Journal of Advances and Scholarly Researches in Allied Education
Vol. 19, Issue No. 1, January-2022, ISSN 2230-7540

A Study the Review of Text Classification with its Applications and Processes

Ankur Pandey[1]*, Dr. Anoop Kumar Chaturvedi[2]

[1] Research Scholar, LNCT University

[2] Guide, LNCT University

Abstract - Text mining is a variant of data mining that attempts in finding an interesting pattern from larger datasets. The text databases have been growing rapidly owing to the increase in amount of information that is available in the form of electronic publications, from e-mail and World Wide Web. Technically, text mining is one of the automated methods that exploit the large knowledge available from text document. Natural Language Processing (NLP) has several applications in web search, retrieval of information, ranking and classification of documents where text classification is an important task. Text classification intends to provide high quality textual representation accessed from the digital forms of document available online and build high quality classifiers. Current research explores the Machine learning classifiers for text classification.

Keywords - Text classification, Machine learning, text documents, Document classification

----------X----------

INTRODUCTION

Text documents such as the contents of an article or a free-text box in an electronic medical record. Require much processing before they can be queried for IR purposes. Numerous techniques have been developed to achieve various IE tasks. For example, in biomedical publications, IE can be utilized to label articles with proteins or procedures they discuss; news articles may be marked with major topics within them; and the CIA may automatically flag private text messages as containing potential terrorist plots. The IE technique employed depends on the specific task and queries of interest. To label articles with protein names is a fairly simple task that requires a quick keyword search of the free-text for the given terms (protein names) [Patra et al.]. However, if one wants to allow for the query of proteins by their synonyms, additional processes are required; tools based on semantic frameworks are useful for this, which will be discussed at a later section.

Marking articles with their topical content requires a slightly more complex method called bag of words. In this method, only the word frequencies in the text matter, and their order is ignored. If a word or collection of words appears more often in the article than what would be expected, then it can be safely labeled as dealing with a topic that can be represented with those high-frequency words [Joulin Armand 2016]. Of course, some extra work (machine learning) is required to find which words are a good representation of a topic. Machine learning applied to text data will be discussed in detail later. The CIA's task is more complicated than finding just the topic of a text.

One can send a text that contains the topic of terrorism without having any mention of intent in planning an attack. This is when more complex NLP methods become necessary, usually with the intent of natural language understanding; that is, to have the algorithm understand the meaning of the message within the text. Rather than just having some statistical sense of topical correlation [Sarkar 2015]. Some specific tasks of this nature are named-entity recognition (annotating parts of text with classes of entities), conference resolution (finding all references to the same entity), part of speech (such as noun, verb, adjective. etc.), and relationship extraction (extraction of word-word interactions).

Features of Documents

Text mining difficulties can be solved with a variety of methods, all of which include retrieving only the information that is most useful to the user. Some approaches are described below that are grounded in information retrieval techniques.

1) Term based method: A term in a text is a word with a defined meaning. With the advantages of efficient computational performance & well theories for term weighting, the term-based method [Aggarwal, 2012] scrutinizes documents on the basis of terms. Over the past few decades, the Information Retrieval & ML Community has created these strategies. Polysemy & synonymy are two problems with this approach. Both polysemy and synonymy

include words with many meanings. Many of the detected terms have permissible meanings that are vague and don't directly address the questions people have. Many term-based approaches to the posed problem can be found within the framework of information retrieval.

2) Phrase Based Method: The phrase provides more semantic information & uncertainty [Vijayarani, 2015]. Here, we use a phrase-by-phrase approach to estimating the quality of the document, rather than a term-by-term method. There are a few things holding the performance down:

1) Due to secondary analytical properties to terms
2) Less occurrence
3) Massive duplicate & noisy phrases

3) Concept Based Method: Sentence & document-level estimates are used for terms in this technique. Many Text Mining methods center around dissecting and analyzing certain words or phrases. Analytical analysis perfectly sums up the significance of the spoken word apart from any written record. Although two terms may appear interchangeably in a given document with equal frequency, one may provide a more contextually relevant meaning [Vijayarani 2015]. To learn the meaning of texts, a new concept-based mining method is presented. The three parts of this model are as follows. The first part is an analysis of the sentences' semantic order. In the second, a conceptual ontological graph (COG) is assessed for its ability to characterize semantic structures, & the third, the most salient concepts are extracted for use in the construction of feature vectors in accordance with the canonical vector space model. A key feature of this approach is its capacity to identify & isolate the essential terms from among the many that collectively characterize a meaningful phrase. The answer to this question often rests with NLP techniques. To improve the representation and get rid of ambiguity & noise, the concepts in the queries are compelled to have a certain selection applied to them.

4) Pattern Taxonomy Method: Documents are graded using patterns in pattern taxonomy. Application of the is-a connection allows taxonomists to build patterns [Jindal 2015]. Data mining has been reviewing pattern mining for a long time. Association rule mining, frequent item set mining, sequential & closed pattern mining are only some of the data mining techniques that may be used to identify patterns. Because of the lack of support for some essential long patterns with high selectivity, the application of identified knowledge in the field of text mining is both crucial & inefficient. Misconceptions about patterns, in which the assumption that all short patterns are useful leads to subpar results, are common. To combat the low-frequency & misconstruction issues plaguing text mining, a productive pattern finding technique has been suggested. Two mechanisms, pattern deployment & pattern evolution, are used in the pattern-related method. This method improves upon the previously found patterns. The pattern-based model outperforms all other methods that are based solely on data mining.

OVERVIEW OF TEXT CLASSIFICATION

Natural Language Processing (NLP) has several applications in web search, retrieval of information, ranking and classification of documents where text classification is ansignificant task. Text classification intends to provide high quality textual representation accessed from the digital forms of document available online and build high quality classifiers [Xu, JS 2007]. The phases involved in text classification are database collection, preprocessing of the data, reduction in dimensionality of the dataset and implementation of the classifier. Current research explores the Machine learning classifiers for text classification. Some of the effective classifiers are Naive Bayesian, support vector machine, k-nearest neighbor classification, neural network and so on. Neural network based models are widely used and outperforms other models but they take more time for training, thereby limiting their usage on large datasets. A neural network obtains state of the art performance when they are trained with the suitable ideal features and scales to a larger corpus [OzgurLevent 2010].

The process of text classification begins with identifying ideal features and selection of machine learning classifiers. The conventional text classification methods are based on the word combinations like n-grams which normally perform well. Several researchers have identified convolutional networks (the ConvNets) that are found to be useful in the extraction of information from raw signals that range from the applications of computer vision to speech recognition.

TEXT CLASSIFICATION PROCESS

The process of assigning classes to the input document where every document belongs to one or more classes based on the content is called text classification. The advent of Internet, has paved way for heavy flow of information resulting in promoting the growth of automated text classification.

Sufficient computing power is provided through the construction of computer hardware to use automated text classification for practical applications. Text classification helps in handling spam e-mails, classifying large text collections into various categories based on topics and knowledge management and it also helps Internet search engines.

Artificial Intelligence consist of (a) reasoning in spite of uncertainty; (b) task's complexity; (c) heuristics which are required for approximating human judgment. The term 'text' when not properly defined

causes dispute when the results of AI is compared to human perception.

In a text classification process, a classification scheme needs to be proposed. An individual label predicted from a set of documents is termed as class. Designing of a classification scheme does not involve strict protocol/scheme but it is implemented with a strict perseverance so that it receives all ideal data from training.

TEXT CLASSIFICATION APPLICATIONS

Text classification finds its use in various domains such as:

News filtering and Organization: In recent times, extensive news articles are generated by electronic news services like news web portals. However organizing large volume of news articles manually is a nightmare. This paved way for automatic categorization of news articles in various web portals called as text filtering.

Document Organization and Retrieval: Grouping or organization of the documents including digital libraries, scientific literature, web collections or social feeds may be done through a range of supervised method. Sequential organization or hierarchical organization of documents finds its application particularly in browsing and retrieval.

Opinion Mining: Reviews or opinions of the customer can be short documents that can help in determining useful information [Bolaj 2016].

Email Classification and Spam Filtering: Spam or e-mail filtering is nothing but determining whether the mail is a genuine one or a junk/spam mail.

DOCUMENT CLASSIFICATION

Document Classification (DC) stands as the process of analyzing a group of documents and labeling each one of them with an appropriate category as per its relevance towards one of the pre-defined collection of categories (Joorabchi& Mahdi 2011). The DC is rife with potential for several modern document-centric applications, like Document Summarization, essay scoring, organizing documents for query-based information dissemination, email management and topic-specific search engines. DC is usually a 2-step process. First, the text contained in a document is analyzed and a compact set of features which characterize the document are generated. Next, centered on their extracted features, the documents are classified to their respective categories by employing a suitable ML technique. DC undergoes a series of tasks to reach its aim. The tasks involved in the DC process for supervised and un-supervised techniques are depicted in Figure 1.5

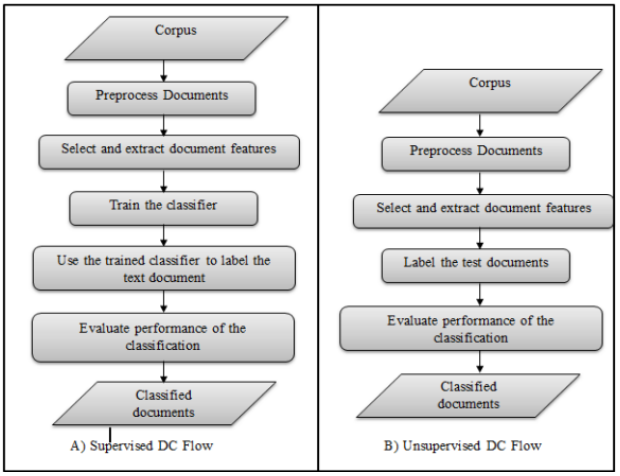

Figure 1: General flow for DC

Presently, DC becomes more imperative. The internet has contributed significant augmentation of un-structured data produced and consumed (AnkitBasarkar 2017). Therefore, there is an acute requirement for content-centered DC in order that these documents are competently located by the consumers who desire to consume it. Search engines were exactly developed for this job. Search engines say Yahoo, HotBot, et cetera, in their former days, utilized to work by forming indices and ascertain the data requested by the user. Nevertheless, it was not eminently un-common that search engines sometimes might return a list of documents with bad correlation. This has brought about development and research of intelligent agents which makes utilization of ML in classifying documents.

If a supervised approach is followed, the classifiers have to be trained by a group of training documents. Thus the corpus is bifurcated into i) training and ii) testing sets. This step is not needed in unsupervised classification.

The individual steps are briefly described below:

- **Pre-process Documents:** The raw documents must be the pre-processed to convert them into an array of non-trivial tokens. Pre-processing comprises the following steps:
  a. **Stop word removal:** Trivial words, like ‘a’, ‘am’, ‘above’, ‘across’, ‘alongside’ ‘but’ etc, which do not play any significant role in classifying a document are removed.
  b. **Stemming:** Inflected words are transmuted to their root forms. These are termed as tokens. The goal is to evade treatment of the same words in different forms. For example, both ‘celebration’ and ‘celebrating’ are converted to the same base form ‘celebrate’, thereby dealing with a single word. Hence, Stemming lessens the size

of the feature sets by utilizing such tokens.

- **Feature Selection and Extraction:** This step is required to lessen the large dimensionality of a feature space that is typical of text documents. It aids to eliminate the unimportant tokens from documents. Dimension reduction is accomplished by two methods:
    a. **Feature selection:** In this method, a subset is selected from the original features. Since the number of chosen features is less than the original features, it reduces a great dimension of feature space which helps to restrain cost and also time complexity.
    b. **Feature Extraction:** In feature extraction, new promising features are added to the original. As a resultant, a higher dimension projected feature space is received. Now a group of features is extracted from this feature space.
    c. **Train the Classifier:** The labeled features sets are applied to a classifier which accordingly regulates its internal decision parameters so as to be able to discern the category of any new documents.
    d. **Label Test Documents:** The test documents are inputted to a trained classifier. These documents are labeled with their relevant categories as per the classifier's acquired discerning ability.
    e. **Evaluate the Classifier:** Now, the predicted categories of test documents are matched with their originally labeled categories. Centered on the number of correct & also incorrect matches, the classifier's performance is evaluated by calculating the various performance indices explained in the above subsection.

LITERATURE REVIEW

Guang Wu et al. (2022) Many sports training researchers in China are now looking to artificial intelligence tools to better understand the many facets of athlete preparation. However, in practice, these approaches frequently use varying basic training principles, which limits the generalization capability of AI networks. Using a neural network model developed by AI., this research investigates the multifaceted nature of effective sports training practices. This research offers an AI sports training node prediction approach that uses an upgraded version of the dropout optimization algorithm in conjunction with LSTM to circumvent the need for elaborate sports training models. The maximum node static estimation of AI sports training is achieved through the use of operational and maintenance logs and experimental data on the core capacity of AI-based sports training. According to the findings, the node prediction model may be created with the help of the procedure outlined in this work. The model has been shown to have great prediction accuracy through experimental comparison & analysis. Advantages in long-term prediction of 2000 data can be attributed to LSTM's state memory function. Prediction findings had an absolute error of less than 3.4% on average and less than 5.2% at their most extreme. The network model used by the AI in this paper offers strong generalization capabilities. The model developed in this research effectively mitigates the issue of overfitting& yields more accurate prediction outcomes in sports training for distinct groups than competing models. Consequently, traditional stadiums & gymnasiums should take a more proactive approach to introducing artificial intelligence technology in order to realize advancements and innovations in technology application, service innovation, management efficiency, or function integration.

Muhammad Imran et al. (2022) One of the key problems for researchers and network managers is anomaly detection in network traffic. A quick and reliable network IDS is needed when anomalies in network traffic indicate a network intrusion. The research community is becoming more interested in intrusion detection systems that utilize artificial intelligence (AI) techniques as these techniques have advanced recently. With the use of artificial neural networks (ANNs) that have been improved using the cuckoo search algorithm, this research suggests a fresh approach to anomaly identification. The NSL-KDD dataset has been used with a 70:30 ratio for simulation purposes, where 70% of the data is used for training & remaining 30% is utilized for testing. The proposed model is then assessed in terms of accuracy, root-mean-square error, mean absolute error, and mean square error. The suggested research's findings are contrasted with accepted techniques found in the literature, such as fuzzy clustering artificial neural network (FC-ANN), intrusion detection using a bee colony made of artificial bees, neural network IDS, and feature selection.The outcomes unequivocally demonstrate that the suggested strategy outperforms the mentioned conventional methods.

Arati Paul et al. (2022) An efficient preprocessing step for hyperspectral image (HSI) analysis is dimensionality reduction (DR). In the current study, an unsupervised DR approach based on PSO is

proposed, where informative bands are chosen using both spatial gradient information & spectral divergence. In generally, low signal-to-noise ratio (SNR) has a detrimental impact on HSI.Hence, in the proposed method, a noise filter is applied to minimize the effect of noise in band selection. Clustering is introduced to reduce spatial redundancy and extract distinct patterns from the data. This enables improvement in the computation performance of each iteration in PSO. On two common datasets, the suggested strategy is used, & performance is assessed utilizing overall classification accuracy.Lastly, outcomes are associated with other recent state-of-the-art methods where the proposed method performed reasonably better than other tested methods in terms of consistency and classification accuracy.

Basanti Pal Nandi et al. (2022) The study of sentiment, sometimes known as opinion mining, covers a wide expanse of academic territory. Today, when trying to form an opinion on a topic, we take into account the vast amounts of both structured & unstructured data that can be found on the web. Big data refers to the large amounts of data that need to be processed, which necessitates the development of new tools. This research takes into account the need for efficient processing time while performing sentiment analysis on such a large data set. This approach (FFT-TIFS) speeds up sentiment classification by using fast Fourier transform on a temporal intuitionistic fuzzy set derived from the text. The signal is transformed by Fourier analysis from the time domain to the frequency domain. It is the first time a frequency domain technique based on a temporal intuitionistic fuzzy set has been applied to sentiment analysis. This method can classify binary sentiment at the document level and the sentence level, making it helpful for the short texts that are common on Twitter. On average, it achieves 80% accuracy when tested with the aclImdb, Polarity, MR, Sentiment 140, and CR datasets. When compared to the sequential fuzzy C-means (FCM) approach, the suggested method is 17 times faster in processing, and when compared to the distributed FCM method in the literature, it is at least 7 times faster. This study presents a novel method for conducting text sentiment analysis that is both fast and flexible enough to be used to texts of varying sizes.

Deepak KumarJain et al. (2022) Recent developments in networking & information technology have always occurred as a natural phenomena. With people's data production skyrocketing, Big Data Analytics (BDA) has become increasingly popular. Brain-computer interface problems can be mitigated with the use of cognitive computing, an AI-based system. Meanwhile, Sentiment Analysis (SA) is used to interpret these linguistically based tweets, extract features, compute subjectivity, and analyze the sentimental contents embedded in these tweets. Companies that apply SA to large data sets are able to glean useful commercial insights from hitherto untapped sources of textual information. This research introduces a novel cognitive computing framework alongside a big data analysis tool tailored to SA. Procedures including pre-processing, feature extraction, feature selection, and classification are all a part of the suggested model. Hadoop'sMapReduce program is used to manage massive amounts of data. The suggested model first goes through pre-processing to filter out irrelevant terms. Then, the set of feature vectors is extracted using Term Frequency-Inverse Document Frequency (TF-IDF) as a feature extraction technique. In addition, the FS process is optimized with the help of a Binary Brain Storm Optimization (BBSO) method, leading to better classification results. Also, Fuzzy Cognitive Maps (FCMs) are used to categorize how often positive or negative emotions are expressed. This improved performance on the benchmark dataset is guaranteed by a thorough examination of experimental findings using the given BBSO-FCM model. The obtained experimental values demonstrate the proposed BBSO-FCM model's superior classification performance across multiple metrics.

V. Vaissnave et al. (2022) Due to the plethora of available data nowadays, text segmentation is a powerful tool for finding and extracting relevant information from large volumes of documents. Text categorization is the process of dividing a document into smaller, semantically related pieces of text and labeling them. Given the abundance of semantic information in legal texts, text segmentation plays a critical role in the retrieval of data from court pleadings and other legal documents. Furthermore, in order to construct an effective system, supervised classification requires a massive amount of training data. It would be very expensive to collect and manually categorize such a large number of data. The doors to automation have been opened by recent developments in information technology, especially artificial intelligence. In this paper, we offer a model that employs DL techniques to disentangle a verdict's text into its component parts: the problem, the evidence, the analysis, the conclusion, the votes, and the winners and losers. After running trials to test the efficacy of the suggested model, we found that the proposed bidirectional long short-term memory (Bi-LSTM) categorization technique achieved 97% accuracy

and obtained top-notch performance in its target task. We've developed an automated method to classify historical judgments and extract useful legal reference information from them to save you time and money.

Pallavi Grover et al. (2020) The input features and the method of extracting and selecting them have a significant impact on the learning of a text categorization system. The primary motivation for engaging in feature selection is to reduce the dimensionality of the problem at hand, which in turn makes classification easier. Feature selection is critically important in many different problem domains, and text categorization is just one of them. The curse of dimensionality is notoriously problematic for text classification. This leads to the development of a bad classifier based on a feature space that may contain redundant or unnecessary information. Therefore, selection is a crucial step in the development of a smart classifier feature. In order to accomplish its four goals, this paper will first apply a popular score to the task of translating words into vectors. Second, it plans to use a nature-inspired approach to maximize the potential of the textual feature space. Finally, it seeks to evaluate the strengths and weaknesses of three widely-used classifiers for text classification: SVM, Naive Bayes, and k-Nearest Neighbors. As a final goal, it seeks to contrast various measures. To gain insight into the repercussions of optimizing feature space with a nature-inspired algorithm, in addition to improving accuracy. Classification accuracies for SVM, Naive Bayes, & k-Nearest Neighbors all reached 95.07% when the standard text classification dataset Reuters-21578 was employed. The performance indicators included not just accuracy but also precision, recall, & F-measure. When looking at the positive outcomes obtained using the ABC algorithm, it appears that this technique holds promise for other uses of text classification.

CONCLUSION

Nowadays the incremental growth of text documents on the web emphasizes the importance of text document classification. Text document management has become an essential methodology due to the rapid growth in document digitization like World Wide Web. Machine learning applied to text data. Cataloguing text documents into one or more predefined classes or categories is called text categorization. Based on the feature set, various categorization stake place and the relationship of a given document to a category is assessed. Text classification is a fundamental task with the overarching goal of categorizing a batch of documents into a predetermined number of classes.

REFERENCES

1. Abualigah, L. M., Khader, A. T., Al-Betar, M. A., &Alomari, O. A. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Systems with Applications, 84, 24-36.
2. Abualigah, L. M., Khader, A. T., Hanandeh, E. S., &Gandomi, A. H. (2017). A novel hybridization strategy for krill herd algorithm applied to clustering techniques. Applied Soft Computing, 60, 423-435.
3. Adrian Bilski 2011, 'A review of artificial intelligence algorithms in document classification', International Journal of Electronics and Telecommunications, vol. 57, no. 3, pp. 263-270.
4. Aggarwal, CC &Zhai, C 2012, 'A survey of text classification algorithms in mining text data', Springer, Boston, MA, pp. 163-222.
5. Aghdam, M. H., &Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research, 5.
6. Agnihotri, D., Verma, K., &Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. Expert Systems with Applications, 81, 268-281.
7. Ahmad BasheerHassanat, Mohammad Ali Abbadi, GhadaAwadAltarawneh& Ahmad Ali Alhasanat 2014, 'Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach', arXiv preprint arXiv:1409.0919.
8. Ahmed H Aliwy&Esraa H Abdul Ameer 2017, 'Comparative study of five text classification algorithms with their improvements', International Journal of Applied Engineering Research, vol. 12, no. 14, pp. 4309-4319.
9. 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.
10. Albashish, D., Hammouri, A. I., Braik, M., Atwan, J., &Sahran, S. (2021). Binary biogeography-based optimization based SVM-RFE for feature selection. Applied Soft Computing, 101, 107026.
11. AliakseiSeveryn& Alessandro Moschitti 2015, 'Learning to rank short text pairs with convolutional deep neural networks', in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 373-382.
12. AlperKursatUysal&SerkanGunal 2012, 'A novel probabilistic feature selection method

for text classification', Knowledge-Based Systems, vol. 36, pp. 226-235.
13. AlperKursatUysal, &SerkanGunal 2014, 'The impact of preprocessing on text classification', Information Processing & Management, vol. 50, no. 1, pp. 104-112.
14. Amit Rathore 2015, 'Review of structure and unstructure based web document classification', International Journal of Computer Security & Source Code Analysis (IJCSSCA), vol. 3, no. 2.
15. Andrea Esuli, TizianoFagni&FabrizioSebastiani 2008, Boosting Multi-Label Hierarchical Text Categorization, Springer, pp. 1-27.
16. AnkitBasarkar 2017, Document Classification using Machine Learning, Master's Projects, P. 531.
17. AnshuBharadwaj, ShashiDahiya&Rajni Jain 2012, 'Discretization based support vector machine (D-SVM) for classification of agricultural datasets', International Journal of Computer Applications, vol. 40, no. 1, pp. 8-12.
18. Antoniou G. (1997). Nonmonotonic Reasoning, MIT Press.
19. AnuradhaPurohit, DeepikaAtre, PayalJaswani&PriyanshiAsawara 2015, 'Text classification in data mining', International Journal of Scientific and Research Publications, vol. 5, no. 6, pp. 1-7.
20. Arar, Ö. F., &Ayan, K. (2017). A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Applied Soft Computing, 59, 197-209.
21. ArashJoorabchi&Abdulhussain E Mahdi 2011, 'An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata', Journal of Information Science, vol. 37, no. 5, pp. 499-514.

Corresponding Author

Ankur Pandey*

Research Scholar, LNCT University