A Review on different classification techniques for text classification

Megha Singla,Punjabi university, Patiala, Punjab, India.

Brahmaleen K.Sidhu, Punjabi university, Patiala, Punjab, India.

Abstract: In current era there are various applications related to social media or it may be related to certain large organization that are producing large amount of data. This data can be text data, images, videos etc. For any processing engine it will be very difficult to store and process the data generated from various sources. Text classification is one such solution to reduce the problem of the size of the text. It can be considered as the automatic way of classifying the text data into multiple classes. Each class of the text will be having single type of information. For any processing engine such as sentiment analyzer, it will be very easy to process single type of data for identification of the sentiment of the text data. There are various techniques and tools that are being used for the text classification. Each technique followed by the researcher has its own set of drawbacks and advantages. These require large amount of further research on the text classification for increasing the success rate.

Keywords: text classification, indexing, dataset of news

I. INTRODUCTION

Automatic text classification always remains the fascinating research topic for various researchers. This text classification come into the existence as the digital documents comes into the picture.

In current time there are various application of the digital world which are producing large amount of multimedia data. This multimedia data is so large that it is very difficult to process the data while having single class. For the reduction of the size of the text data there is a process required called as text classification. Various techniques based on machine learning exists which can be used for the classification with higher rate of success.

We can categories the text classification into two categories.

a. Topic based text classification.
b. Genre based text classification.

a. **Topic based classification** is the first type of classification. It classifies the text into multiple sub categories. These sub categories are pre defined sub categories. For example

S is the set of the text taken as sample for the classification.

C(c1,c2,c3…….cn) are the predefined sub classes for the text classification

Si is the ith set of the total sample.

Si->Ck means allocating the sub class Ck to the set Si.

This process of classification goes on until the set S gets empty. Classification of the text based on pre defined classes are further of two types.

a. Soft classification.
b. Hard classification

a. In soft classification each class is ranked as per the classification. For example high, Medium and low. High means the class text is highly related, medium is the medium level relation and low is the low level relation. This means the level of the relation here helps in applying the class based on the requirements.

b. Hard classification is another type of traditional classification of the text. It considers the pre set classes for the classification. Whole sample of the dataset containing text will be classified into these categories. Sometimes hard classification wrongly classify the text with weak belongings to the class.

b. **Genre classification** is sophisticated type of classification process. It usually classifies the text based on document type. For example conference paper, journal paper, book topic, news article, speaker notes. Genre classification will helps in applying the rules for editing, building and format styles of the document. In current time various publishers are getting the documents containing text related to different Genre. There is high requirement for the document classification based on the Genre of the document. For example

S is the sample of the documents

G is the categories of the Genre which are pre set.

C->{g1,g2,…….gk,……..gn} Take

element si from S. Classify si->gk

This will proceed till the set is empty.

1.1 General steps for the text classification

a. Input the sample set.

b. Pre processing of the document from the sample set.

c. Features selection.

d. Classify the document based on features.

Input the sample set is regarding entering the text documents whose classification is to be performed. This sample set contains random documents. All the documents can be collected from the single sources or from different sources.

Preprocess the document text: It involves tokenization of the text contained into the document. All the unrequired words like stop words are to be removed. This will reduce the processing time and also increase the efficiency of the results.

Features selection: Once all the normalization process is undertaken, there comes the features selection. These features are the base features on to which classification has to be performed.

Classify: Based on some Machine learning technique perform the classification of the text. The classification technique of the text depends upon the requirement of the researcher.

1.2. Features selection techniques

a. Term Frequency-Inverse document frequency

It is the most popular scheme. Where each term frequency is calculated in the document.

Like $TF_{ij}=f_{ij}/Max_k f_{kj}$

In this the fij is the frequency of the term j in the document i. where the maxk is the frequency for the most common term. It is the kth term.

$IDF_i=\log_2((N+1)/(n_i+1)+1$

In this N is the total number of the text document. ni is the documents count which have I term.

$TFIDF_i=TG_{ij}.IDF_i$

b. Singular value decomposition

It is calculated after the term frequency-inverse document frequency. This is done using

$A_{mn}=U_{mn}* S_{mn} * V^T_{nn}$

II. LITERATURE SURVEY

Rini Wongso(2017) et. al: author in this paper has worked on the text classification into multiple classes. Each class will be having pre set features. The whole process is done using technique which is the combination of two techniques that is the TF-IDM and SVD. Success rate for the classification using this combined approach is much better compared to the other individual techniques. The whole process of the classification is spanned into various sections or phases. First is the input the text dataset, second step includes removing the noise in the text and then after features extraction. In last the classification is performed onto the features. The dataset is taken with Indonesian languages with accuracy of 85%. Wen Zhang(2011) et. al: in this paper author has worked on the comparison of the TF*IDD and the LSI based techniques. According to study two techniques for the classification for the text is used. The performance of the LSI is better compared to the TF*IDM. The score allocated to the words into the text is not biased using LSI. The scheme of the scores allocations is done by the LSI using merit based process.

Davood Mahmoodi(2011) et. al: author in this paper has proposed SVM based classification technique. It uses the dataset of the Persian based language. They have classified the dataset of the Persian into three categories. Small set is sub divided into the training set and remaining elements are kept as testing set. They have achieved the accuracy of 98.67% for the true classification.

Hao Lin (2014): author in this paper has proposed a classification process for the text being mined using any of the mining technique. For the classification they have used Naïve Bayes based classification. The result generated is much better compared to the SVM. Author in this paper has mainly focused on the efficient way of the classification. The energy lost while classification should be minimized. That technique has to be implemented which is much efficient technique compared to the other techniques. Thus Naïve Bayes is the best technique for the classification of the text.

Krina Vasa (2016): author in this paper has studied the need for the text classification. Text classification is important process as far as current data need is there. These require various types of classification tools which can classify the text to summarize the text for various

applications like medical diagnosis, sentiment analysis. Researcher has studied various research techniques which are based on machine learning and statistical classification techniques like K-nearest neighbor, Naïve Bayes etc.

Vangelis Metsis(2006) et. al: Author in this paper has studies the technique based on Naïve Bayes on the dataset having different types of the messages. These messages are having various spam messages. Using Naïve Bayes the classification of the text messages are performed. These text messages are classified into two categories. One is the true messages and other are the spam messages. This will enhance the security for the system to recognize the spam messages.

III. COMPARATIVE ANALYSIS

Author Name	Year	Technique	Constraints
Rini Wongso et. al	2017	Naïve Bayes, and Su	To test the technique for the language dataset which is tested by the experts.
Davood Mahmoodi	2011	SVM	SVM is included for the optimization problem equation which is time consuming process. So better equation can be developed which can reduce the time complexities.
Wen Zhang	2011	TFIDF, LSI	These require some evaluation methods that can determine the accuracy for the indexing technique.

IV. CONCLUSION

In current digital word there are various applications which are producing tons of data. This data will be very difficult to process due to its size. We want to overcome the complexities of the size by having classification of the text into multiple sub classes. Each class will be having its own set of the text based on the features. Various researchers are using various types of techniques for the text classification. TFIDF is the best technique used with higher accuracy for the text classification. This

accuracy can be enhanced and the dataset can be further increased by adding multiple classes into the dataset. The technique can be enhanced by including the text classification using multi SVM based technique. Dataset can be enhanced by adding sports and other sub categories to check the result accuracy for the current TFIDF, LSI techniques and expected technique of multi SVM based.

V. FUTURE WORK

In current time the text classification is the area for the research by various researchers. There are various techniques with different success rate that has been applied onto the system of classification. In the whole system the results can be enhanced by hybridization of the technique and also enhancing the dataset to test the results.

References

Rini Wongso*, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, Rudy," News Article Text Classification in Indonesian Language", ICCSCI,issue:116, pp:137-143,2017.

Wen Zhang a,⇑ , Taketoshi Yoshida b , Xijin Tang c," A comparative study of TFIDF, LSI and multi-words for text classification", Expert Systems with Applications**,issue 38, pp:2758- 2765,2011.**

Davood Mahmoodi1 , Ali Soleimani1 , Hossein Khosravi1 , Mehdi Taghizadeh2," FPGA Simulation of Linear and Nonlinear Support Vector Machine", Journal of Software Engineering and Applications,issue 4,pp:320- 328,2011.

Hao Lin," Research on Energy-Efficient Text Classification", ICITEC, 2014.

Krina Vasa," Text Classification through Statistical and Machine Learning Methods: A Survey", IJEDR,vol. 4,issue 2, pp:655-658,2016.

Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras," Spam Filtering with Naive Bayes – Which Naive Bayes?",issue 27-28, 2006.

C. C. Aggarwal and C. Zhai, Mining Text Data, 2012.

M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.

R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.

R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Trans. Knowl. Data Eng., vol. 27, no. 9, pp. 2349-2361, 2015.

A. Chaudhuri, "Modified fuzzy support vector machine for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.

E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," Tkde, vol. 25, no. 12, pp. 2780- 2795, 2013.

X. Fang, "Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive," vol. 25, no. 10, pp. 2302-2314, 2013.

C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K- nearest neighbor and support vector machine," Expert Syst. Appl., vol. 39, no. 15, pp. 11880-11888, 2012.

L. H. Lee, R. Rajkumar, and D. Isa, "Automatic folder allocation system using Bayesian-support vector machines hybrid classification approach," Appl. Intell., vol. 36, no. 2, pp. 295-307, 2012.