© 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Science and Technology

Internal News Classification Using Deep Learning

Shivani Kundra
Department of Computer Science, DAV University, Jalandhar, Punjab, India

ABSTRACT

For the last few years, text mining has been gaining significant importance. Since Knowledge is now available to users through variety of sources i.e. electronic media, digital media, print media, and many more. Due to huge availability of text in numerous forms, a lot of unstructured data has been recorded by research experts and have found numerous ways in literature to convert this scattered text into defined structured volume, commonly known as text classification. Focus on full text classification i.e. full news, huge documents, long length texts etc. is more prominent as compared to the short length text. In this paper, we have discussed text classification process, classifiers, and numerous feature extraction methodologies but all in context of short texts i.e. news classification based on their headlines. Existing classifiers and their working methodologies are being compared and results are presented effectively.

Keywords: RSS, Support Vector Machine, Web Page Classification Method, SIU, Classifier, JAVA JDK, SDK

I. INTRODUCTION

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done ”manually” (or ”intellectually”) or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is mainly in information science and computer science. The problems are over- lapping, however, and there is therefore interdisciplinary research on document classification [9].

With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the WWW, and to guide a user’s search through hypertext. Since building text classifiers by hand is dificult and time-consuming, it is advantageous to learn classifiers from examples[1].

In this proposed work, we are going to develop a classifier in form os a software prototype that will use the deep learning for the internal news classification. The Software prototype will consist of 5 modules are demonstrated in Figure 1.1 :

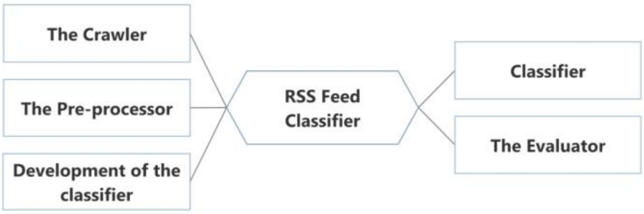


Figure 1.1 : Software Prototye of Proposed Work

1. The crawler: It will be develop in order to crawl the full length RSS feeds from the web automatically. In this way we will produce the data for the analysis of proposed work.
2. The Pre-processor: The work of this module is to pre-process the data in order to remove the outliers and vectorization of the news articles. Basically the development of the input or data frame for the classifier.
3. Development of the Classifier: The data frame that is an output of the pre-processor module is

used and a classifier is developed using the deep learning technique.

4. Classifier: This module basically applies the classifier on the un-categorized news articles and categorizes them.
5. The Evaluator: This module basically evaluates the performance of the classifier developed. The evaluation will be done using some evaluation metrics.

The importance of text classification is even more apparent when the information space is huge such as the World Wide Web. Examples of web classification systems include [2] Yahoo! directory and Google web directory Service [3].

II. METHODS AND MATERIAL

A. Literature Survey

In the last decade, a great deal of scientific researchers and studies have been performed on news classification, Brief descriptions of their work are given below:

1. Ee, Chee-Hong Chan Aixin Sun, and Peng Lim. ”Automated online news classification with personalization.” 4th international confer- ence on asian digital libraries. 2001.

   They proposed a categorizer system, they have experimented an automated approach to classify online news using the Support Vector Machine (SVM) technique. SVM has been shown to deliver good classification results when ample training documents are given. In their search, they had applied SVM to personalized classification on online news. In personalized classification, users can define their personalized categories using a few keywords. By construct- ing search queries using these keywords, categorizer obtains both positive and negative training documents required for construction of personalized clas- sifiers. In this paper, they described the preliminary version of Categorizer and present system architecture [1].

2. Selamat, A.; Yanagimoto, H.; Omatu, S., ”Web news classification using neural networks based on PCA,” SICE 2002. Proceedings of the 41st SICE Annual Conference , vol.4, no., pp.2389,2394 vol.4, 5-7 Aug. 2002.

   In this paper, they proposed a news web page classification method (WPCM). The WPCM uses a neural network with inputs obtained by both the principal components and class profile-based features (CPBF). The fixed number of regular words from each class will be used as a feature vectors with the reduced features from the PCA. These feature vectors are then used as the input to the neural networks for classification. The experimental evaluation demonstrates that the WPCM provides acceptable classification accuracy with the sports news datasets [4].

3. Burkepile, A.; Fizzano, P., ”Classifying RSS Feeds with an Artifi- cial Immune System,” Information, Process, and Knowledge Man- agement, 2010. eKNOW ’10. Second International Conference on, vol., no., pp.43,47, 10-15 Feb. 2010

   Artificial Immune Systems (AIS) have been used in a number of applications from autonomous navigation to computer security because of their ability to rapidly adapt and evolve. In this paper, they examined the application of an AIS for the purpose of determining which news articles from a set of RSS feeds are relevant. Because the articles we are examining come from RSS feeds, the articles can vary greatly in length and detail. Their training set is composed of a set of news articles that represent articles a user has already deemed relevant. Then they have the AIS determine which articles from another set are related to the relevant articles. they show that the AIS performs well regardless of the diversity of the subjects in the data set and can even make fairly fine grained distinctions with high accuracy [8].

4. Tufekci, P.; Uzun, E.; Sevinc, B., ”Text classification of web based news articles by using Turkish grammatical features,” Signal Pro- cessing and Communications Applications Conference (SIU), 2012 20th , vol., no., pp.1,4, 18-20 April 2012.

   In this study, they explained how to reduce the dimension of the feature vec- tor by using Turkish’s grammar rules without compromising success rates. The feature vector is weighted on the basis of the word frequency as the word stems have been selected as features. During this selection the effects of selection of the word stems with different length and type to the classi- fication are investigated and when the word stems with noun type and the maximum length are selected as features, the success rate has been found to be at the highest level. When this selection is applied with the other methods which reduce the dimension, the dimension of the feature vector is decreased to 97.46%. Using the reduced feature vector the better success rates generally have been obtained from Naive Bayes, SVM, C 4.5 and RF classification methods and the best performance achieved is 92.73% which has been obtained using the Naive Bayes method [6].

5. Agarwal, S.; Singhal, A.; Bedi, P., ”Classification of RSS feed news items using ontology,” Intelligent Systems Design and Ap- plications (ISDA), 2012 12th International Conference on , vol., no., pp.491,496, 27-29 Nov. 2012.

   In their approach, they are using weighted Concept Frequency-Inverse Docu- ment Frequency (CF-IDF) with background knowledge of domain Ontology, for classification of RSS feed News Items. Metadata information of news items has been used to assign weight to the identified concepts. No trained classifiers are required as Ontology itself acts as a classifier. they have de- signed ontology based on news industry standards. This classification approach considers relations among the concepts and properties. It results in reduction of noise in final output. It considers only the key concepts of a domain for classification instead of all the terms, which curbs the problem of dimensionality. Evaluation of experimental results reveals that proposed approach gives better classification results [7].

6. Dilrukshi, I.; De Zoysa, K.; Caldera, A., ”Twitter news classifica- tion using SVM,” Computer Science & Education (ICCSE), 2013 8th International Conference on , vol., no., pp.287,291, 26-28 April 2013.

   They classified news into different groups so that the user could identify the most popular news group in a given country for a given time. The short messages were extracted from Twitter micro blog. Several active news groups were chosen to extract the short messages. Each short message was classified manually into 12 groups. These classified data were used to train the machine learning techniques. The data were trained using SVM (Support Vector Ma- chine) machine learning techniques. Current research is a high dimensional problem as a large number of features will be collected using short messages. F-Measure was calculated to obtain a single value measurement [5].

7. Aparicio, Roxana, and Edgar Acuna. ”Using Ontologies to Im- prove Document Classification with Transductive Support Vector Machines.” International Journal of Data Mining & Knowledge Management Process 5.3 (2013).

   In this paper, they proposed the use of ontologies in order to improve the accuracy and efficiency of the semi-supervised document classification. They used support vector machines, which is one of the most effective algorithms that have been studied for text. Their algorithm enhances the performance of transductive support vector machines through the use of ontologies. They reported experimental results applying their algorithm to three different datasets. Their experiments show an increment of accuracy of 4% on av- erage and up to 20%, in comparison with the traditional semi-supervised model [3].

8. Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib. ”On- line News Classification: A Review.”, International journal of In- novation in Engineering and Technology (IJIET), Vol. 2, Issue. 2, 2013.

   In this presented work, they created a novel algorithm which can classify inner structure of simple classified news. The work had just been done to identify the outer clusters of the cluster but no work till now had been done for the inner cluster of dataset. Their proposed work will be creating inners structure of each and every field of the proposed system like SPORTS, ENTERTAINMENT and MATRIMONIAL’S [2].

B. Problem Formulation

Classification is the one of the most exiting scientific task of this proposed work. Crawling some RSS feeds, the classifier will be developed that will classify the unclassified news articles into some suitable category. The Generalized structure of the proposed classifier is demonstrated in the Figure 3.1.

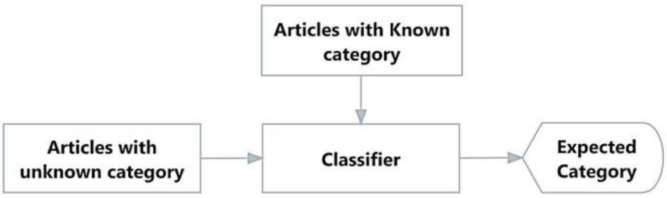


Figure 3.1: Problem Formulation

The first step in text categorization is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. The Strings are broken and Frequency table is devel-oped. Each distinct word wi corresponds to a feature, with the number of times word wi occurs in the document as its value.

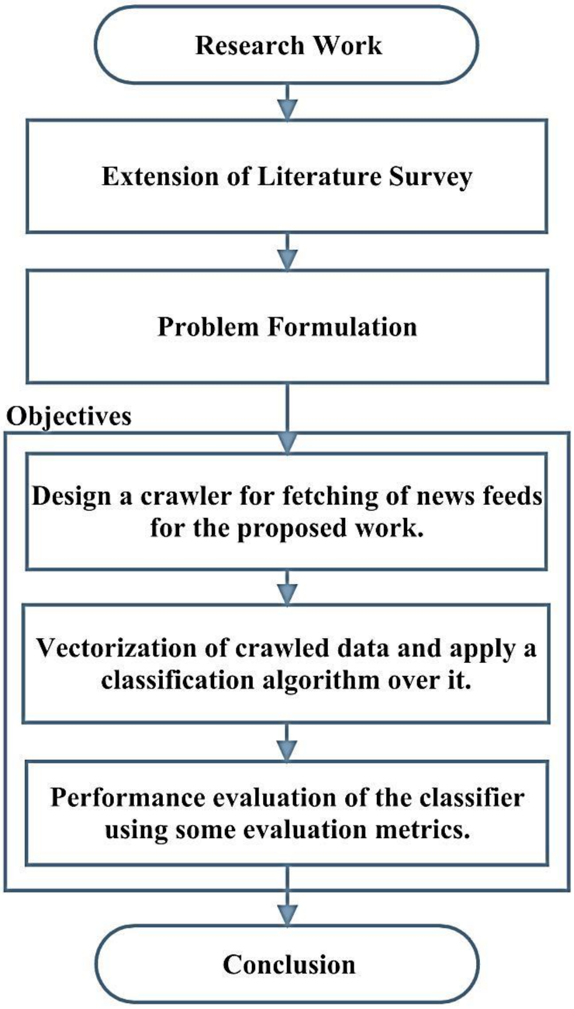
C. Objectives

In this proposed work, it is proposed to model the news patterns in a systematic way. The work will be performed by crawling some RSS Feeds using a crawler, vectorization then model them using deep learning. The main objectives of the thesis work are:

1. Design a crawler for fetching of news feeds for the proposed work.
2. Vectorization of crawled data and apply a classification algorithm over it.
3. Performance evaluation of the classifier using some evaluation metrics.

D. Methodology

The motivation to the proposed work is to Internal news classification using deep learning. The detailed methodology is presented in the form of flow chart in Figure 5.1.


Research Work
Extension of Literature Survey
Problem Formulation
Objectives
Design a crawler for fetching of news feeds for the proposed work.
Vectorization of crawled data and apply a classification algorithm over it.
Performance evaluation of the classifier using some evaluation metrics.
Conclusion

Moreover, we will also suggest future work on internal news classification at the end.

E. Facilities Required for Work

All the calculations will be performed on 5Neuroph tool and JAVA, which requires:

- System

  – Processor : Intel Core i5
  – RAM : 4 GB
  – Harddisk : 500 GB
  – Operating System : Windows 7 Professional

- Neuroph 2.7
- JAVA JDK, SDK
- JDeveloper 11.1.2.3

Here Neuroph is an open-source, object-oriented neural network framework written in Java. It can be used to create and train neural networks.

III. REFERENCES

[1] http://www.cs.cornell.edu/people/tj/publications/joachims 98a.pdf

[2] Yahoo! Driectory, http://www.yahoo.com

[3] Google Web Directory, http://www.google.com/dirhp

[4] Ee, Chee-Hong Chan Aixin Sun, and Peng Lim. ”Automated online news clas- sification with personalization.” 4th international conference on asian digital libraries. 2001.

[5] Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib. ”Online News Classification: A Review.”, International journal of Innovation in Engineering and Technology (IJIET), Vol. 2, Issue. 2, 2013.

[6] Aparicio, Roxana, and Edgar Acuna. ”Using Ontologies to Improve Docu- ment Classification with Transductive Support Vector Machines.” Interna- tional Journal of Data Mining & Knowledge Management Process 5.3 (2013).

[7] Selamat, A.; Yanagimoto, H.; Omatu, S., ”Web news classification using neural networks based on PCA,” SICE 2002. Proceedings of the 41st SICE Annual Conference , vol.4, no., pp.2389,2394 vol.4, 5-7 Aug. 2002.

[8] Dilrukshi, I.; De Zoysa, K.; Caldera, A., ”Twitter news classification using SVM,” Computer Science & Education (ICCSE), 2013 8th International Con- ference on , vol., no., pp.287,291, 26-28 April 2013.

[9] Tufekci, P.; Uzun, E.; Sevinc, B., ”Text classification of web based news arti- cles by using Turkish grammatical features,” Signal Processing and Commu- nications Applications Conference (SIU), 2012 20th , vol., no., pp.1,4, 18-20 April 2012.

[10] Agarwal, S.; Singhal, A.; Bedi, P., ”Classification of RSS feed news items using ontology,” Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on , vol., no., pp.491,496, 27-29 Nov. 2012.

[11] Burkepile, A.; Fizzano, P., ”Classifying RSS Feeds with an Artificial Immune System,” Information, Process, and Knowledge Management, 2010. eKNOW ’10. Second International Conference on , vol., no., pp.43,47, 10-15 Feb. 2010.

[12] Wikipedia contributors. ”Document classification.” Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 14 Jan. 2015. Web. 18 Feb. 2015.