International Journal of Scientific Research & Engineering Trends
Volume 5, Issue5, Sep-Oct-2019, ISSN (Online): 2395-566X

A Survey on Techniques and Features of Document Classification

Ph.D Scholar Vinod Sharma
sharma01.scope@gmail.com

Dr. Shiv Shakti Shrivastava
shivshakti18@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. PRE-PROCESSING AND FEATURES OF TEXT MINING

Preprocessing

All words passes to preprocessing level. Irrelevant terms are eliminated there. This process is also called as tokenization process. It consists of two kinds operations such as stop list removal, stem word removal [8].

Stop List Removal: It saves the system resources. Stop word has list of words. That are deemed or irrelevant and then it is removing .It consists of articles (a, an, the), preposition (for, in, at, etc), and so on. A text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces [3]. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of document collection.

Stem Word Removal: The group of different words may share the same word is called as stem. For example drug, drugged, drugs, Different occurrences of the same word. Terms with a common stem would have same meaning. So it is filtering from the concern text documents. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem. A well-known rule based stemming algorithm has been originally proposed by Porter [4]. He defined a set of

© 2019 IJSRET

production rules to iteratively transform (English) words into their stemming algorithm Every word is identified and the word co-occurrences are calculated with a score is calculated for each word.

Features

The main function of document representation phase is to convert terms which are strings to feature IDs which are integers of greater than or equal to 0 [5].

Term Frequency: The TF is the count of category-of-words of every category in each document. So the documents term frequency for a category is the occurrence of the words in single document or article.

Document Term Frequency: It is the number of documents in the collection that contain a term.

IDF: Inverse Document Frequency, is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word.

$$IDF(t) = log\left(\frac{N}{n}\right)$$

N represents the total number of documents in the dataset, n represents the number of documents that term t appears

TF-IDF: TF-IDF[5,6](Term Frequency-inverse Document Frequency), puts weighting to a term based on its inverse document frequency. It means that if the more documents a term appears, the less important that term will be, and the weighting will be less.

$$TFIDF(t) = TF_t * log\left(\frac{N}{n_t}\right)$$

TF-IDF-CF: As per the Shortcomings of TF-IDF has, [5] introduce a new parameter to represent the in-class characteristics, and we call this class frequency, which calculates the term frequency in documents within one class.

$$TFIDFCF(t) = \log(TF_t + 1) * log\left(\frac{N+1}{n_t}\right) * \frac{n_{c,t}}{N_c}$$

the number of documents where term t appears within the same class c document. Nc represents the number of documents within the same class c document.

III. TECHNIQUES OF DOCUMENT CLASSIFICATION TECHNIQUES

Voting

In [6] 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 is explicit for each class in this weighting strategy, mistake of every classifier is determined.

Centroid based classifier

The centroid-based characterization calculation is exceptionally basic. [7] For each arrangement of text documents having a place with a similar class, this paper figures 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.

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.

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.

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

© 2019 IJSRET

vector. There are issues with this technique like it don't work well for multiclass dataset.

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 utilizes the single-layer perception, because of its straightforwardness of working [12]. The multi-layer perceptron which is progressively complex, additionally generally actualized for arrangement errands.

IV. RELATED WORK

In [13] presented an approach using closest neighboring algorithm with cosine analogy to classify research papers and patents published in several fields and stored in different conferences and journals database. Experimented results proves that user get better outcomes by traversing research paper or patent in specific category. The primary advantage of presented technique is that search area become compact and waiting time for query's solution has reduced. They have calculated the threshold depending upon similarity of terms of query, patent and research paper. Threshold calculation was not numerical value based. Hence the presented technique categorize more precisely than existing approach.

In [14] examined that social media posts can analyze the personal intelligence. Primary base of human behavior is personality. Personality tests elaborate the individual's persona that influences the relations and priorities. User reveals their opinions on social media. The text classification was exploited to predict the character and nature on the basis of their comments. Indonesian and English language were used for this test. Naïve Bayes, SVM and K-Nearest Neighbor are executed methods for classification. Naïve Bayes performed better than other techniques. The research work uses My Personality dataset. In this dataset used to classify the personality based-on an online ques

In [15] traversed internet for huge data to gather knowledge. It consists of huge unstructured data like text, image and video. Challenging issue is organization of big data and gathers useful knowledge that could be used in bright computer system. Ontology covers the big area of topic. To construct ontology with specific domain, big dataset on web was used and arranging with particular domain before the completion of organization. Naïve Bayes classifier was implemented with Map reduce model to organize big dataset. Plant and animal domain articles from encyclopedia available online were used to experiment. Proposed technique yielded robust system with high accuracy to classify data into domain specified ontology. In this research work, datasets use plant and animal domain animal's article in online encyclopedia and Wikipedia as dataset.

In [16] presented a Bayesian classification technique for text categorization using class-specific characteristics. Unlike regular approaches of text categorization proposed method chose a particular feature subset in every class. Applying such class-dependent characteristics for classification, a Baggenstoss's PDF Projection Theorem was followed to recreate PDFs from class-specific PDFs and construct a Bayes classification rule. The importance of suggested approach is that feature selection criteria, like: MD (Maximum Discrimination), IG (Information Gain) is included easily. Evaluated the performance on several actual benchmark data set and compared with feature selection approaches. The experiments , they tested approach for texture classification on binary real time benchmarks : 20- Reuters and 20-Newgroups.

In [17] proposed a BI-LSTM (Bidirectional long short term memory) network to inscribe the short text classification with 2 settings. The short-text classification is required in applications of text mining, especially health care applications in short texts mean linguistic ambiguity bound semantic expression due to which traditional approaches fails to capture actual semantics of limited words. In health care domains, the text includes infrequent words, in which due to lack of training data embedding learning is not easy. DNN (Deep neural network) is potential to boost the performance as per their strength of representation capacity. Initially, a common attention mechanism was adopted to guide network training with domain knowledge in dictionary. Secondly, direct cases when knowledge dictionary is unavailable. They presented a multi-task model to learn domain knowledge dictionary and performing text classification task in parallel. They applied suggested technique to existing healthcare system and exclusively available ATIS dataset to get better results.

In [18] surveyed the process of text classification and existing algorithms. Large amount of data is stored as e-documents. Text mining is a technique of extracting data from these documents. Classifying text documents in specific number of pre-defined classes is Text classification. Its application consists of email routing, spam filtering, language identification, sentiment analysis, etc.

In [19] introduced a fuzzy logic based technique to solve text classification. Data inserted in proposed model are extracted from twitter's message. Social media offers plenty of data to study human behavior. Hurricane Sandy 2012 was used to extract information and classifying text. It's beneficial to analyze the

© 2019 IJSRET

relation between human influenced events and social media. Several fuzzy rules are designed and de-fuzzification methods were combined to get desired results. Suggested technique was compared to popular search method as per rate and quantity correctness. Results show that proposed technique is suitable for classification of twitter messages. The experimental uses the twitter review using social media.

In [20] proposed a technique which uses the connection between lexical things and labels before finishing Latent Dirichlet Allocation (LDA) theme display. They modified parameters of SVM (Support Vector Machine) to locate optimized values by K-crease cross approval. It's an awesome test that comprehending high-measurement and content scarcity issues in short content arrangement. Also, utilizing piece SVM as classifier, we effectively arrange named short Chinese content reports. Contrasting and other two regular techniques k-Nearest Neighbor and Decision Tree of short content arrangement, the exploratory outcomes demonstrate that our strategy outflanks them on order exactness, accuracy, review and F-measure.

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 has 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 detects the spam but spammer profile as well.

REFERENCES

[1]. 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.

[2]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana. "Relevance Feature Discovery for Text Mining". IEEE transaction knowledge and Data ENGINEERING, VOL. 27, NO. 6, JUNE.

[3]. Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song. "Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues".IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 12, DECEMBER 2013.

[4]. Wang, J. Wang, et al., "Labelled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo." In 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 428-432. IEEE, 2017.

[5]. Mingyong Liu1 and Jiangang Yang. "An improvement of TFIDF weighting in text categorization". 2012 International Conference on Computer Technology and Science (ICCTS 2012).

[6]. 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

[7]. 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.

[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]. 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.

[14]. 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.

© 2019 IJSRET

[15]. 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.

[16]. 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.

[17]. 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.

[18]. 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.

[19]. K.. Y. Wu, M. Zhou, et al., "A fuzzy logic-based text classification method for social media data," Systems, Man, and Cybernetics (SMC), IEEE International Conference on, vol.13,no.3 pp:23-32. 2017.

[20]. X. Wang, J. Wang, et al., "Labelled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo." In 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 428-432. IEEE, 2017.

© 2019 IJSRET