International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Text Mining For Multiclass Research Paper Categorization

Amber Saxena, Anamika, Bhaskar Pant, Vikas Tripathi

Abstract*: *A research paper is a rich source of academic and innovative writing on a particular topic, and they are unstructured in nature. Categorization of documents refers to classification of documents in classes that are predefined. It is arduous for a user to categories research paper in different domains: because extracting meaningful and relevant words from the research paper is a challenging task.
For extracting important information we have used certain methods and classifiers. Methods like bag of words and tfidf is used for processing data. Prepossessing the data includes string tokenizing and stop-word removal. Then the processed data is classified using SVM classifier. For multiclass classification; since predefined classes are 4, therefore 1-v-r classifier is used. The system performance is 88% with 800 training and 200 testing documents. It is analyzed that the model performs better when the training data is more. The aim of this work is to categorize the documents and allocate set of predefined tag to them. It also evaluates the performance of the model by considering different percentages for training and testing sets of documents.

Keywords*: *Categorization, tf-idf, bag of words, SVM, one-versus-rest.

I. INTRODUCTION

Dramatically quick and rapid growth of data and research works available, make arduous for a user to categories those research paper in different domains. A research paper is a rich source of academic and innovative writing of an author on a particular topic. A research paper is generally unstructured in nature. The domain or the category of the research finding in the research paper is not properly mentioned. This arises to the need for research paper categorization.Categorization of text documents refer to automatic classification of a set of documents in predefined classes. With exponential growth of the data it has become more conceptual and relevant to categorize things/documents e.g. categorization of news, indexing of patients data on the basis of multiple factors like disease, surgical procedure, medication given, organization of emails into various groups (such as social, primary and promotions).

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Amber Saxena*, Department of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.
Anamika, Department of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.
Bhaskar Pant, Department of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.
Vikas Tripathi, Department of Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.

An interesting application of text categorization is Research Paper Categorization.Unstructured data in the form of text is everywhere: emails, web pages, research papers, social media and many more. The data inside the unstructured data is an extremely rich source of information and statistics. Extracting insights from it is not easy and also time-consuming due to its unstructured nature. Proper categorization of research papers requires machine learning, text retrieval and NLP (natural language processing).Research paper categorization is the process of assigning category to research paper (text document) according to its content. For example a research paper containing the words like computer, recommendation, cloud etc can be categorized under Computer Science category, whereas papers containing words like machine, engine, motor etc can be categorized under Mechanical category. Extracting meaningful and relevant words from the research paper is a challenging task. Researchers handled the text categorization problem in many ways. Naïve Bayes (NB)[4], Probabilistic Bayesian models [15, 7, 12], K Nearest Neighbors (KNN) [18], Decision Tree (DT) [6, 9, 15], Decision Rules [11, 4], Support Vector Machine (SVM)[8] and hidden Markov model (HMM) are fundamentally commonly used text classification and categorization.The aim of this research is to categorize the research papers considering its textual content. The task is to automatically categorize text documents into its predefined classes based on content.

II. MATERIAL AND METHOD

The need of research paper categorization system is because of information overloading. This over loading of information i.e. the overabundance of research papers makes the information seeking a challenging task. It increases the time to access and categorize a research in a desired domain. The different approaches used in categorization of research paper are: Bag of words, Tf-id.

A. BAG OF WORDS:

The bag of words is a simple method of feature extraction of the textual data. Machine learning algorithms are used to model the extracted features from text. The bag-of-words is simple model and can be used in solving problems such as language modeling and document categorization.
Algorithms of machine learning cannot work with unstructured and raw text directly therefore the text should be converted into numeric structured data. The text inside the document is converted into vectors of numbers. The vectors derived from textual data mimic various linguistic properties of the text. This is called feature extraction.

For example, we assume every line as a separate document here.
i like biscuits.
i like coffee.
i like biscuits and coffee.
The unique words here are: “i”, “like”, “biscuits”, “coffee”, “and”
Now create a document vector of fixed length as that of the number of unique words. And convert it to binary vector where 1 shows the presence of the word in the document, and 0 represents that the particular word is not present in the document.
Binary vector for the above listed documents are:
i like biscuits [1, 1, 1, 0, 0]
i like coffee [1, 1, 0, 1, 0]
i like biscuits and coffee [1, 1, 1, 1, 1]
Bag of words represents two things:

- A vocabulary of known distinct words.
- A depiction of the presence of known words (vector).

The model is only concerned with whether word occur in the document or not, and is not concerned about the position of word in document.

B. TF-IDF:

Term frequency-inverse document frequency is a statistical approach to evaluate the importance of a word to a document, and that document is in the collection of other documents. Tf-idf is commonly used in text mining, information retrieval and user modeling. Google also uses tf-idf to assign rank to the web pages, i.e. uses tf-idf as web crawler, because search engine focus more on term frequency rather than counting words. It has been observed that more than 83% of text based recommendation system uses tf-idf.Tf-idf typically defines two terms: Term Frequency (TF) and Inverse Document Frequency (IDF).
TF (Term Frequency) is the ratio of the number of times a word appears in a document by the total number of the words in the document. It measures how frequently a word appears in a document.
IDF (Inverse Document Frequency) calculated as the logarithm of ratio of the number of the documents in the database to the number of documents where the individual term occurs. It measures how important a word is to a document. It is observed that the words like: “is”, “a”, “the” may appear lot of time, but these words are less important to document. Thus IDF helps to find out frequent terms and rare ones.
TF-IDF = tf * idf ; higher the tf * idf weight (value), unique the word.
Calculating Tf-idf:
For example in a document of 1000 words recommend repeat 20 times.
Then tf = 20/1000 => 0.020
Now if we have dataset of 1200 documents in which recommend appears 50 times.
Then idf = log (1200/50) => 1.380
Therefore, tf-idf = 0.020 * 1.380 => 0.0276.

III. DATA COLLECTION

The dataset consists of 1000 research papers. These research papers are broadly classified into 4 major categories.
The four categories are:

i. Computer Science
ii. Mechanical
iii. Biotechnology
iv. Electrical

Each category contains 250 research papers. The dataset is collected from different online sites and even Graphic Era University provides the dataset for this research work.
Data is categorized into two subsets: training set: the set of the data used to create and train the model for desired output. Test set: this set of the data is used for testing the model. For example if training percentage is 80% and testing percentage is 20%. Then out 250 papers in each domain 200 are classified under training dataset while other 50 are used as testing dataset. That is there are total 800 training research papers and 200 testing research papers.

IV. IMPLEMENTATION

- Tokenization: Tokenization is the approach of segmenting text into words and sentences. Digital text is a stream of characters or words. Text needs to be segmented into basic units such as words, numbers, punctuation, alpha-numerics etc. before actual text processing. Pre-processing of a text document and an identification of basic units of text to be processed is known as tokenization.
- Eliminating Stop Words: Stop words, or stop words, should be filtered out prior to the processing of document text. These are functional words which do not have informational meaning. Even Search Engines do not consider stop words in order to save space in the disk and to speed up search results.
- Information Extraction: Information extraction is the process of extracting structured information from semi-structured and unstructured documents. It processes texts with the help of natural language processing (NLP).
- Term weighting for text extraction: Text in Research Paper is collection of different sentences and paragraphs. For text extraction we represent our textual data in vector space. The data in document in vector space is represented as vectors where components correspond to terms contained in the document and their value indicates weight-age for the term. The weight is considered to be the important concept for text categorization. The importance of the term to that document is represented by the weight of that term. Term frequency (TF) of word in a document is represented by the number of times a word appears in a document divided by the total number of word in the document. Inverse document frequency (IDF) measures the distribution of each word in the collection of documents i.e. it is the logarithm of the number of documents in the database divided by the number of documents where that particular term appears. IDF shows the importance of word in the corpus.

The weight (wt) of term t in document d is represented as :-

Wt = TF (t, d) * IDF (N, n)

TF (t, d) represent the ratio of the number of times t appears in the document d.

IDF (N, n) represents the logarithm of the total number of documents in the corpus(N) to the number of documents where that term appears(n).

The terms in the documents are ranked by their weights using TF-IDF function. The TF-IDF value should be high so as to show the importance of the word to the document. It also interprets that the word is rarer. The rare word signifies that the word is significant in summarizing the document in which it is present.

- SVM Classifier: SVM (Support vector machine) is an approach of decision plane to depict decision boundaries. Set of objects with different class labels are separated by a decision plane. SVM sets a decision plane to separate the training set data into different classes and make decisions based on support vectors. Therefore the goal of SVM is to define optimal separating hyper-plane that categories new examples. The optimal separating hyper plane is that which have maximum margin between two classes.

  SVM are designed to classify linear separating binary classes. For multi categorization we use 1-V-R approach i.e. one-verse rest. It is necessary to change the problem to multiple binary classes in order to apply SVM for multi classification. The 1-V-R concept involves constructing of a binary classifier for each unique class. Each class separates objects belonging to that class from other that do not come. In 1-V-R there are N classes, thus 1-V-R will create N binary classifier. The categorized label for document in 1-V-R SVM is the class/label that corresponds to the SVM with the highest value.

PROCESS:

A document is taken and based on TF-IDF values of words, the most recurring words are extracted. For deciding the recurring words, TF-IDF threshold is varied to get the number of recurring words in a document in the range of 30 to 50 words. This same procedure is applied to all 1000 documents. Then, these documents are converted into vectors using 'Bag of words' approach considering recurring words only.These vectors are split into training and testing in the ratio 70 to 30. Then, using the training data, SVM model is trained and the accuracy is measured using the testing data.The text categorization is defined as a categorization model to assign predefined class to new document. A classifier for categorizing research document is built by applying one-versus-rest method to a training set of objects. This classifier is henceforth used to predict the predefined tags to new papers.

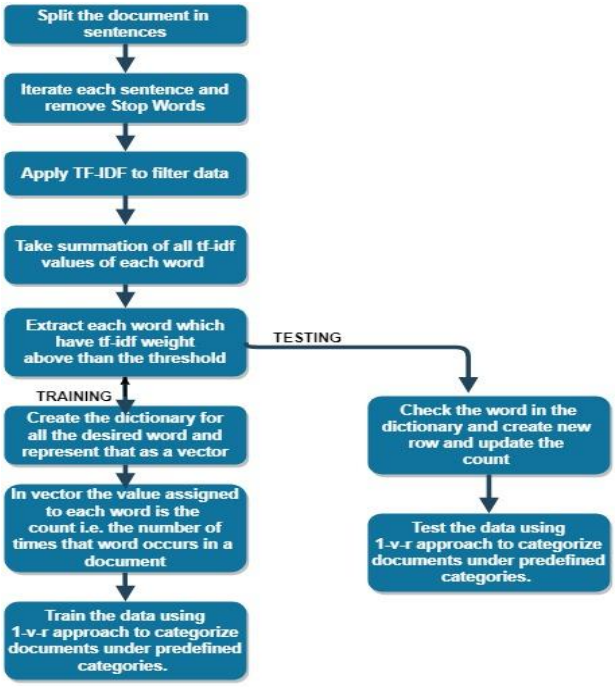

Figure: A flow to depict overall working model of categorization system.

V. RESULT

Table 1: Documents considered for training and testing purpose.

DOCUMENT CONSIDERED	No. OF DOCUMENTS
Training set	700
Testing set	300

The documents are classified as training and testing set of data. The training data is used to train the model, while testing data are used to analyze model performance.
The model is analyzed with different percentage of training and testing dataset.

Table 2: Comparison of accuracy with respect to training and testing dataset.

Training set percentage	Testing set percentage	Accuracy
70	30	86.67
80	20	88.00

It is analyzed that the model performs better when the training data is more. And also as the system performance is 88% with 800 training and 200 testing documents which shows that the system is feasible to categorize documents in different predefined categories. Also this accuracy can be increased when the size of training dataset increases.

VI. CONCLUSION

This paper introduces a technique for research paper categorization. We have used four categories for a total of 1000 documents divided as follows: Computer Science (250 documents), Mechanical (250 documents), Biotechnology (250 documents) and Civil (250 documents).

Research papers have data unstructured in nature. The data inside the unstructured data is an extremely rich source of information and statistics. Research paper categorization helps in extracting insights from it. For extracting significant information we have used certain methods and classifiers. Prepossessing the data includes string tokenizing and stop-word removal. For processing data Bag of words and Tf-idf approach is used. Then the processed data is classified using SVM classifier. Support vector machine classifies binary classes. For multiclass classification; since our predefined classes are 4, we have used 1-v-r classifier. As there are 4 classes then this classifier will create 4 binary classifiers. The research paper is categorized under that predefined class which has highest weight. Documents are divided into training sets and testing sets in different ratios i.e. 70-30 and 80-20. The accuracy increases when the size of training dataset increases i.e. when we partition in 80-20 ratio.The number of research papers and text documents in the past decades has increased rapidly. Aftermath of this rapid growth emphasized to classify these documents into classes or groups that define the content of the document. The proposed system could be used for categorizing research paper and text documents to user.In future, text categorization can be extended for web page categorization, electronic-mail categorization. Performance can be improved by designing and implementing hybrid networks to categorize a huge set of documents.

REFERENCES

1. Shugufta Fatima, Dr. B. Srinivasu, "Text Document categorization using support vector machine", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 02 | Feb - 2017.
2. Sihem Khemakhem; Younes Boujelbene, "Support vector machines for credit risk assessment with imbalanced datasets", International Journal of Data Mining, Modelling and Management, 2018 Vol.10 No.2, pp.171 – 187.
3. Pawan Lingras, Cory Butz, "Evaluation and Simplification of rules created by 1-v-r Rough SVM multiclassification", Fuzzy Information Processing Society 2006. NAFIPS 2006. Annual meeting of the North American, pp. 553-558, 2006.
4. D. Lewis, " Na¨ıve (bayes) at forty: The independence assumption in information retrieval", 10th European Conference on Machine Learning (ECML-98), pages 4–15.
5. S.S. Desai; D.N. Kashid, "Estimation of regression parameters using SVM with new methods for meta parameter", International Journal of Data Mining, Modelling and Management, 2015 Vol.7 No.3, pp.239 – 256.
6. Zhi-kun Hu, Wei-hua Gui, Chun-hua Yang, Peng-cheng Deng, Steven X. Ding, "Fault classification method for inverter based on hybrid support vector machines and wavelet analysis", International Journal of Control, Automation and Systems, vol. 9, pp. 797, 2011.
7. W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules", IEEE International Conference on Data Mining (ICDM'01), San Jose, California.
8. B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining", ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'98), pages 80–86, New York City.
9. T. Joachims, "Text categorization with support vector machines: learning with many relevant features", 10th European Conference on Machine Learning (ECML-98), pages 137–142.
10. W. Cohen and Y. Singer, "Context-sensitive learning methods for text categorization", ACM Transactions on Information Systems, 17(2):141 –173.
11. M.-L. Antonie, O. R. Za¨ıane, and A. Coman, "Application of data mining techniques for medical image classification", Second International ACM SIGKDD Workshop on Multimedia Data Mining, pages 94–101, San Francisco, USA.
12. The reuters-21578 text categorization test collection. http://www.research.att.com/˜lewis/reuters21578.html.
13. Hima Suresh; Gladston Raj. S, "An innovative and efficient method for Twitter sentiment analysis" International Journal of Data Mining, Modelling and Management, 2019 Vol.11 No.1, pp.1 – 18.
14. Rajni Jindal; Shweta Taneja, " A lexical-semantics-based method for multi label text categorisation using word net", International Journal of Data Mining, Modelling and Management, 2017 Vol.9 No.4, pp.340 – 360.
15. H. Li and K. Yamanishi, "Text classification using esc-based stochastic decision lists", 8th ACM International Conference on Information and Knowledge Management(CIKM99), pages 122 –130, Kansas City,USA.
16. Aaqib Saeed, "http://aqibsaeed.github.io/2016-07-26-text-classification/".
17. Ahmed H. Aliwy1 and Esraa H. Abdul Ameer, "Comparative Study of Five Text Classification Algorithms with their Improvements", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 14 (2017) pp. 4309-4319.
18. O. R. Za¨ıane and M.-L. Antonie. Classifying text documents by associating terms with text categories. In Thirteenth Australasian Database Conference (ADC'02), pages 215–222, Melbourne, Australia, January 2002.

AUTHORS PROFILE

Amber Saxena, I am persuing B Tech in Computer Science and Engineering from Graphic Era Deemed to be University, Dehradun. I have done internship in Analytics Vidhya as a data science intern, GeeksforGeeks, O.C.F(Ordnance clothing factory) and RWX technology. I have done several research publication and projects in data science, machine learning and natural language processing such as Recommendation System, Air Quality Prediction, Text Summarization.

Anamika, I am pursuing B.Tech. in Computer Science and Engineering. I have worked as an intern in Ministry of Electronics & Information Technology in Research and Development Department, National Institute of Technology Delhi and also as a content writer for geeksforgeeks.
I had done research in predictions and recommendation system. I have worked on projects like Emotion based Music and Quotation recommendation system, Movie Genre recommendation system.

Dr.Bhasker Pant, Currently working as Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal.
Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded) and 5 candidates are in advance state of work. He has also guided 28 M.Tech. Students for dissertation. He has also supervised 2 foreign students for internship.
Dr.Bhasker Pant has more than 70 research publication in National and international Journals.
He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.

Dr. Vikas Tripathi has done BE in information technology from Technocrats institute of technology, Bhopal, M. Tech in Software engineering from Indian institute of information technology Gwalior and PhD from Uttarakhand technical university, Dehradun.
He is actively involved in research related to Software engineering, Computer Vision, Machine learning and Video Analytics. He has published many papers in reputed international conferences and journals. Currently he is working as an associate professor in Graphic era deemed to be university Dehradun, India.