International Journal of Management, IT & Engineering
Vol. 13 Issue 02, February 2023,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

VARIOUS METHODS OF NEWS CLASSIFICATION:A Review

Harmandeep Kaur(Assistant professor , Rayat Bahra University, Mohali)

Ranvir Kaur(Assistant professor , Rayat Bahra University, Mohali)

ABSTRACT:

Data mining has gained quite a significant importance during the past few years. Now a days data is available through many sources like electronic media, digital media and many more.

Mostly data is available in unstructured form but there are various ways to convert the data in to structure form. In real life it is highly desirable to classify information into various sections. News contents are the most significant factors that have influence on various sections. In this paper we have considered the problems of classifications of news and discuss some algorithms for classification of news.

I INTRODUCTION:

Data mining is process of discovering interesting knowledge such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in database, data warehouse, or other information repositories. Data to the wide availability of huge amount of data in electronic form, and imminent need for turning such data into useful information and knowledge for broad application including market analysis, business management and decision support, data mining has attracted a great deal of attention in information industry in recent year. Data mining has popularly treated as synonym of knowledge discovery in database, although some researchers view data mining as an essential step of knowledge discovery. A knowledge discovery process consist of an iterative sequence of following step .

1.Data cleaning: which handle noisy, missing or irrelevant data.

2.Data integration: it related with integrated the multiple, heterogeneous data into one

3.Data transformation: in this data is transformed into form appropriate for mining by performing some aggregate operations.

4.Data mining: It is essential process , where intelligent methods are applied for extract data patterns.

5.Pattern evaluation :it identify the truly interesting pattern represent knowledge based on some interesting measures.

6.Knowledge presentation: in this various visualization and knowledge representation techniques are used to present the mined knowledge to the users.

II TEXT CLASSIFICATION

Mostly the information store in the form of text like e mails, web pages, newspaper article, market research reports, complaint letter from customer and internally generated reports. News papers provide news under various categories like national, international, politics, finance, sports, entertainment etc. Text classification is also an important part of text mining . Text classification based on expert knowledge, how to classify the document under the given set of categories. Data mining classification start with training set of document that are already label with class. Text classification has two flavors as single label and multi label. A single label document is belong to only one class and multi label document may be belong to more than one class. Data stored in most text databases are semi structure data in that they are neither completely unstructured not completely structured. For example a document may contain a few structured field such as title, authors, publication date, category But also contain some largely unstructured text components such as abstract, contents.

III NEWS CLASSIFICATION PROCESS

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.

A.NEWS COLLECTION

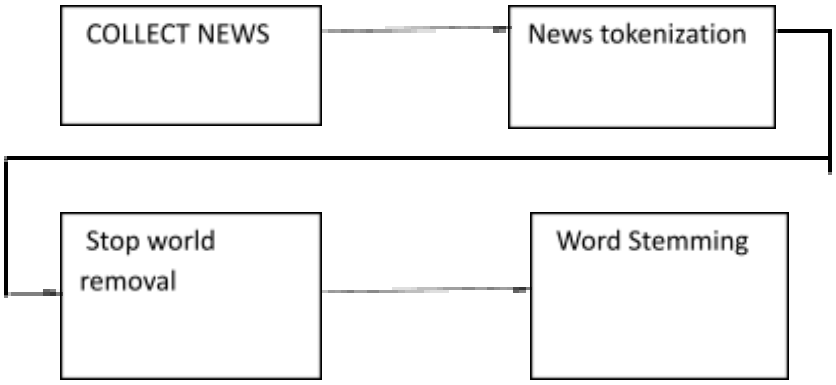
-

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.

B.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.



1.News Tokenization:

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.

2.Stop word removal:

In stop word mostly 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.
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.

3.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.
NEWS CLASSIFICATION:
The next phase is the classification phase which is an important phase in which the aim is to classify the news to their respective categories. The most common news classification methods are Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, K-Nearest Neighbours.

Vol. 13 Issue 02, February 2023,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

1.Naive Bays:
Naive Bays is a problistic classifier based on text features. It calculates class labels and probability of classes. It 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 bays were used. The best thing about Naive bays 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.

2.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. In this positive data is represent as 1 and negative data represent as 0.it is used to distinguish the keywords.

3.Artificial neural network:
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 Bays 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

4.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 problems very easily.

5.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.

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:

1.Hyeran Byun1 and Seong-Whan Lee2,"Applications of Support Vector Machines for Pattern Recognition: A Survey,"SVM 2002, LNCS 2388, pp. 213-236, 2002.
2.D. Morariu, R. Cre¸tulescu and L. Vin¸tan,"Improving a SVM Meta-classifier for Text Documents by using Naïve-Bayes,"Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844.
3.Krishnalal G, S Babu Rengarajan, K G Srinivasagan ,"A new text mining approach based on HMM -SVM for web news classification,"International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19,2010.
4.Vandana Korde, C Namrata Mahender,"Text classification and classifier a survey," International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.
5.Mita K. Dalal, Mukesh A.Zaveri,"Automatic text classification,"International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.
6.Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad,"Financial news classification using SVM",International Journal of Scientific and Research Publications, Volume 2, Issue 3, March 2012 .
7.Chee-Hong Chan Aixin Sun Ee-Peng Lim,"Automated Online News Classifcation with Personalization,"4th international conference on asian digital libraries, Dec 2001.
8.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.
9.Hassan, Malik Tahir, et al. "CDIM: Document Clustering by Discrimination Information Maximization." , Information Sciences, 316,2015,87-106.
10.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.
11. 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