Jitendra Yadav. 2019, 7:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey on Digital Document Classification Various Features and Techniques

Jitendra Yadav, Prof. Sumit Sharma, Prof. Pranjali Malviya
Department of Computer Science & Engineering
Vaishnavi Institute of Technology and Science (VTTS), Bhopal, India.
Jtndr.gkg@gmail.com

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.

I. 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.

II. FEATURES OF DOCUMENTS

1. Title include:

The word in sentence that present in title gives high score to be an keyword. This is calculate by checking the quantity of matches between the document

International Journal of Science, Engineering and Technology

An Open Access Journal

word in a sentence and word in the title. In [4] evaluation of the score for this feature was done which is the proportion of number of words in the sentence that happen in the title over the total number of words in the title.

2. Sentence Length:

This feature is valuable to separate short sentence, for example, datelines and writer names ordinarily found in the news articles the short sentences are not expected to has a place in the summary. In [5] utilize the length of sentence, which is the proportion of the quantity of words happening in the sentence over the words happening in the longest sentence of the text documents.

3. Term Weight:

The reappearance of the term event within a text documents has been utilized for computing the significance of sentence. The score of a sentence can be determined as the whole of the score of words the sentences. The score of critical word can be determined by well known text feature TFIDF strategy.

4. Sentence position:

Whether it is the starting five sentence in the passage, sentence position in content gives the significance of the sentences. This features can include a few things, for example, the placement of the sentence in the text documents, segment, passage, and so on, proposed the primary sentence of most noteworthy information. The score for this features in [6] think about the initial five sentence in the selected documents.

5. Sentence to sentence comparability:

This element is a similitude between sentences for each sentence S , the similarity among S and each other sentence is figured by the cosine closeness measure with a subsequent incentive somewhere in the range of 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are spoken to as the vector. The similitude of each sentence pair is determined dependent on closeness.

III. RELATED WORK

In [4] Given approach utilizing nearest neighboring algorithm with cosine analogy to categorize analysis papers and patents revealed in many fields and keep in several conferences and journals information. Conduct experiment results proves that user reclaim outcomes by traversing analysis paper or patent in specific set. The first advantage of given technique is that search space become compact and waiting time for query's resolution has reduced. They need calculated the edge relying upon similarity of terms of question, patent and analysis paper. Threshold calculation wasn't numerical worth primarily based. Thus, the given technique categorized more accurately than active one.

In [5] inspected that social media posts will analyse the personal intelligence. Key base of human behaviour is nature. Nature tests detailed the individual's persona that influences the relations and main concern. Users share their opinions on social media. The text categorization was demoralized to forecast the character and nature on the idea of their comments. Indonesian and West Germanic language were used for this take a look at. Naïve bayes, SVM and K-Nearest Neighbor are performed method for arrangement. Naïve bayes performed higher than different techniques. The analysis work uses Personality dataset. During this dataset used classify the nature based-on an internet queue.

In [6] author navigate web for vast information to collect data. It comprises of big unstructured information like text, image and video. Tricky issue is organization of massive information and gather helpful data that would be utilized in bright computer system. Ontology covers the massive space of topic. To build associate degree ontology with specific domain, massive dataset on net was used and arrangement with specific domain before the completion of organization. Naïve bayes classifier was enforced with Map reduce model to arrange massive dataset. Plant and animal domain articles from encyclopaedia are online easily available for experiment. Planned technique yielded robust system with high accurateness to classify information into domain specified ontology. During this analysis work, datasets use plant and animal domain animal's article in online encyclopedia and Wikipedia as dataset.

International Journal of Science, Engineering and Technology

An Open Access Journal

In [7] projected a Bayesian categorization technique for text categorization utilizing class-specific characteristics. In contrast to regular approaches of text classification planned methodology chosen a selected feature set in each category. Applying such class-dependent characteristics for classification, a Baggenstoss's PDF Projection Theorem was pursued to recreate PDFs from class-specific PDFs and construct a Bayes classification rule. The significance of instructed approach is that feature choice criteria, like: MD (Maximum Discrimination), IG (Information Gain) are enclosed simply. Estimating the performance on much actual benchmark information set and compared with feature choice approaches. The experiments, they tested approach for texture categorization on binary real time benchmarks: 20-Reuters and 20-Newgroups.

In [8] presented a BI-LSTM (Bidirectional long short term memory) network to engrave the short text categorization with a pair of settings. The short-text categorization is needed in applications of text mining, particularly health care applications briefly texts mean linguistic ambiguity bound semantic expression because of that ancient approaches fails to capture actual linguistics of restricted words. In health care domains, the text includes rare words, during which because of lack of training information embedding learning isn't simple. DNN (Deep neural network) is potential to spice up the performance as per their strength of illustration capability. At first, a typical attention mechanism was adopted to lead network training with domain data in wordbook. Secondly, direct cases once data wordbook is out of stock. They gave a multi-task model to find out domain data wordbook and performing arts text categorization task in parallel. They applied instructed technique to existing aid system and completely obtainable ATIS dataset to induce higher results.

In [9] reviewed the method of text categorization and active algorithms. Great amount of information is keep as e-documents. Text mining could be a technique of taking out information from these documents. Categorizing text documents in specific variety of pre-defined categories is Text categorization. Its application comprises of email routing, spam filtering, language identification, sentiment analysis, etc.

IV. TECHNIQUES OF DOCUMENT CLASSIFICATION

As text document is accumulation of sentences. Passages are gathering of sentences. While sentences are gathering of words. So entire preprocessing center around word in the text document with no hard grammatical puntuations. So in pre-handling of text document there are two basic advances first was stopword evacuation, and second was stemming [7]. Each dataset in research need some pre-processing steps, so text mining have following arrangement of steps:

1. 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:

International Journal of Science, Engineering and Technology

An Open Access Journal

1. K-Nearest Neighbors

K-NN classifier is a case-based learning [8] calculation that depends on a separation or closeness work for sets of perceptions, for example, the Euclidean separation or Cosine comparability measure's. This technique was used for some application in [9] because of its viability, non-parametric and simple to usage properties. But this technique have some set of issues like the grouping time is long and hard to discover ideal estimation of number of cluster that means value of k .The best decision of k relies on the information for the most part, bigger estimations of k diminish the impact of noise on the arrangement, yet make limits between classes less particular.

2.Naïve Bayes

Naïve technique is somewhat module classifier [10] under known priori likelihood and class restrictive likelihood .it is essential thought is to figure the likelihood that text document D is has a place with class C. There are two occasion display are available for credulous Bias as multivariate Bernoulli and multinomial model. Out of these model multinomial model is progressively appropriate when database is substantial, yet there are distinguishes two significant issue with multinomial model first it is unpleasant parameter evaluated and issue it lies in taking care of uncommon classes that contain just couple of preparing archives.

3.SVM

The use of Support vector machine (SVM) technique to Text Classification has been proposed by [11]. The SVM need both positive and negative preparing set which are extraordinary for other characterization techniques. These positive and negative preparing set are required for the SVM to look for the choice surface that best isolates the positive from the negative information in the n dimensional space, this was shown in the hyper plane. The text document agents which are nearest to the choice surface are known as the support vector. There is issues with this technique like it don't work well for multiclass dataset.

4. Neural Network

A neural system classifier is a system of units or neurons, where the input units as a rule speak to terms, the last layer neuron(s) speaks to the classification. For identifying the text document class its feature term weight are put in the trained neural network input layer where the output information layer consist of the enactment of these neuron of any type of neural network like feed forward through the system, and the value that the yield unit(s) takes up as a result decides the classification choice. A portion of the researcher utilize the single-layer perceptron, because of its straightforwardness of working [12]. The multi-layer perceptron which is progressively complex, additionally generally actualized for arrangement errands.

5. Voting

In [13] calculation depends on strategy for classifier boards of trustees and depends on thought that given assignment that requires master opinion for learning. Here k number of specialists feeling might be superior to anything one if their individual decisions are properly consolidated. Distinctive mix rules are available as the most straightforward conceivable guideline is lion's share casting a ballot (MV)If a few classifiers are concede to a class for a test text document, the aftereffect of casting a ballot classifier is that class. Second weighted dominant part casting a ballot, in this technique, the loads are explicit for each class in this weighting strategy, mistake of every classifier is determined.

6. Centroid based classifier

The centroid-based characterization calculation is exceptionally basic. [14] For each arrangement of text documents having a place with a similar class, this paper figure their centroid vectors. In the event that there are k classes in the preparation set, this prompts k centroid vectors (C1, C2, C3...) where each Cn is the centroid for the stream class. The class of another text document x is resolved as, First the archive frequencies of the different terms registered from the preparation set Then, figure the likeness between x to all k centroid utilizing the cosine measure. At long last, in view of these likenesses, and relegate x to the class relating to the most comparable centroid.

V. 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.

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] B. Gourav& R. Jindal, "Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold," International Journal of Computer Applications, vol. 33, no. 5. 2011.

[5] B.P. Yudha, and R. Sarrno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," In Data and Software Engineering (ICoDSE), in proceedings od International Conference on, pp. 170-174. IEEE, 2015.

[6] J. Santoso, E. M. Yuniarno, et al., "Large Scale Text Classification Using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building." In Intelligent Human-Machine Systems and Cybernetics (IHMSC), in proceedings of the 7th International Conference on, vol. 1, pp. 428-432. IEEE,2015.

[7] B.Tang, H. He, et al., "A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering 28, pp: 1602-1606,no. 6, 2016.

[8] S. Cao, B. Qian, et al.," Knowledge Guided Short-Text Classification for Healthcare Applications", 2017 IEEE International Conference on Data Mining (ICDM) vol. 2, no. 6,pp: 234-289. 2017.

[9] V. K. Vijayan, K. R. Bindu, et al., "A comprehensive study of text classification algorithms." IEEE Advances in Computing, Communications and Informatics (ICACCI),, vol 12, no. 1 pp: 42-53. 2017.

[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.

International Journal of Science, Engineering and Technology

An Open Access Journal

[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.