International Journal of Research in Advent Technology, Vol.7, No.6S, June 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Survey on Classification and Summarization of Documents

Naresh E, Vijaya Kumar B P, Pruthvi V S, Anusha K, Akshatha V

Abstract—Text classification and summarization has become one of the important concepts in analyzing and understanding various text documents related different fields. It is used in various different fields like spam email detection, twitter sentiment analysis, summarizing textual documents, etc. There have been few works in dataset formation and analysis for different complex documents like research papers, survey papers etc. This survey concentrates on reviewing and comparing various approaches used for analysis of such complex documents.

Index Terms— Text classification; text summarization; pre processing; machine learning; neural networks; domain specific documents; research papers.

I. INTRODUCTION

With growth of digitalization and easy availability of electronic resources, a lot of textual data is available all over the world. Documents are being made digitally, with the increase in complex documents containing lot of information in them and due to exponentially increasing volume of data, understanding them has in-turn increased the popularity of text summarization and classification techniques and there is need for efficient tools to handle such data [1].

Text classification is one of the tasks in supervised machine learning where available text data is classified into different categories based on the content of the data [2] and text summarization is the process of shortening a text document, in order to create a summary of the major points of the original document. There are two methods of text summarization, abstractive and extractive[3]. There are various applications of these two techniques which include email classification, social media monitoring. Natural Language Processing(NLP) is commonly used computational technique for the understanding of text documents[4]. The above two techniques can be applied using NLP. In general, all text documents go through preprocessing before going through any kind of analysis. With increase in availability of documents, noise or generation of unwanted data also increases, hence Pre-processing of documents is very important.

Preprocessing transforms the input text data before any further text mining analysis, into a text with meaning, including the needed linguistic features. It discards the unnecessary data from the text. Bad preprocessing deteriorates the further document processing effects[5]. As we are speaking about huge and complex text documents, thorough pre processing of such documents is necessary before further processing.

There have been different works on showing the importance of pre processing techniques and are coming up with more efficient techniques. Few such techniques include stop word removal[7], lemmatizing[8], tokenization[9][10], Parts of Speech tagging, stemming[11][12] and many more.

Some of these techniques are the most essential techniques, without which the analysis cannot take shape. Once the document is processed and ready for further analysis, two important tasks need to be performed on these documents, classification and summarization. Natural Language Processing provides many prominent algorithms for the above tasks. A number of statistical and machine learning classification methods have been used for text classification, which includes Naive Bayes Algorithm[14], this algorithm works well for email spam classification and categorizing[15], another important algorithm is the Decision Tree, it has been used for different applications, but this algorithm has mainly proved to be very efficient in Arabic and Urdu text classification[16][17][18].

SVM is another strong algorithm and different researches in the field of SVM have aimed at improving the efficiency of the algorithm and towards making it better[19][20][21], KNN [22][23] is another simple but widely used algorithm and is very efficient in some cases. All these algorithms have their own advantages for different classification problems.

The newly emerging techniques in the field of text classification as well as summarization include Neural Networks such as CNN and RNN, these can prove to be of great importance for the analysis of complex text documents where a huge amount of text is involved and highly computational algorithms like CNNs and RNNs can be used for effective results[24][25][26].

Document summarization is also an important step for analyzing and easily understanding the content of the documents. Different technique like Extractive summarization[27], Latent Semantic Analysis(LSA) [28],

Manuscript revised May 13, 2019 and published on June 5, 2019
Naresh E ,Assistant Professor, Department of ISE, M S Ramaiah Institute of Technology Bangalore, India
Vijaya Kumar B P, Professor and Head, Department of ISE, M S Ramaiah Institute of Technology Bangalore, India
Pruthvi V S, Department of ISE, M S Ramaiah Institute of Technology Bangalore, India
Anusha K, Department of ISE, M S Ramaiah Institute of Technology Bangalore, India
Akshatha V, Department of ISE, M S Ramaiah Institute of Technology Bangalore, India

doi: 10.32622/ijrat.76S201906

Knowledge based and automatic summarization[29], genismetc have been used for this purpose in different works.

This survey works on analysis of research documents and domain specific documents has been concentrated on, as applying the general algorithms to such documents do not give expected results. There has been works where analysis of biomedical documents have been done using different supervised, semi-supervised and self supervised techniques for relation extraction[30]. Another significant work has been found where the research paper topic or subject has been classified based on interrelationship between them, like common author, references, citations etc[31].

There has been exclusive work in scientific document classification where the system accepts the entire document in pdf format as input and does classification of the document[32]. This system also differentiates between different parts of the document and analyses them. Finally, it produces an ontology based on the analysis which can be used to draw inferences.

Another important work which concentrates on the domain specific classification of documents and provides a simple algorithm which takes into consideration the domain specific nature as well as the use of certain local scientific terms while analyzing the documents[33].

There has also been development of a system called papits which shares research information related to a particular domain specific topic or paper. This research information includes pdf files of research documents. This system also classifies documents into research topics[34].

The main goal of this paper is to compare the different techniques used for such domain specific text categorization and summarization and to compare the algorithms which can be used for this purpose.

II. PRE PROCESSING TECHNIQUES

A. Removal of Stopwords

Stopwords in a language are those words that are less significant than the other tokens. They are the commonly used words like prepositions and conjunctions. Stopwords elimination is another important task of the preprocessing technique in NLP and text mining applications. This process impacts on the further processing of the text documents differently based on the application implementing it. Say, in case of text classification this process reduces the number of dimensions in terms of space but in the case of Machine Translation, this could have a negative impact on the accuracy. When it comes to the case of text summarization, the stop words can be removed based on the requirements. Even though there is no list of universal stop words yet, according to the convention, the stop words do not have any important meaning, and they don't have any significant effect on the result of a text mining technique. So, it is better for the text mining techniques' performance and speed if they are removed. Even so, there has been no prominent examination on the real effects of the stopwordselimination[6] and there are a few research that imply that the stopwords can have a positive impact on the text mining techniques as in [7].

B. Lemmatizing

The fundamental modules of Natural Language Processing are lemmatizer and stemmer. Lemmatizer applies linguistic rules to remove the affixes in the inflected words. It later returns the lemma, which is the dictionary form(base form) of the word. As the split lemma is a meaningful valid root, a lemmatizer needs more linguistic knowledge than a stemmer. In the linguistics, the objective of a lemmatizer is to make groups of inflected word forms for each word, and consider each group to be a common term. A lemmatizer checks for long suffix/prefix, and then, the input inflectional word is stripped off of the suffix/prefix from it. Next, to get the proper lemma, rules are applied on the stem that is obtained. The root-directory is searched for the obtained lemma. In case it is not present, it is automatically added into it, and if present, it is displayed. The performance of the lemmatizer is thus increased due to the increase in the number of entries in the root dictionary.

C. Tokenization

Tokenizing is the basic text processing technique. Tokenization breaks down the text sequence into into words, keywords, phrases, symbols and tokens. Tokens can be individual words, phrases or even sentences. Tokenizer identifies and separates the tokens in the text such that each word and each and every punctuation mark is considered a different token. It considers acronyms, abbreviations, numbers with decimal, date in numerical format to separate dot or comma or slash from previous or the following elements [9]. To function well, two separate dictionaries are used for acronyms and abbreviations and specific rules are used for detecting dates and numbers. A regular expression tokenizer uses regular expression to split substrings[10]. Next, the tokenizer matches either the tokens or the separators between them. It considers sequence of blank lines as a delimiter.

D. POS Tagging

The Parts of Speech tagging technique annotates the terms in the text corresponding to a particular parts of speech according to the interpretation and circumstances, like the relationship with the neighbouring and the associated words in the sentence, phrase, statement sequence and paragraph. POS assigns each word with the tags that represents syntactic role. This can be achieved by the use of features such as the multi words bigrams and trigrams.

E. Chunking

Chunking is used for the removal of unused words in the sentences. It groups together the similar words in the chunks. And the remaining words are eliminated. Segments of the sentences are assigned syntactic roles like verb or noun phrases which is nothing but the tagging of the text by the chunking. A unique tag is assigned by chunking and are called begin-chunk or inside-chunk tags.

F. Stemming

Yet another significant preprocessing technique in text mining is stemming. It is the technique used to reduce the derived and inflected forms of the words into their basic forms also called the stem or the root words[11]. Stemming replaces a number of words that occur in the text document and those which are semantically close by their root words. This reduces the terms vector's dimension, which increases

doi: 10.32622/ijrat.76S201906

the resultant document's quality. Stemming can be categorized into three methods: truncating, statistical and mixed [12]. The truncating method finds the stem by removing affixes of the words.

The second group of stemmer is based on the statistical techniques and is called the statistical method. After implementing some of the statistical procedure, the affixes are removed. N-Gram Stemmer is one such stemmer which is language independent. Mixed method is the third category that contains the Inflectional and Derivational Methods, these are based on word variants and also based on parts of speech. One of the most popular approaches is the Porter Stemmer[13].

III. APPROACHES FOR TEXT CLASSIFICATION

A. Naive Bayes

Naive Bayes is a probability based classifier which is based on the Bayes' Theorem. In this classifier the feature value is not dependent on the value of any other feature that is related to the document. Hence, it takes the different mutually independent features of a document as input and classifies it into one of the different classes available[14].

This algorithm has been used for different classification techniques. One important application includes spam email classification [15].

B. Decision Tree

This is a kind of classifier which divides data in the hierarchical way that has the tree structure. In a decision tree, the features or the attributes are represented as nodes, the decision or the rules are represented by each link also called the branch, and the outcomes(categorical or the continuous values) are represented by the leaf nodes. So, for classification of the text data, each node of the decision tree represents the features of the text document and the final outcomes signifies which class it belongs to. Therefore we traverse from the root of the decision tree, till we reach one of the leaf nodes which classifies the document into a category. It has been used for different purposes like text classification for inappropriate web content blocking[16], image classification[17], Urdu text classification[18] etc.

C. SVM

The classification in Support Vector Machine is based on the hyperplanes. Here the different classes are separated by the planes and each partition represents a class. In SVM algorithm, for a class with n number of classes, the data items in the classification data are plot in the n-dimensional space as points, so the value of a coordinate corresponds to the value of a particular feature. Next, the classification is done by determining the hyperplane that differentiates two classes distinctly[19]. For applications like question classification[20], web document classification[21] and many other text classification applications the SVM is used.

D. KNN

KNN is a supervised learning algorithm which is non-parametric and a lazy learning algorithm. The KNN algorithm predicts and classifies a new sample point into a class using the database with the separated data points of various classes. This algorithm classifies based on the feature similarity which classifies a given data point. When classifying a document it computes the distance between documents and then the word features are chosen based on that. Each document is categorized according to its most K-nearest neighbours then it will be assigned to the category with that neighbours[22]. It can be used for various text classification application in different fields. It has been proved to be better algorithm for Persian text classification in one of the works[23].

E. Neural Networks

There are multiple approaches under neural network. In this study we are going to focus on the Convolution Neural Networks(CNN). CNN is a feed forward network. Input for CNN model must be encoded. The popular encoding techniques are one hot encoding, word2vec, Latent semantic analysis. In this one hot encoding is the basic where a particular word corresponds to a dimension in the vector. But the drawback of the one hot encoding is as the unique words increases, the vector dimension also increases. This in turn reduces the performance when there are more unique words. One hot encoding does not take semantics into consideration, therefore documents with similar meaning will have different vectors with high cosine values[24].

An alternative for the one hot encoding can be word2vec [25]. In word2vec feature vector of words are formed. This feature vector help in establishing semantic relationship. Hence similar meaning words ave a high cosine values. There are two submodels for the word2vec.

Submodels are CBOW and Skip Gram. Both the models are used to train the vectors. CNN basically consists of a number of convolution layers and pooling layers.

Convolution layer- this layer's parameters are the Kernels. Kernels along with the input produce a dot product to produce a matrix which is the feature map. The input layer of CNN consists of number of text(n), fixed text length(s) and dimension of the word vector(k).

Pooling layer- the main purpose of the pooling layer is to avoid overfitting. This is done by taking input from previous and dividing it into non overlapping regions. Maximum value of the feature map would extract the most useful feature. In CNN non-linearity is the important. In order to introduce non-linearity several functions like sigmoid, tanh and most popular ReLu are used.

F. RNN

Recurrent Neural Networks (RNN) are robust and powerful neural networks. As they have internal memory, they are very considered as strong algorithms. They precisely predict what is coming next as they can remember the significant facts regarding the input received and this is because of the existence of the internal memory. And for this reason the time series, speech, text, financial data, video, weather and many more sequential data use this algorithm.

In the NLP systems, the words are conventionally treated as distinct atomic symbols. NLP models can predict very little information regarding the relationship between different words. These models can have major drawbacks of

doi: 10.32622/ijrat.76S201906

ignoring the semantics and the order of the words. RNN helps in reducing loss of details and improving input dependencies which is considered to be a very powerful tool in the predictions for text [26].

IV. APPROACHES FOR TEXT SUMMARIZATION

A. Extractive Summarization

This technique is based on the selection of paragraphs, important sentences and so on to produce the summary of the original documents precisely. And according to the linguistics and statistical features, the sentence implication is determined. Word and sentence level features are the few features that are considered for the extractive summarization. Few supervised and unsupervised approaches are present for extractive summarization [27].

B. Genism

It is an unsupervised algorithm. A corpus of text which is to be summarized is the only requirement. The semantic structure of the document is automatically discovered by inspecting the statistical co occurrence pattern with the text corpus. Genism is an open source, free algorithm to extract semantic summarization of text. An extractive summarization[27] is performed in this algorithm. It can process large text corpus due it memory independent feature. Due to its effective implementation of the word2vec[25] and LSA[28] it is one of the popular choice for text summarization.

C. Latent Semantic Analysis (LSA)

The LSA method applies the statistical computations to extract and represent the contextual-usage meaning of the words. The LSA assumes the distributional hypothesis, according to which, the semantically closer words occur in the similar text pieces. The matrix represents source text document in which the columns represents the sentences and the rows represent the unique words. And such a matrix is constructed from huge pieces of text and a linear algebra theorem called Singular Value Decomposition (SVD). It preserves the similarity structure among the columns while reducing the rows. Various words are compared by finding the dot product between the normalization of the two vectors formed by any two rows in the matrix. If the values are close to 1, then the words are considered to be very similar and if the values are close to 0 are considered to be dissimilar words [28].

D. Knowledge Based And Automatic Summarization

Here the knowledge or the data is fed into the system and the system automatically extracts information from the knowledge representation using intelligent algorithms and presents a summary. Bayesian networks have been used in one of the works for knowledge based summarization [29].

V. EXISTING ANALYSIS ON COMPLEX DOMAIN SPECIFIC DOCUMENTS

In this survey, we present different existing works on acquisition of data related to complex documents. In this context, complex documents are huge and concentrate on specific domains. These may be technical documents on specific domains related to different fields like computers, medicine, life science etc. Examples of such documents include research papers, technical blogs, surveys, government documents etc. With the advance in technology there is need for special tools and datasets specific for analyzing such domain specific and complex documents.

Many initiatives have been started in different domains for analyzing and collecting different agricultural documents. One such works in biomedical domain [30]. This work is based on Relation Extraction(RE). Using the advanced approaches which are combination of various methods, RE extracts simple as well as complex relations between different entities in the biomedical literature. Based on the F-Score, the paper compares the methods of supervised, semi-supervised, higher order relation and self-supervised approaches. A detailed data is required in case of the supervised approach, whereas a small knowledge base is enough for the semi-supervised approach to automatically annotate the features. Self-supervised approach is the combination of the supervised and semi-supervised approach. In higher order relation many different biomedical relations are connected together. The relation to relation F-Score performance varies when they are extracted.

Another work concentrates on the classification of the research paper subjects based on the interrelationship between them [31]. Searching and finding the relevant journals and papers for any given research area has been a real concern for researchers, professors and students. As it helps in the research and study, here subject classification of papers is used, subject classification facilitates the search process. The keyword based search or classification of such documents fails when semantically equivalent words are used in the document instead of exact words which happens most of the times in such technical documents. This paper proposes a novel supervised approach for the classification of the scientific articles' subject on the basis of their interrelationships. They have exploited the links such as common authors, citations, and common references to assign subject to papers. It is evident by the paper that the approach works better when the graph of relationships is dense which requires a large number of links between the papers.

There have been research towards classification of scientific document which takes document PDF as input and classification is performed based on the bag of words that is present[32]. In any research institute classification is performed superficially. The document classifying criteria are dependent on several people. The work related to identifying various sections of the paper, and then classifying them automatically and to perform inferences about them, instantiating them in an ontology are a few things that are described in the paper. The final ontology produced by this analysis can be referred to understand and draw conclusions regarding the documents.

Since domains and topics are important focus of such complete and technical documents, there has been work towards document analyzing keeping this in mind [33]. Simple algorithm for feature extraction has been explained in this paper. In case of the development-centric topics this algorithm can achieve high accuracy for document classification. Here, the main aspect is that in a development-centric topics, the feature extraction algorithm

doi: 10.32622/ijrat.76S201906

exploits two distinct aspects: a) unlike the semantically hard classification topics (like banks or chemistry) these tend to be highly focused; b) various region specific features are used by the authentic pages because of the local language and cultural underpinnings in these topics. Hence, a domain specific and development-centric technique has been designed for proper summarization and results.

A system called papits has also been developed in one of the works[34]. It is a research support system, and it shares research information. Different research information and the survey of the corpora of any given research field can be shared among Papits users.

VI. CONCLUSION

In this paper, we presented a survey of different papers with various techniques used for text categorization and summarization. We can see that each of the techniques has its own advantages for specific applications. However, techniques like CNNs and RNNs are emerging to be very popular due to their advantages of back propagation and internal memory.

Here we are concentrating on the application of these techniques for specific purpose of complex data analysis. This complex data may include domain specific research papers, articles etc. We can see from the available survey that some of the researches concentrate on providing useful domain specific and research specific analysis, and propose on using simple models and technical inter relations to provide good results.

REFERENCES

[1] Biancalana, C., &Micarelli, A. (2007, September). “Text categorization in non-linear semantic space”. In Congress of the Italian Association for Artificial Intelligence (pp. 749-756). Springer, Berlin, Heidelberg.

[2] V. K. Vijayan, K. R. Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 ICACCI, Udupi, 2017, pp. 1109-1113.

[3] V. Dalal and L. Malik, "A survey of extractive and abstractive text summarization techniques," 2013 6th International Conference on Emerging Trends in Engineering and Technology, Nagpur, 2013, pp. 109-110.

[4] R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2014, pp. 1-5.

[5] Munková D., Munk M., Vozár M. (2014) “Influence of stop-words removal on sequence patterns identification within comparable corpora”. In: Trajkovik V., Anastas M. (eds) ICT Innovations 2013. ICT Innovations 2013. Advances in Intelligent Systems and Computing, vol 231. Springer, Heidelberg

[6] Manalu, S.R. (2017). “Stop words in review summarization using TextRank”. 2017 14th International Conference on Electrical Engineering/Electronics,Computer,Telecommunications and Information Technology , 846-849.

[7] R. J. Prathibha and M. C. Padma, "Design of rule based lemmatizer for Kannada inflectional words," 2015 ICERECT, Mandya, 2015, pp. 264-269.

[8] F. M. Barcala, J. Vilares, M. A. Alonso, J. Grana and M. Vilares, "Tokenization and proper noun recognition for information retrieval," Proceedings. 13th International Workshop on Database and Expert Systems Applications, Aix-en-Provence, France, 2002, pp. 246-250.

[9] J. A. Bakar, K. Omar, M. F. Nasrudin and M. Z. Murah, "Tokenizer for the Malay language using pattern matching," 2014 14th International Conference on Intelligent Systems Design and Applications, Okinawa, 2014, pp. 140-144.

[10] S. Gadri and A. Moussaoui, "Information retrieval: A new multilingual stemmer based on a statistical approach," 2015 3rd International Conference on Control, Engineering & Information Technology ,Tlemcen, 2015, pp. 1-6.

[11] Jivani, A.G. (2011). “A Comparative Study of Stemming Algorithms Ms”.

[12] Porter, M. F. (1980). “An algorithm for suffix stripping”. Program, 14(3), 130-137.

[13] H. Zhang and D. Li, "Naïve Bayes text classifier," 2007 IEEE International Conference on Granular Computing (GRC 2007), Fremont, CA, 2007, pp. 708-708.

[14] Yu Wanjun and Song Xiaoguang, "Research on text categorization based on machine learning," 2010 IEEE ICAMS 2010, Chengdu, 2010, pp. 253-255.

[15] I. Kotenko, A. Chechulin and D. Komashinsky, "Evaluation of text classification techniques for inappropriate web content blocking," 2015 IEEE 8th International Conference on IDAACS: Technology and Applications, Warsaw, 2015, pp. 412-417.

[16] H. Liu, M. Cocea and W. Ding, "Decision tree learning based feature evaluation and selection for image classification," 2017 ICMLC, Ningbo, 2017, pp. 569-574.

[17] F. Harrag, E. El-Qawasmeh and P. Pichappan, "Improving arabic text categorization using decision trees," 2009 First International Conference on Networked Digital Technologies, Ostrava, 2009, pp. 110-115.

[18] Z. Wang, X. Sun, D. Zhang and X. Li, "An optimal SVM-Based text classification algorithm," 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 2006, pp. 1378-1381.

[19] X. Lin, H. Peng and B. Liu, "Support Vector Machines for text categorization in Chinese question classification," 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), Hong Kong, 2006, pp. 334-337.

[20] Jiu-Zhen Liang, "SVM multi-classifier and Web document classification," Proceedings of 2004 International Conference on Machine Learning and Cybernetics, Shanghai, China, 2004, pp. 1347-1351 vol.3.

[21] M. Farhoodi and A. Yari, "Applying machine learning algorithms for automatic Persian text classification," 2010 6th International Conference on Advanced Information Management and Service, Seoul, 2010, pp. 318-323.

[22] S. Patil, A. Gune and M. Nene, "Convolutional neural networks for text categorization with latent semantic analysis," 2017 ICECDS, Chennai, 2017, pp. 499-503.

[23] Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., ... & Dean, J. (2013). “Zero-shot learning by convex combination of semantic embeddings”.

[24] A. Hassan and A. Mahmood, "Efficient deep learning model for text classification based on recurrent and convolutional layers," 2017 16th IEEE ICMLA, Cancun, 2017, pp. 1108-1113.

[25] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," 2017 ICCCSP, Chennai, 2017, pp. 1-6.

[26] Laudauer, T., Foltz, P., &Laham, D. (1998). “Introduction to latent semantic analysis”. Discourse processes, 25(2/3), 259-84.

[27] Malviya, S., &Tiwary, U. S. (2016). “Knowledge based summarization and document generation using Bayesian network”. Procedia Computer Science, 89, 333-340.

[28] F. Alshuwaier, A. Areshey and J. Poon, "A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining," 2017 4th IEEE ICETAS, Salmabad, 2017, pp. 1-13.

[29] Taheriyan, M. (2011, August). “Subject classification of research papers based on interrelationships analysis”. In Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation (pp. 39-44). ACM.

[30] J. C. Rendón-Miranda, J. Y. Arana-Llanes, J. G. González-Serna and N. González-Franco, "Automatic classification of scientific papers in PDF for populating ontologies," 2014 International Conference on Computational Science and Computational Intelligence, Las Vegas, NV, 2014, pp. 319-320.

[31] Power, R., Chen, J., Kuppusamy, T. K., & Subramanian, L. (2010, March). Document “Classification for focused topics”. In AAAI Spring Symposium: Artificial Intelligence for Development.
Ozono, T., &Shintani, T. (2006). “Paper classification for recommendation on research support system papits”. IJCSNS , 6(2006), 17-23.

doi: 10.32622/ijrat.76S201906