A Survey on Various Features and Techniques of Text Content Classification

Vishal Sahu
M.Tech Scholar

Vivek Kumar
Assistant Professor

ABSTRACT
Traditional information retrieval methods become inadequate for increasing vast amount of data. Without knowing what could be in the documents; it is difficult to formulate effective queries for analyzing and extracting useful information from the data. This survey focused on some of the present strategies used for filtering documents. Starting with different types of text features this paper has discussed about recent developments in the field of classification of text documents. This paper gives a concise study of methods proposed by different researchers. Here various pre-processing steps were also discussed with a comprehensive and comparative understanding of existing literature.

Keywords
Content filtering, Fake Profile, Online Social Networks, Spam Detection.

1. INTRODUCTION
Unstructured data remains a challenge in almost all data intensive application fields such as business, universities, research institutions, government funding agencies, and technology intensive companies. Eighty percent of data about an entity (person, place, or thing) are available only in unstructured form. They are in the form of reports, email, views, news, etc. Text mining/ analytics analyzes the hitherto hidden relationships between entities in a dataset to derive meaningful patterns which reflect the knowledge contained in the dataset. This knowledge is utilized in decision making [1]. Text analytics converts text into numbers, and numbers in turn bring structure to the data and help to identify patterns. The more structured the data, the better the analysis, and eventually the better the decisions would be. It is also difficult to process every bit of data manually and classify them clearly. This led to the emergence of intelligent tools in text processing, in the field of natural language processing, to analyze lexical and linguistic patterns. Clustering, classification, and categorization are major techniques followed in text analytics [2]. It is the process of assigning, for example, a document to a particular class label among other available class labels like "Education", "Medicine" and "Biology". Thus, text classification is a mandatory phase in knowledge discovery [2]. The aim of this article is to analyze various text classification techniques employed in practice, their spread in various application domains, strengths, weaknesses, and current research trends to provide improved awareness regarding knowledge extraction possibilities.

Whole of this paper are sorted out as following: in the second area, the necessity of text features were also examined. Third section list various techniques adopt by researcher to increase the classification accuracy. While fourth section provide related work of the current approaches applied by different researchers to correct class of document. Research problem is pointed out, and then the proposed problem is formalized in detail. The conclusion of the whole paper is made in the last section.

2. FEATURES OF DOCUMENTS
1) Title feature: The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2) Sentence Length: This features is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belongs to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3) Term Weight: The frequency of the term occurrence with a documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

4) Sentence position: Whether it is the first 5 sentence in the paragraph, sentence position in text gives the importance of the sentences. This features can involve several items such as the position of the sentence in the documents, section and the paragraph, etc, proposed the first sentence of highest ranking. The score for this features in [6] consider the first 5 sentence in the paragraph.

5) Sentence to sentence similarity: This feature is a similarity between sentences for each sentence S , the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are represented as the vector. The similarity of each sentence pair is calculated based on similarity.

3. TECHNIQUES OF DOCUMENT CLASSIFICATION
As document is collection of paragraphs. Paragraphs are collection of sentences. While sentences are collection of words. So whole preprocessing focus on word in the document without any punctuations. So in pre-processing of document there are two common steps first is stop word removal, and second is stem word removal [7]. Each dataset in research need some pre-procesing steps, so text mining have following set of features:

Stop Word Removals: As sentence is frame with number of words but some of those words are just use to contruct a proper sentence although it does not make any information in the sentence. So identification of those words then removing is term as Stop word removal. So a list of words is store by the researcher which help in identifying of stop words. This removal of stop words help in reduce the execution time of the algorithm, at the same time noisy words which not give any fruitful information is also removed. Stop words are like {a, the, for, an, of, and, etc.}. So text document is transform into collection of

words which is then compare with these words and then each match word is removed from the document.

Inorder to understand this assume an sentence {India is a great country in the world} then after pre-processing it become {India, great, country, world} while stop words {is, a, in, the} in the sentence are removed. Let

Stem Word Removal In this words which are almost similar in prefix are replace by one word. This can be said collection of words share same word is term as stem. So there occurrence in the document make same effect but while processing in text mining algorithm it make different so update each word from the collection into single word is done in this stem word removal pre-processing step. Let us assume an collection of words for better understanding of this work. Collection of word is {play, plays, playing} then replace each with word {play}.

Some techniques of text document classification are list:

K-Nearest Neighbors

K-NN classifier is a case-based learning [8] algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measure's This method is try for many application [9] Because of its effectiveness, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k .The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct.

Naïve Bayes

Naïve bias method is kind of module classifier [10] under known priori probability and class conditional probability .it is basic idea is to calculate the probability that document D is belongs to class C. There are two event model are present for naive Bias as multivariate Bernoulli and multinomial model. Out of these model multinomial model is more suitable when database is large, but there are identifies two serious problem with multinomial model first it is rough parameter estimated and problem it lies in handling rare categories that contain only few training documents.

SVM

The application of Support vector machine (SVM) method to Text Classification has been propose by [11]. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector.

Neural Network

A neural network classifier is a network of units, where the input units usually represent terms, the output unit(s) represents the category. For classifying a test document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit(s) takes up as a consequence determines the categorization decision. Some of the researches use the single-layer perceptron, due to its simplicity of implementing [12]. The multi-layer perceptron which is more sophisticated, also widely implemented for classification tasks.

Voting

In [13] algorithm is based on method of classifier committees and is based on idea that given task that requires expert opinion knowledge to be performed. k experts opinion may be better than one if their individual judgments are appropriately combined. Different combination rules are present as the simplest possible rule is majority voting (MV)If two or three classifiers are agree on a class for a test document, the result of voting classifier is that class. Second weighted majority voting, in this method, the weights are specific for each class in this weighting method, error of each classifier is calculated.

Centroid based classifier

The centroid-based classification algorithm is very simple. [14] For each set of documents belonging to the same class, we compute their centroid vectors. If there are k classes in the training set, this leads to k centroid vectors (C1, C2, C3...) where each Cn is the centroid for the jet class. The class of a new document x is determined as, First the document-frequencies of the various terms computed from the training set Then, compute the similarity between x to all k centroid using the cosine measure. Finally, based on these similarities, and assign x to the class corresponding to the most similar centroid

4. RELATED WORK

Seyyed M. H. Dadgar et al. (2016) [15] aimed to classify news to different categories, using SVM classification technique. They used TF-IDF and SVM classifier in their work. They followed the process of text preprocessing, feature extraction on the basis of TF-IDF and at the end classification by using SVM classifer. They evaluated their results using two different data sets os BBC news and 20newsgroup dataset. The precision was 97.84% & 94.93% for both the datasets respectively.

Adel M Hamdan et al. (2016) [16] came up with a comparative study for text classification. They took Arabic language textual documents as their data set. They performed the task of text classification using three different techniques which are SVM classifier, Multilayer perceptron Neural Networks and Naïve Bayes classifier. They end up with the result that SVM technique is the best classifier for text classification while using Arabic text documents.

June Ling O. et al. (2017) [17] came up with an idea of analyzing & determining the relevant news content on the basis of sentiment-based classifier. They took 250 English news documentaries which are in text form as their dataset. These text documents were labeled with different sentiments on the basis of which classification is done. They used knn approach in their work to classify the new content.

Roopesh S. et al. (2017) [18] proposed a work to classify multiclass documents in the case of text documents. They used Naïve Bayes classifier to solve the problem of text classification. The approach is first applied in linear and then in hierarchical manner to get the efficient results. They concluded that hierarchical approach is more effective as compared to linear one. They improved the accuracy & efficiency of the classifier.

In [19] propose to raise the problem of automatic classification of scientific texts as an optimization problem, which will allow obtaining groups from a data set. The use of evolutionary algorithms to solve classification problems has been a recurrent approach. However, there are a few approaches in which classification problems are solved, where the data attributes to be classified are text-type. In this way, it is proposed to use the association for computing machinery taxonomy to obtain the similarity between documents, where each document consists of a set of keywords.

So following are problem identified in this work:

- Document should not be in particular format before inserting in proposed algorithm [3].
- Reduce document classification algorithm execution time by arranging data in specific structure.
- Improve accuracy of the predicted class of document.
- Security of the data owner dataset before classification was done by AES algorithm.
- Cluster document storage privacy maintain by placing encrypted keyword for searching.

5. CONCLUSIONS

Text Classification using analytical approach project proposed a design of the application that can effectively classify text files into appropriate folder depending upon the theme of the file, using the training data to model the classifier. So this paper have summarize current methodologies that have been basically created. Here it was obtained that people develop high social networking sites than create various document set. It was obtained that most of work use clustering techniques for segregating content from other class of contents. In future it is desired to develop the highly accurate algorithm which not only detect the spam but spammer profile as well.

6. REFERENCES

[1] Brindha, S., Sukumaran, S., & Prabha, K. (2016). A survey on classification techniques for text mining. Proceedings of the 3rd International Conference on Advanced Computing and Communication Systems. IEEE. Coimbatore, India.

[2] Vasa, K. (2016). Text classification through statistical and machine learning methods: A survey. International Journal of Engineering Development and Research, 4, 655-658.

[3] Farman Alia, Kyung-Sup Kwaa,Yong-Gi Kimb," Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification", Applied Soft Computing-2016.

[4] Isidro Peñalver-Martinez, Francisco GarciaSanchez, Rafael Valencia-Garcia," Featurebased opinion mining through ontologies", Expert Systems with Applications-2014.

[5] RuiXia,FengXu,JianfeiYu," Polarity shift detection, elimination and ensemble: A three stage model for document-level sentiment analysis" Information Processing and Management 52 (2016) 36– 45.

[6] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu," Sentence Compression for spect-Based Sentiment Analysis" IEEE/ACM transactions on audio, speech, and language processing, vol. 23, no. 12, December 2015.

[7] Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1/13/$31.00 ©2013 Ieee.

[8] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", Proc. ODBASE pp- 986 – 996, 2003

[9] Eiji Aramaki and Kengo Miyo, "Patient status classification by using rule based sentence extractionand bm25-knn based classifier", Proc. of i2b2 AMIA workshop, 2006.

[10] SHI Yong-feng, ZHAO, "Comparison of text categorization algorithm", Wuhan university Journal of natural sciences. 2004.

[11] Joachims, T. "Text categorization with support vector machines: learning with many relevant features". In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137–142 1998.

[12] MIgual E .Ruiz, Padmini Srinivasn, "Automatic Text Categorization Using Neural networks", Advaces in Classification Research, Volume VIII.

[13] Yiming Yang Christopher G. Chute "A Linear Least Squares Fit Mapping Method For Information Retrieval From Natural Language Texts" Acres De Coling-92 Nantes, 23-28 AOUT 1992

[14] B S Harish, D S Guru, S Manjunath " Representation and Classification of Text Documents: A Brief Review" IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition"RTIPPR, 2010.

[15] [3] Seyyed Mohammad Hossein Dadgar et al "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification" 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016.

[16] [4] Adel Hamdan Mohammad et al "Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network" GSTF Journal on Computing (JOC) ,Volume 5, Issue 1; 2016 pp. 108-115.

[17] [13] Omar Al-Momani, Tariq Alwada et al. "Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study" International Journal of Current Engineering and Technology 2016.

[18] [14] E Jadon, R Sharma et al. "Data Mining: Document Classification using Naive Bayes Classifier" International Journal of Computer Applications (0975 – 8887) Volume 167 – No.6, June 2017.

[19] Alan Díaz-Manríquez , Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. "An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy". accepted March 9, 2018, date of publication March 15, 2018, date of current version May 9, 2018.