Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 04 (April -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i04.pp53 - 59*

Text Classification for Newsgroup using Deep Learning

Mr.Ch.Mani Kanta Kalyan, [1],Koppana Satya [2],Mallipamula Lakshmi Prasanna [3],Kamadi Jyothi Manasa [4],Gedda Gowthami [5],Samanasa Navya Satya Manisri [6]
ASSISTANT PROFESSOR
DEPT OF COMPUTER SCIENCE AND ENGINEERING
PRAGATI ENGINEERING COLLEGE(A),SURAMPALEM(EAST GODAVARI)A.P,INDIA

To Cite this Article
Mr.Ch.Mani Kanta Kalyan, ,Koppana Satya ,Mallipamula Lakshmi Prasanna ,Kamadi Jyothi Manasa ,Gedda Gowthami ,Samanasa Navya Satya Manisri **Text Classification for Newsgroup using Deep Learning ”** *Journal of Science and Technology, Vol. 08, Issue 04,-April 2023, pp53-59*
Article Info
Received: 21-02-2023 Revised: 11-03-2023 Accepted: 23-03-2023 Published: 17-04-2023

Abstract:

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 labelled 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 region 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.

I. 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.

II. LITERATURE SURVEY

2.1 Machine Learning in Automated Text Categorization

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert manpower, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely document representation, classifier construction, and classifier evaluation.

2.2 Automatic Arabic text categorization: A comprehensive comparative study

Text categorization or classification (TC) is concerned with placing text documents in their proper category according to their contents. Owing to the various applications of TC and the large volume of text documents uploaded on the Internet daily, the need for such an automated method stems from the difficulty and tedium of performing such a process manually. The usefulness of TC is manifested in different fields and needs. For instance, the ability to automatically classify an article or an email into its right class (Arts, Economics, Politics, Sports, etc.) would be appreciated by individual users as well as companies. This paper is concerned with TC of Arabic articles. It contains a comparison of the five best known algorithms for TC. It also studies the effects of utilizing different Arabic stemmers (light and root-based stemmers) on the effectiveness of these classifiers. Furthermore, a comparison between different data mining software tools (Weka and RapidMiner) is presented. The results illustrate the good accuracy provided by the SVM classifier, especially when used with the light10 stemmer. This outcome can be used in future as a baseline to compare with other unexplored classifiers and Arabic stemmers.

2.3 A Comparative Study on Feature Selection in Text Categorization

This paper is a comparative study of feature selection methods in statistical learning of text categorization. The focus is on aggressive dimensionality reduction. Five methods were evaluated, including term selection based on document frequency (DF), information gain (IG), mutual information (MI), a Ø 2 -test (CHI), and term strength (TS). We found IG and CHI most effective in our experiments. Using IG thresholding with a k nearest neighbor classifier on the Reuters corpus, removal of up to 98% removal of unique terms actually yielded an improved classification accuracy (measured by average precision) . DF thresholding performed similarly. Indeed, we found strong correlations between the DF, IG and CHI values of a term. This suggests that DF thresholding, the simplest method with the lowest cost in computation, can be reliably used instead of IG or CHI when the computation of these measures is too expensive. TS compares favorably with the other methods with up to 50% vocabulary redo.

2.4 Text Document Pre-processing with the Bayes Formula for Classification Using the Support Vector Machine

This work implements an enhanced hybrid classification method through the utilization of the naïve Bayes classifier and the Support Vector Machine (SVM). In this project, the Bayes formula was used to vectorize (as opposed to classify) a document according to a probability distribution reflecting the probable categories that the document may belong to. The Bayes formula gives a range of probabilities to which the document can be assigned according to a pre-determined set of topics such as those found in the "20 newsgroups" dataset for instance. Using this probability distribution as the vectors to represent the document, the SVM can then be used to classify the documents on a multi – dimensional level. The effects of an inadvertent dimensionality reduction caused by classifying using only the highest probability using the naïve Bayes classifier can be overcome using the SVM by employing all the probability values associated with every category for each document. This method can be used for

any dataset and shows a significant reduction in training time as compared to the square method and significant improvement in classification accuracy when compared to pure naïve Bayes systems and also the TF-IDF/SVM hybrids

III.SYSTEM ANALYSIS

1.1. EXISTING SYSTEM

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 subsampling. 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 receptive field. A distinguishing characteristic of convolution layer is weight sharing. For input x, a unit associated with the region calculates, where is a region vector expressing the region of input x at location 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.

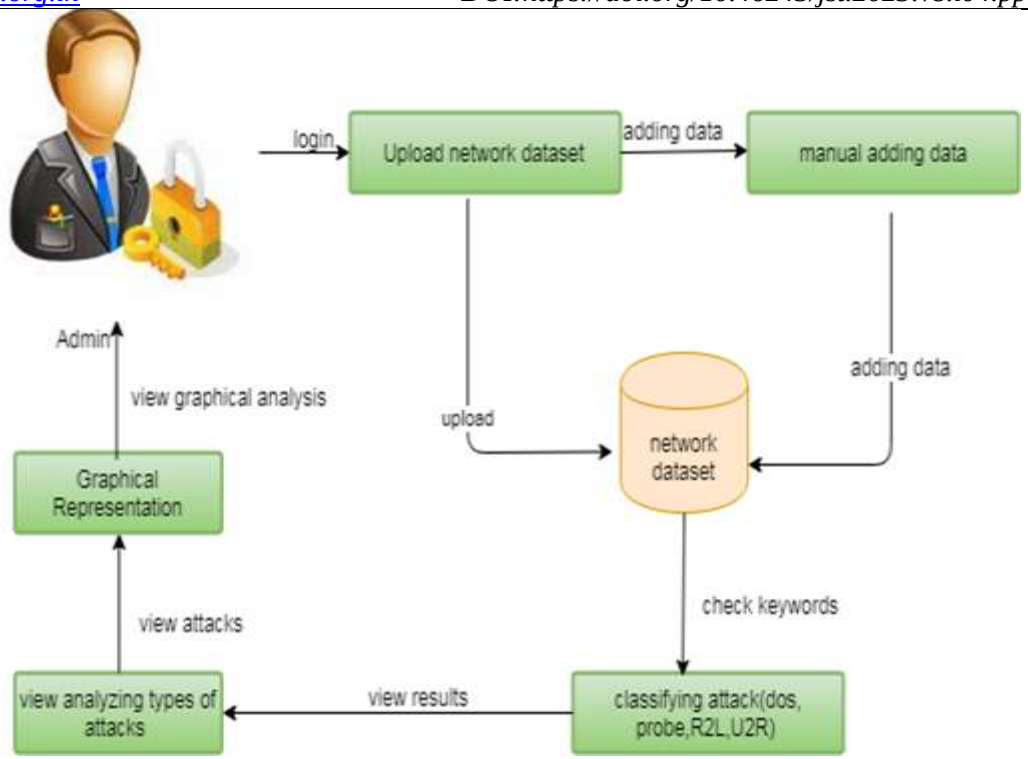
1.2. PROPOSED SYSTEM

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 average pooling 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 and it learns useful feature vectors from no-label text data. The tv-embedding model mainly implements the following three goal.

IV.SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

Below diagram depicts the whole system.



4.1. System Architecture

V. SYSTEM IMPLEMENTATION

5.1. MODULES

- DOCUMENTS TO SERVER
- TEST CLASSIFICATION
- SEMI-SUPERVISED CNN
- GRAPHICAL ANALYSIS

5.1.1 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 containing the details about the document and it is given in the output page of uploading.

5.1.2 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.

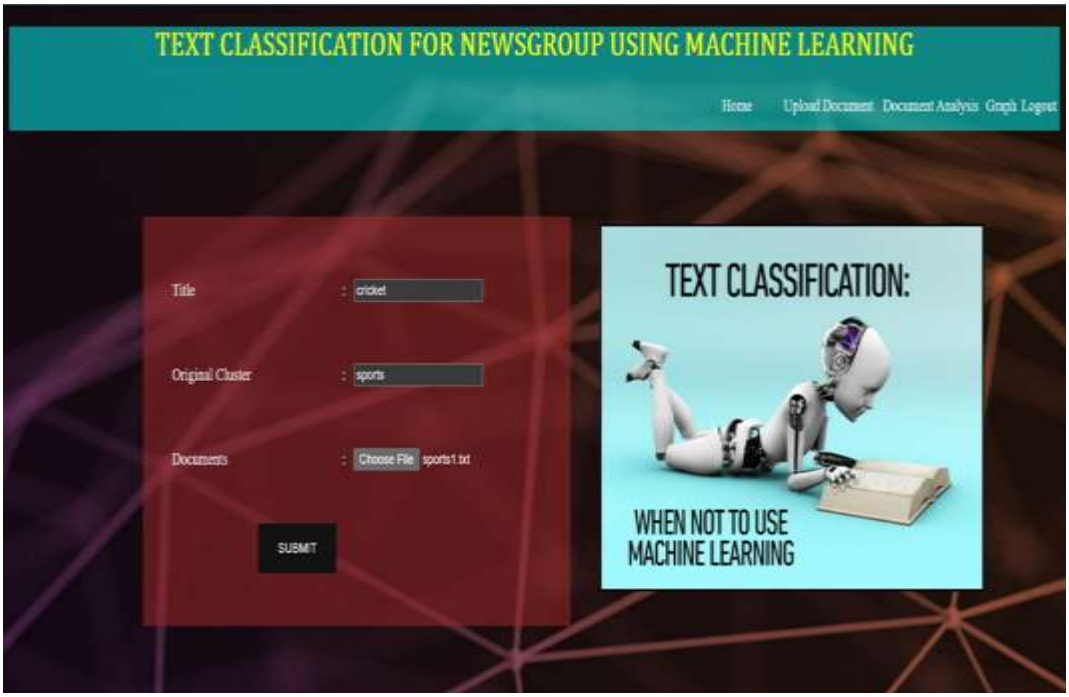
5.1.3 Semi-supervised CNN

The existing Semi-supervised CNN of 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.

5.1.4 Graphical Analysis

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

VI. RESULTS



6.1 Upload the text files

Fig. 6.2 Content matched in the cluster

Fig. 6.4 Graphical Analysis

VII.CONCLUSION AND FUTURE WORK

CNN as a viable approach can accurately achieve text categorization. We have proposed a new architecture for NLP which follows the design principle: tv-embedding of text regions with unlabeled data and then labelled data, that is, a semi-supervised framework. This architecture has been evaluated on a freely available large-scale data sets: the Chinese legal case description. We can show that semi supervised CNNs with tv-embeddings for text categorization improves performance compared with the traditional neural networks. Due to the limited space, this paper only considered the law text classification, therefore we will extend the system so that it is able to another applications, such as, traffic rules, film review, etc.

REFERENCES :

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data', SIGKDD Explor Newsl, vol. 1, no. 2, pp. 12–23, Jan. 2000.

[2] M. Jafari, F. SoleymaniSabzchi, and S. Jamali, 'Extracting Users'Navigational Behaviour from Web Log Data: a Survey', J. Comput. Sci. Appl. J. Comput. Sci. Appl., vol. 1, no. 3, pp. 39–45, Jan. 2013.

[3] N. Soni and P. K. Verma, 'A Survey On Web Log Mining And Pattern Prediction', Int. J. Adv. Technol. Eng. Sci.-2348-7550.

[4] T. R. Ramesh and C. Kavitha, 'Web user interest prediction framework based on user behaviour for dynamic websites', Life Sci. J., vol. 10, no. 2, pp. 1736–1739, 2013.

[5] L. Yi, B. Liu, and X. Li, 'Eliminating Noisy Information in Web Pages for Data Mining', in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 296–305.

[6] A. Dutta, S. Paria, T. Golui, and D. K. Kole, 'Structural analysis and regular expressions-based noise elimination from web pages for web content mining', in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1445–1451.

[7] G. D. S. Jayakumar and B. J. Thomas, 'A new procedure of clustering based on multivariate outlier detection', J. Data Sci., vol. 11, no. 1, pp. 69–84, 2013.

[8] V. Chitraa and A. S. Thanamani, 'Web Log Data Analysis by Enhanced Fuzzy C Means Clustering', Int. J. Comput. Sci. Appl., vol. 4, no. 2, pp. 81–95, Apr. 2014.

[9] L. K. Joshila Grace, V. Maheswari, and D. Nagamalai, 'Analysis of Web Logs And Web User In Web Mining', Int. J. Netw. Secur. Its Appl., vol. 3, no. 1, pp. 99–110, Jan. 2011.

[10] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, 'User profiles for personalized information access', in The adaptive web, Springer, 2007, pp. 54–89.

[11] P. Peñas, R. del Hoyo, J. Vea-Murguía, C. González, and S. Mayo, 'Collective Knowledge Ontology User Profiling for Twitter – Automatic User Profiling', in 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, vol. 1, pp. 439–444.

[12] S. Kanoje, S. Girase, and D. Mukhopadhyay, 'User profiling trends, techniques and applications', ArXiv Prepr. ArXiv150307474, 2015.

[13] H. Kim and P. K. Chan, 'Implicit indicators for interesting web pages', 2005.

[14] J. Xiao, Y. Zhang, X. Jia, and T. Li, 'Measuring similarity of interests for clustering Web-users', in Proceedings 12th Australasian Database Conference. ADC 2001, 2001, pp. 107–114