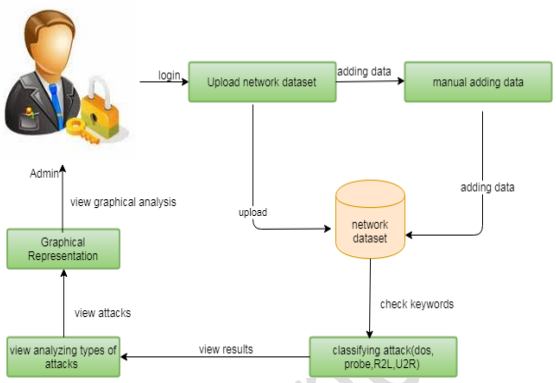
Vol 13, Issue 05, May /2022
ISSN NO:0377-9254

receptive field. A distinguishing characteristic of convolution layer is weight sharing. For input x, a unit associated with the i − th region calculates σ (W · ri (x) + b), where ri (x) is a region vector expressing the region of input x at location i. Here, σ is a nonlinear activation function, (e.g., applying y = f (x) = max (0, x), namely ReLU, to each vector portion). In the end, the goal of the subsampling layer is that condenses the adjacent region vectors of certain size into a vector, and make text regions zoom with a certain proportion. According to the different scaling algorithm, commonly-used scaling algorithms are average-pooling and max-pooling.

4. METHODOLOGY

With the developments of internet technologies, dealing with a mass of law cases urgently and assigning classification cases automatically are the most basic and critical steps. Convolutional Neural Networks (CNNs), has been shown to be effective for text classification. To better apply CNNs into law text classification, this paper presents a new semi-supervised Convolutional Neural Networks (SSC) framework. Our method combines unlabeled data with a small labeled training set to train better models, and then integrates into a supervised CNN. More specifically, for effective use of word order for text categorization, we use the feature of not low-dimensional word vectors but high-dimensional text data, that is, a small text regions is learned based on sequences of one-hot vectors. To better improve the prediction accuracy of the scheme, we seek effective use of unlabeled data for text categorization for integration into a supervised CNN.We compare the proposed scheme to state-of-the-art methods by the real datasets. The results demonstrate that the semi-supervised learning model can get best text classification accuracy.

4.1 Architecture:



4.2 The proposed model :

The output of the convolution layer is put into a pooling layer, which is a no-parameter layer. The essence of pooling layer, as described, is shrunk the data size by merging neighboring region; that is, it brings down the dimension of the data. So that, higher layer can process more abstract/ global information. The reason why it does is because the abstract/global feature information can still describe data, even if it reduces many data. What is more, it can avoid overfitting as well, due to decreasing the dimension of data. Frequently-used merging ways are averagepooling and max-pooling in pooling layer. A pooling layer includes a number of pooling units, where each of pooling units responds to a small region of text data.Semi-supervised learning is the combination method of supervised learning and unsupervised learning. It focuses on the task of using a small amount of label text and a mass of no-label text to make train and classification. We put forward the semi-supervised learning framework including two steps. The step is tv-embedding (tv represents two-view) learning. The definition of tv-embedding is that, if there exists a function g(e.g.P (X2|X1) = g1 (f1 (X1),X2)), the function f1 is the tv-embedding of $\chi 1$w.r.t.$\chi 2$, for any (X1,X2) $\in \chi 1 \times \chi 2$. A tv-embedding of a view(X1) keeps required structure in order

to generate the another view(X2),and it learns useful feature vectors from no-label text data. The tv-embedding model mainly implements the following three goal.

4.3 ALGORITHM: SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is one of the most robust and accurate methods in all machine-learning algorithms. It primarily includes Support Vector Classification (SVC) and Support Vector Regression (SVR). The SVC is based on the concept of decision boundaries. A decision boundary separates a set of instances having different class values between two groups. The SVC supports both binary and multi-class classifications. The support vector is the closest point to the separation hyperplane, which determines the optimal separation hyperplane. In the classification process, the mapping input vectors located on the separation hyperplane side of the feature space fall into one class, and the positions fall into the other class on the other side of the plane. In the case of data points that are not linearly separable, the SVM uses appropriate kernel functions to map them into higher dimensional spaces so that they become separable in those spaces

4.4 MODULES:

Documents To Server

The document which required to analysis is needed to upload to the server. The only uploaded documents will be able to cluster. The uploading page will be contains the details about the document and it is given in the output page of uploading.

Text classification

The centroids are fixed set of words that are actually makes the context of the content to be classified and clustered into folders. The general sets of documents are into their respective clusters based on the separation of centroids.

Semi-supervised CNN

The existing Semi-supervised CNNof document will be shown in the system to compare with the proposed system. In existing system centroids were fixed in programmatically so it cannot change according to user needs. So it is fixed it will only forms certain number of clusters. The document will be analysis by the help of content within the document.

Graphical Analysis

This module will describe the comparison of existing and proposed system in graphical manner. The graphs such as column chart, line chart are shown to display the comparison and efficiency in proposed system is shown evidently from the experiment.

4.5 Software Requirements

For developing the application the following are the Software Requirements:

1. Python
2. Django
3. MySql
4. MySqlclient
5. WampServer 2.4

Operating Systems supported

1. Windows 7
2. Windows XP
3. Windows 8

Technologies and Languages used to Develop

1. Python

Debugger and Emulator

- Any-Browser (Particularly Chrome)
-

4.6 DESIGN

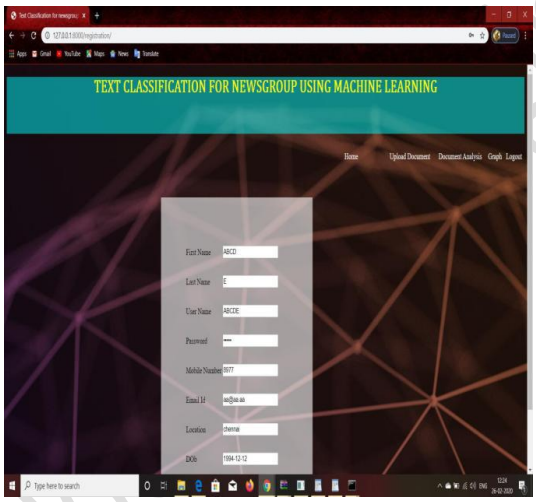
The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it

can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

4.7 Graphical Representation:

The primary piece of the venture is to investigation the assault types in the organization dataset. The client information examination of the information should be possible by diagrams design. This is where administrator have capacity to come for specific arrangement about proposed framework. The pictorial portrayals of gathered information are appeared as charts. The various diagrams give the best examination of the framework.

4.8 RESULTS



5.CONCLUSION

CNN as a reasonable methodology can precisely accomplish text order. We have proposed another engineering for NLP which follows the plan rule: television implanting of text locales with unlabeled information and afterward named information, that is, a semi-managed system. This engineering has been assessed on a uninhibitedly accessible huge scope informational indexes: the Chinese lawful case portrayal. We can show that semi supervised CNNs with television embeddings for text order improves execution contrasted and the conventional neural organizations. Because of the restricted space, this paper just considered the law text order, consequently we will broaden the framework so it can another application, for example, traffic rules, film audit, and so forth.

References:

Sebastiani, F.: Machine learning in automated
text categorization. ACM Comput. Surv. 34(1), 1–47 (2002)
Al-Harbi, S., Almuhareb, A., Al-Thubaity, A.,
Khorsheed, M., Al-Rajeh, A.: Automatic Arabic text classification. In: JADT'08, France, pp. 77–83 (2008)
Forman, George: An extensive empirical study of feature selection metrics for text
classification. J. Mach. Learn. Res. 3, 1289–1305 (2003)
Yang, Y., Pedersen, J.O.: A Comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412–420, 08–12 July 1997
Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R.: Text document pre-processing with the
Bayes formula for classification using the support vector machine. IEEE Trans. Knowl. Data Eng. 20(9), 1264–1272 (2008)
Yan, X., Gareth J., Li J.T., Wang, B., Sun, C.M.: A study on mutual information based feature selection for text categorization'. J. Comput. Inf. Syst. 3(3), 1007– 1012 (2007).
Porter, M.F.: An algorithm for suffix stripping.
Program 14(3). 130–137 (1980)
Nigam, K., Mccallum, A.K., Thrun, S.,

Vol 13, Issue 05, May /2022
ISSN NO:0377-9254

Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. 39, 103–134 (2000)

Joachims, T.: A statistical learning model for text classification for support vector machines.
In: 24th ACM International Conference on Research and Development in Information Retrieval (SIGIR) (2001)

Dong, Tao, Shang, Wenqian, Zhu, Haibin: An improved algorithm of Bayesian text categorization. J. Softw. 6(9), 1837–1843 (September 2011)

Kumar, C.A.: Analysis of unsupervised dimensionality reduction techniques. Comput.
Sci. Inf. Syst. 6(2), 217–227 (Dec. 2009)

Soon, C.P.: Neural network for text classification based on singular value decomposition. In: 7th International conference on Computer and Information Technology, pp. 47–52 (2007)

Muhammed, M.: Improved k-NN algorithm for text classification. Department of Computer Science and Engineering University of Texas at
Arlington, TX, USA Ikonomakis, M., Kotsiantis, S., Tampakas, V.: Text classification using machine learning techniques. IEEE Trans.
Comput. 4(8) 966–974 (2005)