IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. III (Jan – Feb. 2016), PP 22-26
www.iosrjournals.org

News Classification and Its Techniques: A Review

Gurmeet Kaur[1], Karan Bajaj[2]
[1](Chitkara University,Himachal Pradesh,India)
[2] (Chitkara University,Himachal Pradesh,India)

Abstract : *Text mining has gained quite a significant importance during the past few years. Data, now-a-days is available to users through many sources like electronic media, digital media and many more. This data is usually available in the most unstructured form and there exists a lot of ways in which this data may be converted to structured form. In many real-life scenarios, it is highly desirable to classify the information in an appropriate set of categories. News contents are one of the most important factors that have influence on various sections. In this paper we have considered the problem of classification of news articles. This paper presents algorithms for category identification of news and have analysed the shortcomings of a number of algorithm approaches.*

Keywords: *Classification model, News classification,Text classifiers, Text Mining*

I. Introduction

There exists a large amount of information being stored in the electronic format. With such data it has become a necessity of such means that could interpret and analyse such data and extract such facts that could help in decision-making. Data mining which is used for extracting hidden information from huge databases is a very powerful tool that is used for this purpose. News information was not easily and quickly available until the beginning of last decade. But now news is easily accessible via content providers such as online news services. A huge amount of information exists in form of text in various diverse areas whose analysis can be beneficial in several areas. Classification is quite a challenging field in text mining as it requires prepossessing steps to convert unstructured data to structured information. With the increase in the number of news it has got difficult for users to access news of his interest which makes it a necessity to categories news so that it could be easily accessed. Categorization refers to grouping that allows easier navigation among articles. Internet news needs to be divided into categories. This will help users to access the news of their interest in real-time without wasting any time. When it comes to news it is much difficult to classify as news are continuously appearing that need to be processed and those news could be never-seen-before and could fall in a new category. In this paper a review of news classification based on its contents and headlines is presented. A variety of classification has been performed in past, their performance and various flaws have also been discussed.

II. News Classification Process Workflow

There are different steps involved in news classification. Classification is a difficult activity as it requires pre-processing steps to covert the textual data into structured form from the un-structured form. Text classification process involves following main steps for classification of news article. These steps are data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures.

2.1 News Collection

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, radio, television and World Wide Web and many more. But with the widespread network and information technology growth internet has emerged as the major source for obtaining news. Data may be in available in any format i.e. it may in .pdf, .doc, or in .html format.

2.2 News Pre-processing

After the collection of news text pre-processing is done. As this data comes from variety of data gathering sources and its cleaning is required so that it could be free from all corrupt and futile data. Data now needs to be discriminated from unrelated words like semicolon, commas, double quotes, full stop, and brackets, special characters etc. Data is made free from those words which appear customarily in text and are known as stop words.

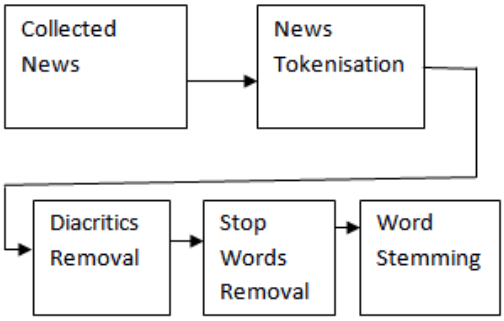

Fig: News Pre-processing

I. News Tokenisation

News tokenization involves fragmenting the huge text into small tokens. Each word in the news is treated as a string. The output of this step is treated as input for the next steps involved in text mining.

II. Stop Word Removal

The stop words language specific and does not carry any information. It generally includes conjunctions, pronoun and prepositions. They are contemplated of low worth and are removed eventually. These words need to be percolate before the processing of data.

Stop words can be removed from data in many ways. There removal can be on the basis of concepts i.e. the removal will be of the words which provide very fewer information about classification.

Another way of removal of stop words is the removal of the words that are present in the list of English stop words. The list is made up of approx 545 stop words and is provided by Journal of Machine Learning Research.

Stop words can also be abolished depending upon the frequency of their occurrence. In this method frequency of occurrence of words is computed and then weights are assigned to words. Then depending on these weights the stop words are dropped.

III. Word Stemming

After the removal of stop words the next activity that is performed is stemming. This step reduces a word to its root. The motive behind using stemming is to remove the suffixes so that the number of words would be brought down. For example the words like user, users, used, using all can be reduced to the word "USE". This will reduce the required time and space.

For stemming there exists many stemmers like S-Stemmers, Lovins Stemmer , Porter Stemmer , Porter Stemmer, Paice/Husk Stemmer. Among these stemmers M.F. Porter is mostly used.

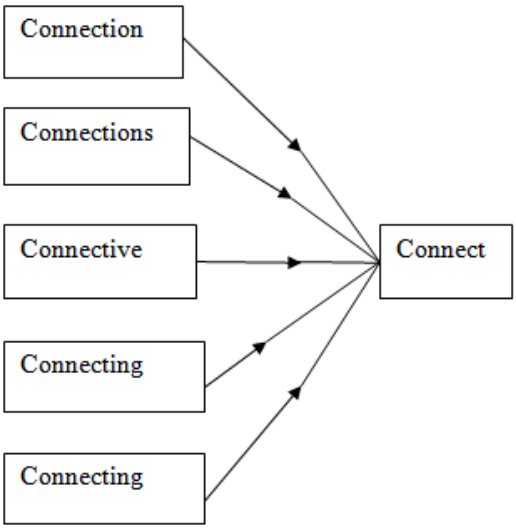

Fig: Stemming Process

S-Stemmer: This stemmer is relevant to the words whose length is greater than three. The motive of this stemmer is to contemplate both forms of news i.e. singular and plural.

Lovins Stemmer: This was the first stemmer proposed. Lovins stemmer has an edge over various other stemmers and that is its speed. It is faster than other stemmers. This stemmer has 294 endings,29 conditions and 35 transformation rules. Initially longest ends are discovered which satisfies the condition and are removed and then 35 rules are applied to transform the ending.

Porter Stemmer: This is the most widely used stemmer because of its high precision and simple algorithm. It includes five steps which are easy to implement and are applied to each and every word.

Paice/Husk stemmer: This stemmer implements an algorithm which is based on iteration and implements same rules and suffixes for each loop. Each rule has five steps among whom three are compulsory to implement while rest two are voluntary.

2.3 Feature Selection

When there exist a large number of features and each of the features is a well known descriptive word for each class, a lot of time may be required in classification and it may be possible that expected accuracy may not be achieved and to overcome these issues, a process named as feature selection is adopted in which only those relevant and highly effective features are chosen, which may prove more noticeable for better news classification. A large number of techniques exists in literature for selecting appropriate features like Boolean weighting, Class Frequency Thresh holding, Term Frequency Inverse Class Frequency, Information Gain.

I. Boolean Weighting

Boolean Weighting looks for each word present in the news. The presence or absence of the words will depend on the word frequencies across the classes separately. Boolean weighting has two possible results. If the word appears in class it is assigned a value 1 or 0 otherwise. This method is not a very useful technique in every case because there exists a lot of irrelevant and useless words.

II. Information Gain

A common technique for feature selection is Information Gain. Information gain predicts if a word appears or not in a certain news. The presence or absence of words empowers us to select features easily which makes classification more robust and reliable.

III. Term Frequency Inverse Class Frequency

Term Frequency Inverse Class Frequency is another method of handling the class frequency issues It is used for feature selection but also provides a means to eradicate stop words.

Equation for Term Frequency Inverse Class Frequency is:

$TF\text{-}ICF_w = TF_w * ICF_w$

Where TF is a term frequency of a word "w" and ICF is an inverse class frequency of a word "w".

IV. Class Frequency Thresh holding

Class frequency threshold gives all those news classes which hold all the words at least once as output. The major use of this technique is that we can remove all those words completely which have class frequencies less than a given threshold value. The issue with this technique is that it removes those words which are important but have less frequency.

2.4 News Classification

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories. The most common news classification methods are Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, Support Vector Machines, K-Nearest Neighbours

I. Naive Bayes

Naive Bayes is a problistic classifier based on text features. It calculates class labels and probability of classes. Naive bayes isn't made up of a single algorithm for classification but it includes a large number of algorithms that work on a single principal for training classifiers and the principal states that the value of a particular feature is autonomous of value of any other feature specified in a class. In the past classification of news article naive bayes were used. But due to its incorrect parameter assessment revamped accuracy was reported. The best thing about Naive bayes algorithm is that it works equally well on both textual as well as numeric data and it is easy to implement and calculate. But it shows poor performance when the features are correlated like short text classification.

II. Support Vector Machines

SVM has been used a lot for news text classification.SVM has a unique feature that it includes both negative and positive training sets which is generally not preferred by other algorithms.

III. Artificial Neural Networks

This network drew its concepts from neurons in which huge calculations are performed very easily by providing sufficient input and are used to estimate functions which are based on large number of inputs. Neural network when used with Naive Bayes presented a new approach known as Knowledge based neural network which is efficient in managing noisy data as well as outliers. Artificial neural network yields good results on complex domains and allows performing fast testing. But the training process is very slow.

IV. Decision Tree

Decision tree is a classifier for text categorization represented in form of a tree in which each node can act as leaf or decision node. Decision tree can make appropriate decisions in situations where decisions are to be taken quickly and allowing slight delay may lead to significant difficulties.

Decision Trees are quite easily perceived and rules can be easily produced through them. Decision Trees can be used to solve intricate problems very easily. It comes with a clause that training decision tree is an expensive task. Besides this one news can be connected to one branch only. If there occurs a mistake at the higher upper level it can cause the whole subtree go invalid.

V. K-nearest neighbors.

K-nearest neighbors is a simple algorithm and a non-parameterized way of classification and regression in case of pattern recognition. For using this algorithm we need to refer K-similar text documents. It reckons the similarity against all documents that exists in the training set and uses it for making decisions about presence of class in the desired category. Neighbour that have same class are the most probable ones for that class and the neighbours with highest probability are assigned to the specific class. K-nearest neighbours is effective and non-parameterized algorithm. The biggest pitfall is that it requires a lot of classification time and it is also difficult to find a optimal value of K.

K-nearest neighbour is a type of lazy learning where function Generalization beyond the data is delayed until a query is made to the system. K-nearest neighbour is one of the simplest machine learning algorithms.

III. Conclusion

A review of news classification is bestowed in this paper. All the steps i.e. pre-processing, document indexing, feature selection, and news headlines classification are examined in detail. In addition, stop words filtering using frequency based stop words removal approach is also discussed. In future these algorithms can be tested on larger corpora. Moreover these algorithms can be improved so that efficiency of categorisation could be improved. A combination of algorithm can be used in order to achieve clustering in a faster way.

References

Journal Papers

[1] Punitha, S. C., and M. Punithavalli. "Performance evaluation of semantic based and ontology based text document clustering techniques." , Procedia Engineering, 30, 2012,100-106.

[2] Chen, Chun-Ling, Frank SC Tseng, and Tyne Liang. "An integration of WordNet and fuzzy association rule mining for multi-label document clustering." , Data & Knowledge Engineering, 69(11),2010,1208-1226.

[3] Hassan, Malik Tahir, et al. "CDIM: Document Clustering by Discrimination Information Maximization." , Information Sciences, 316,2015,87-106.

[4] Li, Yanjun, Soon M. Chung, and John D. Holt. "Text document clustering based on frequent word meaning sequences." ,Data & Knowledge Engineering,64(1),2007,381-404.

[5] Luo, Congnan, Yanjun Li, and Soon M. Chung. "Text document clustering based on neighbors." ,Data & Knowledge Engineering, 68(11),2009,1271-1288.

[6] Yan, Yang, Lihui Chen, and William-Chandra Tjhi. "Fuzzy semi-supervised co-clustering for text documents." ,Fuzzy Sets and Systems, 215,2012,74-89.

[7] Zheng, Hai-Tao, Bo-Yeong Kang, and Hong-Gee Kim. "Exploiting noun phrases and semantic relationships for text document clustering ,179(13) ,2009,2249-2269.

[8] Kirange, D. K. "Emotion classification of news headlines using SVM." , asian journal of computer science & information technology 2(5), 2014

[9] Ramasubramanian, C., and R. Ramya. "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm.", International Journal of Advanced Research in Computer and Communication Engineering 2(12),2013

[10] Bracewell, David B., et al."Category classification and topic discovery of Japanese and English news articles." , Electronic Notes in Theoretical Computer Science, 225, 2009,51-65.

Proceedings Papers

[11] Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar. "News Classification Based On Their Headlines: A Review" 17th IEEE International Conference on Multi-Topic Conference (INMIC), Karachi,Pakistan,,2014,211-216

[12] Drury, Brett, Luis Torgo, and J. J. Almeida. "Classifying news stories to estimate the direction of a stock market index." , IEEE 6th Iberian Conference onInformation Systems and Technologies (CISTI),Chaves ,Portugal, 2011,1-4