REVIEW ON APPLICATIONS OF TEXT MINING

[1]Saurav Chaurasia
[1]Student of Master of Science
[1]Department of Computer Science and Application
St. Aloysius' College (Autonomous), Jabalpur (M.P.), India

Abstract :The amount of online unstructured data is exponentially increasing in relevance and quantity,the applications of text mining is believed to possess high business price. Text mining provides us with highly potential textual data from abundant amount of unstructured data. This paper provides a glimpseof various applicationsused in text mining. Finally we tend to offer other application areas wherever text mining may be helpful for extracting purposeful information.

***IndexTerms* - Categorization, Clustering, Information Extraction, Information Retrieval, Natural Language Processing (NLP), Text mining.**

I.INTRODUCTION

Online surveys have opened many opportunities for data collection and retrieval. Text mining has come into existence from information extraction, retrieval and summarization. Sometimes text mining is also referred as text analytics. The process of text mining undergoes different steps such as cleaning the data, integrating the data and selecting the appropriate data.

At present data is available for text mining purpose is not structuredas maximum amount of online data is unstructured.Text mining techniques performs various operations that breaks the unstructured data into structured tokens. These tokens help identifying the required information. In this era of advanced technology, data is continuously being generated through different means like twitter, whatsapp, facebook and etc. To produce meaningful and structured information from meaningless and unstructured data text mining is being widely used.

II.TECHNIQUES OF TEXT MINING

To gain useful information from raw data various techniques of text mining are used which are as follows:

2.1 Information Extraction

Information Extraction (IE) extractsunstructured textual matter and produces structured information and this process is done automatically. Further extracted information is then stored in a database for future access. IE process involves various subtleties and complex techniques. IE process involves following main subtasks: Pre-processing of the text, Finding and classifying concepts, connecting the concepts, Unifying, Getting rid of noisy data.

2.2 Information Retrieval

Information Retrieval (IR) extracts relevant information or documents from a large pool of data. Internet has almost all kind of data stored at various locations. IR systems retrieve the data which is most relevant to the user query. IR system tracks the utility of the displayed information i.e., the information useful for user or not. Google and Yahoo search engines are the two most renowned IR systems.

2.3 Categorization

Text categorization/text classification is a form of supervised machine learning that performs the task of assigning predefined categories to normal language text depending upon their genre. It classifies your content and products into categories to help users search and navigate easily within website or application. Text Classification helps Google crawl websiteswhich helps in Search Engine Optimization (SEO).

2.4 Clustering

Clustering performs division of objects to make group of objects showing similar features into a clusterin an unsupervised way. Clustering algorithms takes set of texts and provides list of detected clusters. Clusters are assigned with a name, a related value and size. Various clustering methods are available such as: Standard clustering methods, specialized text clustering methods and Joint cluster analysis and these methods further contains sub-methods.

2.5 Summarization

Text summarization presents a summary of documents with only main points outlined. There are various types of text summarization methods based on input type, output type and purpose. Single document and multiple document methods are based on input type. Extractive and Abstractive methods are based on output type. Generic, Domain-specific and Query-based methods are purpose based. Open Text summarizer, Classifier4J, NClassifier, CNGLSummarizer are few widely used open source summarization tools.

III.LITERATURE REVIEW

Author improves traditional text mining drawbacks using ontology methods e.g., domain ontology. A space vector model based on domain ontology is employed to analyze text data and obtain potential valuable knowledge. Data is taken from agricultural encyclopedia column of Chinese agriculture Academy science website which contains 400 documents. In cluster analysis, results of accuracy rate, recall rate and F-measure are compared and it is found that SVM based on domain ontology (CSVM) produce much more accurate results than SVM. [1]

Twitter is most generally used social media platform for expressing individual thoughts everywhere in the globe. The target of author is to predict sentiments inclination of Indian individuals towards political state of affairs and issue by using various classifiers. Twitter API v1.1 is employed for assortment of raw tweets, once preprocessing thirty five percent of original tweets are remains. Author used various classifiers: k-nearest neighbor, Random Forest, Naïve Bayes, Bays net and ensemble of classifiers for analyzing the feelings of users. Among all data processing classifiers, k-nearest neighbor classifier is the best for sentiment prediction of tweets as a result it provides terribly high accuracy or terribly correct prediction of individual's sentiments as compared to alternative data processing classifiers on an individual basis or ensemble of classifiers with or without stopwords.[2]

Author presented a broad analysis of risks involved in engineering projects and describes various methods for risk management. Methods of Risk management are divided into two categories: Methods of risk identification, methods of risk assessment and evaluation. These categories have further sub-categories that contain various methods for risk management. At the end of this paper, author suggested several risk management strategies to deal with various kinds of risks in engineering project. [3]

This paper relies on the rail accidents happened in USA from year 2001 to 2012 (i.e.11 years). Author(s) covers incidents with extreme accident damage solely. Boosting and Bootstrap aggregation or fabrics are two strategies used for analyzing information. Information is categorized into various tables for higher understanding the contributors of rail accidents. Random forests, Gradient boosting and Ordinary Least Squares(OLS) techniques conjointly helps in rising the probabilities of higher understanding of the contributors to the rail accidents. [4]

The target of authors is to build an application that generalize the resumes of candidates and provide Knowledge Profile (KP) for software engineering positions. Author introduced a prototype i.e., KP Generator prototype with the aim of identifying Technical Knowledge (TK) of candidates for hiring process of software engineering positions. It has a knowledge base which is used to create Knowledge Profiles(KP) and to find candidate TKs. Authors selected 40 resumes to test the KP Generator prototype. Each resumes was searched and taken from the web. After analyzing the results it is concluded that with the help of KP Generator it is possible to identify the Knowledge Profiles (KP) for software engineering positions. [5]

Author performed sentiment analysis on various smartphone opinions collected online. In the sentiment analysis process, first sentiments are identified then features are selected so that sentiment can be easily classified. Data is taken from a very well-known marketing site known as amazon. Dataset consists of 1000 smartphone reviews with approximately 3000 sentences.Naive Bayes provides 40% of accuracy which is not satisfactory for text mining. SVM increased the accuracy of result up to 90% which is satisfactory and more reliable than Naive Bayes classifier.[6]

The researchers created their own information set by gathering tweets regarding GST exploiting numerous resources. Tweets are collected from 27th June to 7th July 2017 i.e., 10 days when this topic is trending at social media platforms. Authors employed Naive Bayes algorithm to perform sentiment analysis on GST tweets. RapidMiner tool is additionally used for merging the dataset in an excel file. The conclusion of this paper says that quite five hundred of the folks have positive opinion regarding GST in India during June-July 2017. Whereas quite two hundred fifty of the folks have negative opinion and quite two hundred fifty of the folks have neutral opinion. [7]

Data is taken from OpinRank dataset which contains data from 80 to 100 hotels in 10 different cities across the world. A sub dataset of reviews of selected hotels from London, Beijing and Montreal is created. Author used Sentiment Polarity Based Model (SPBM) for classification of sentiments generated from reviews. After analyzing, it is found that NBM algorithm has the highest precision, followed by CNB algorithm which has the second highest precision and CHIRP has the lowest precision. [8]

This paper is based on sentiment analysis of text data available on web. CNN is employed as associate automatic feature learner associated; SVM is employed as an emotional classifier. The work given is said to be classification of emotions and deep learning. The Author has used Word2vec tool introduced by Google that turns the text in vector form which are understood by deep learning. Data is taken from NLPCC2014 dataset based on deep learning technology. The data contains 10000 of the training data and 2500 test data. CNN, NLPCC_SCDL_best and CNN-SVM methods are used for analyzing the result, after analyzing CNN-SVM method has given highest accuracy in comparison to other applied method. [9]

10000 PubMed abstracts were identified with anti-epileptic drugs (AED) provided from 1st January 2007 to 2017 were taken. R-programming based PubMed scraper is utilized to download 10000 abstract positive to either of these watchwords: 'anti-epileptic drugs', 'anti convulsant drugs' and 'AED'. The given dataset was joined with US FDA sedate store to discover the most vital territories of research. Subsequent to breaking down this paper it is discovered that regardless of potential utility of medications in treatment of different ailments, their utilization can be vastly affected by human sentiments. [10]

The above literature review is summarized as:-

Table. List of studies for Applications of Text Mining

S. No.	Year	Title	Dataset	Technique(s)	Tool(s)	Results	Future work
1	2014	Research on Text Mining Based on Domain Ontology	400 documents are taken fromagricultural encyclopedia column of Chinese Agriculture Academy Science website	k-means based on space vector model (SVM) and k-means based on conception space vector model(CSVM)	English WordNet and Chinese HowNet Dictionaries	k-means based on conception space vector model(CSVM)produce much more accurate results than k-means based on space vector model(SVM)	
2	2015	Sentiments Analysis Of Twitter Data Using Data Mining	Training and testing tweets collected from twitter by using twitter searched API v 1.1 for various political leaders and parties in India	k-nearest neighbor, Random Forest, Naïve Baysian, Baysnet singly or ensemble of classifiers for analyzing the sentiments of users	Twitter API v1.1, SentiWordNet 3.0.0 dictionary	Among all data mining classifiers k-nearest neighbor classifier is the best for sentiments prediction of tweets. It gives very high accuracy or very accurate prediction of people sentiments in comparison to other data mining classifiers singly or ensemble of classifiers with or without stop-words.	
3	2016	Research on Risk Management of Engineering Project		Methods of risk identification and methods of risk assessment & evaluation		Risk management methods have great affect in quality improvement and cost reduction of engineering project.	Due to development of science, technology and economy day by day risk management and its related technology will have a great future scope. Further development and improvement of risk management project is required because of emergence of new problems.
4	2016	Text Mining the Contributors to Rail Accidents	Federal Railroad Administration (FRA) dataset from year 2001 to 2012 (i.e.11 years)	Boosting and Bootstrap aggregation or bagging		It is determined that the accuracy of models for predicting rail accidents severity can be improved. Random forests, Gradient boosting and Ordinary Least Squares(OLS) techniques also helps in improving the chances of better	Optimization of text mining technique can be done to further improve the result. Study of accidents with extreme number of casualties is needed to determine their contributors. Then compare the similarities and

						understanding of the contributors to the rail accidents.	differences of these contributors to those of accidents with extreme costs.
5	2017	Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions	40 resumes were selected from the web as testing data.	Natural Language Processing (NLP) and Text Mining	KP Generator Prototype	To identify the Knowledge Profiles(KP) for software engineering positions and the time has been reduced that HR recruiters expand reviewing resumes of the candidates.	To analyze how much effort of HR departments of organizations can be reduced in order to fulfill software engineering job positions and also add other functionalities with KP Generator for improving the results.
6	2017	Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews	A Diverse Dataset obtained from Amazon	Naïve Bayes and Support Vector Machine Classifier		Naive Bayes technique provides accuracy of 40% that is not satisfactory for text mining. To increase the accuracy of result support vector machine(SVM) technique is used which increased the accuracy of result upto 90%.	
7	2017	Sense GST: Text Mining & Sentiment Analysis of GST Tweets by Naïve Bayes Algorithm	Gathered two sets of data clusters: (i) Data containing positive and negative words used as training data (ii) Tweets data used as test data	Naïve Bayes Algorithm	RapidMiner Tool	More than 50% of the people have positive opinion about GST in India during June-July 2017. While more than 25% of the people have negative opinion and more than 20% of the people have neutral opinion.	Use Python Natural Language Toolkit(NLTK) manually and compare between the results of future work and current work.
8	2018	A Framework for sentiment analysis with opinion mining of Hotel Reviews	OpinRank Dataset of approximately 2,59,000 unlabelled reviews on cars and hotels from 80 to 100 hotels in 10 different cities across the world	Naïve Bayes Multinomial(NBM), Sequential Minimal Optimization(SMO), Compliment Naïve Bayes(CNB) and Composite Hypercubes on Iterated Random Projections(CHIRP)		The Naïve Bayes Multinomial algorithm had the highest precision which reached 80.9%. It was closely followed by the Compliment Naïve Bayes algorithm which had 80.5% precision.	Fine tuning the feature extraction algorithm of the framework so that classification error is minimized. Improve automatic labeling, feature extraction and perform classification of customer responses based on emotions using deep learning

							algorithms.
9	2018	Research on Text Sentiment Analysis based on CNNs and SVM	NLPCC2014 emotional analysis evaluation task data set based on deep learning technology	Traditional Convolutional Neural Networks (CNN),NLPCC _SCDL_best method and CNN-SVM	Word2vec tool introduced by Google	CNN-SVM model has the highest accuracy in comparison to Traditional CNN and NLPCC_SCDL_best method.	
10	2018	The application of text mining algorithms in summarizing trends in anti-epileptic drug research	Dataset of 10,000 PubMed abstracts related to anti-epileptic drugs(AED) published between 1 January 2007 to 1 January 2017	Modified Latent Dirichlet Allocation (LDA) algorithm	R-software based PubMed scraper	Despite potential utility of drugs in treatment of various diseases, their use could be greatly affected by sentiments.	To understand the casual relationship between the negative sentiments and the pharmacological profile of anti-epileptic drug.

IV.Conclusion

We have introduced text mining and their techniques used in various application areas.Naïve Bayes Classifier is widely used by the researchers for their experimental purposes. Hence text mining plays a very strong role in modern information retrieval software either small or big. We believe that text mining results can be more accurate by further improving and classifying its knowledge base. In future text mining can be apply in the following application areas such as:

1. Selection of the best project report/ thesis among students of any university/college.
2. Predicting the effect of mandatory recharge policy by Telecom companies in India on customers and company itself.
3. Predicting the rise of Online Streaming Services in India.
4. Predicting the outcome of playing PUBG game in India.
5. Predicting the missing essence in Bollywood movies as compared to Hollywood movies.

V.Reference

[1] Li-hua, J., Neng-fu, X. and Hong-bin, Z. 2014. Research on Text Mining Based on Domain Ontology.International Federation for Information Processing (IFIP ‘14), 361–369.

[2] Jain, A.P. andKatkar,V.D. 2015. Sentiments Analysis Of Twitter Data Using Data Mining. International conference on Information Processing (ICIP ‘15) Vishwakarma Institute of Technology, 807-810.

[3] Sun, X. 2016. Research on Risk Management of Engineering Project. IEEE.

[4] Brown,D.E. 2016. Text Mining the Contributors to Rail Accidents. IEEE Transactions on Intelligent Transportation Systems.

[5] Valdez-Almada,R., Rodriguez-Elias, O.M., Rose-Gómez, C.E., Velázquez-Mendoza, M.D. J. and González-López,S. 2017. Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions: Generating Knowledge Profiles from Resumes. 2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT ’17), 97-106.

[6] Chawla,S.,Dubey,G. and Rana,A. 2017. Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews. 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO ’17), 377-383.

[7] Das, S. and Kolya,A. K. 2017. Sense GST: Text mining & Sentiment Analysis of GST tweets by Naïve Bayes Algorithm.2017 Third International Conference on Research in Computational Networks (ICRCICN ’17), 239-244.

[8] Zvarevashe, K. and Olugbara,O. O. 2018. A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews. 2018 Conference on Information Communications Technology and Society (ICTAS ‘18).

[9] Chen, Y. and Zhang, Z. 2018. Research on Text sentiment Analysis based on CNNs and SVM. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA ‘18), 2731-2734.

[10] Singh, S. P.,Karkare, S.,Baswan,S. M. and Singh, V. P. 2018.The application of text mining algorithms in summarizing trends in anti-epileptic drug research.preprint.