Exploring the Field of Text Mining

Radha Guha
DIT University
Dehradun, India

ABSTRACT
Text mining is the technique of automatically deducing non-obvious but statistically supported novel information from various text data sources written in natural languages. In the big data and cloud computing era of today huge amount of text data are getting generated online. Thus text mining is becoming very essential for business intelligence extraction as volume of internet data generation is growing exponentially. Next generation computing is going to see text mining amongst other disruptive technologies like semantic web, mobile computing, big data generation, and cloud computing phenomena. Text mining needs proven techniques to be developed for it to be most effective. Even though structured data mining field is very active and mature, unstructured text mining field has just emerged. Challenges of text mining field are different from that of structured data analytics field. In this paper, I survey text mining techniques and various interesting and important applications of text mining that can increase business revenue. I give several examples of text mining to show how they can be beneficial for extracting business intelligence. Using text mining and machine learning techniques new challenges for business intelligence extraction from text data can be solved effectively.

General Terms
Text mining

Keywords
Text mining, Business intelligence (BI), Unstructured data, Data analytics, Automatic text summary.

1. INTRODUCTION
Big data era is characterized by 5 V's namely volume, velocity, variety, veracity and value. To explain it a little more, the volume of internet or online data generation is increasing at an astronomical rate. Today social media like Facebook and Tweeter are generating terra bytes (TB) of data every day. Also government regulations are recommending to the transformation to a digital world where all service data are to be available online. This will enable consumers to access data anytime from anywhere to get a transparent view of operations by government sectors. There is also an initiative for a universal digital library taken by Carnegie Melon University (CMU) in 2005, to make online all published books of the world in all different languages. Other businesses are also transforming their operation from offline to online so that they can reach more customers easily and be more profitable. Internet is also a very popular media where users can express their emotion freely, be it on a social event, political agenda or product review. Often times all these opinions are put on the internet anonymously.

Total online data amount of today is several Zetta bytes (ZB) and it is growing rapidly. To keep up with this, businesses, processing this data need to be prepared for high velocity data as data growth can happen overnight. Eighty percent of online data is unstructured and in the form of text such as in news articles, research papers, blogs, emails, customer feedbacks, and social media data like in Facebook and Tweeter postings. In addition there are a lot of multimedia data in audio, image and video.

Much business intelligence can be found in this unstructured or semi structured data in variety of format only if it can be extracted efficiently on time. Veracity and value of the extracted information from different sources depends on how filtered the information is and how fast the information is extracted to be useful at the right time. As the data growth is exponential and at a very fast rate, human cannot process this data even after surfing the internet twenty four hours a day and seven days a week. On the contrary, computers are very good at processing huge amount of data quickly, only if humans have developed good methodologies and good algorithms for fast data processing.

Traditional data processing involves structured data in relational database management system (RDBMS). Online transaction processing (OLTP) system in data bases and data ware houses are there for data collection and modification, by customers and administrators. Structured query language (SQL) is there to get any information stored in the database or to update the data. To solve the challenges of structured data processing in RDBMS such as redundancy avoidance, transaction management, concurrency control and database consistency management there are proven techniques. Further data analytics field on structured data is also very active and matured. There are four different kinds of analytics like descriptive, diagnostic, predictive and prescriptive that are performed to transform the raw data to information and then to knowledge. Based on that knowledge, finally strategic decisions are made by the, "*C"* level executives in a business as per their experience and wisdom.

Text data mining paradigm is new and its challenges are different from that of traditional structured data processing. To tackle text data processing challenges, various text mining sub-areas or sub-tasks have evolved. In reference [1-8] text mining challenges and various interesting and important applications have been explored. In unstructured text data, there is no fixed attributes-value pair as in tuples of RDBMS tables. As there are no direct relations between documents, it is also difficult to aggregate concepts from various documents in heterogeneous formats for document classification and clustering. The task of identifying named entities in a single document and establishing their relations, known as information extraction, is another challenging task for text data.

The remaining of the paper has the following sections and sub sections. In sub section 1.1 text mining sub tasks are explored and in sub section 1.2 accuracy measures of these tasks are introduced. In Section 2, I do literature survey to

explore various interesting text mining applications and what text mining algorithms they have used. In Section 3, I have showed results of text summarizations by different techniques. Section 4, concludes the paper.

1.1 Text Mining Sub Tasks

The structured data processing challenges are not there in text data processing, as text data is mainly for reading and not for updating. The problem with text data is that it is unstructured. Various preprocessing techniques are required to transform the text data to numerical data so that powerful structured data analytics techniques can be applied. The broad umbrella term Text Mining consists of several sub tasks such as information retrieval, web mining, document clustering, document classification, information extraction, natural language processing and concept extraction.

Information Retrieval (IR): Collecting all the relevant documents out there mixed with all other irrelevant documents in the World Wide Web (WWW) databases is called information retrieval [9-10]. It is as tough as finding a needle in a haystack. This task of generating more accurate and relevant search results is merely based on approximate key words query. The distance between the document vector and query vector is the basis for information retrieval. The distance is measured as the cosine angle between document vector and query vector. Simple keyword based search or binary lexical search for information retrieval is not very efficient because of the problem of synonymy and polysemy. The result many miss the best response document or make the result too broad in topic selection. Synonymy decreases recall and polysemy decreases precision of information retrieval.

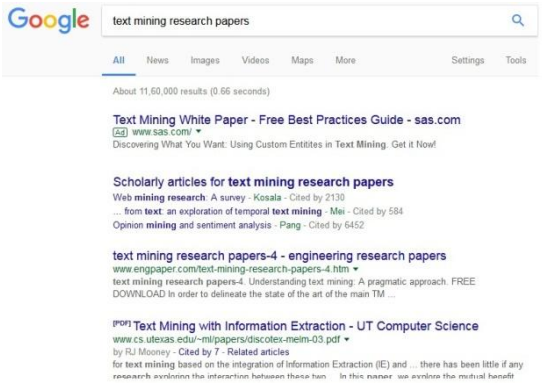

Figure 1: Information Retrieval by Google Search

Google and Yahoo like popular search engine giants try to make document retrieval or information retrieval from the web as efficient as possible with the current available techniques. They are using web crawler like program to crawl the World Wide Web systematically and automatically to collect useful data. These crawlers or robots, after retrieving the relevant documents based on key word search, are indexing them as per document importance. Information retrieval field is additionally employing techniques such as natural language processing, query processing, categorization and clustering of documents. Statistically based latent semantic indexing (LSI) is a popular information retrieval algorithm which can group related key words together to form concepts. LSI will retrieve documents on the same concepts even though the documents do not share any query words.

Even then, for an example, an average length simple key word search such as "Text Mining Research Papers" will list 84, 0000+ papers in a minute (Figure 1). Even though this list is ranked in order of their relevance, the abundance, misrepresentation and non-discrimination in the result set is very frustrating. No human can go through all of the retrieved documents individually and get proper understanding of their content. So again we need a computer to solve the challenging task of summarizing documents and extracting other structural, syntactic and semantic relation information from the documents automatically.

Web Mining: Web mining is a special kind of text mining [3]. As said before, web is a very popular and exponentially growing media where anybody likes to author content freely. World Wide Web is basically a collection of documents where some of them are connected with hyperlinks. But the heterogeneous format of web content makes it hard for mining. Web pages can be mined for topic summarization, sentiment analysis, finding named entities, structure, author, date of creation of document and relation between web pages etc. Another task of web mining is finding hub pages and more authoritative pages depending on number of incoming and outgoing links from/to other web pages.

Usually the plain text documents don't have that much structure information as a web page does. Web pages can insert a lot of metadata information with hypertext markup language (HTML) and extended markup language (XML) using resource descriptor framework (RDF) data model. Metadata means data about data. These metadata describes the text such as author's name, revision date, language, publisher, text length and all reading ease scores etc. of the web page. Web pages written in HTML and especially in XML markup languages can express many hierarchical semantic structures that can be processed by computers automatically. XQuery and Web Ontology Language (OWL) can be used to query the webpages collectively for getting relevant answer automatically. It was the W3C's recommendation to tag all web pages with XML and RDF ontology framework so that the whole web can be a connected database and machine can link all pages automatically [11].

Information extraction from web pages is easier than that from plain text documents. Hierarchical link structure of web pages can also be extracted identifying the more authoritative pages versus hub pages depending on number of incoming or outgoing links to or from the web pages.

Text mining field goes beyond information retrieval or collection of relevant documents. After collecting the unstructured text data from the web they need to be structured for subsequent data analytics tasks using the traditional data analytics algorithms. For this astronomical size data processing from various sources, Google and Yahoo have created open source Hadoop [12] and MapReduce like hardware and software architecture for massive storage and parallel processing using cheap commodity hardware. Hadoop framework achieves faster processing speed and more reliability than traditional RDBMS system as it replicates data in several servers and parallel processes them.

Text Data Preprocessing: All tasks of text mining need numerical representation of the text document. Thus text mining involves additional preprocessing steps to transform the unstructured data to structured data or converting the text format to numerical format first so that data analytics techniques can be applied to them. Heterogeneous sources of text data can be noisy and needs cleansing first. Cleansing means removal of irrelevant data, anomaly

correction, removal of outliers, substituting for missing data, standardization of data format, data normalization and transformation to higher concept data. Unless text data is cleansed and aggregated, the outcome will be as in garbage in garbage out (GIGO). The numerical representation of text happens as follows.

Unstructured to Structured Transformation: Each document is actually a bag of words or terms. A set of documents is called a Corpus or document database. After cleansing the document each word or term is considered as a token. The corpus is represented by Document Term Matrix (DTM) where each row represents a document and each column is either the frequency of the term or token (TF), present in the document or simply a presence (1) or absence (0) of the term in the document. Each term is also weighted according to its importance or discriminative power in the corpus. Thus if a term is present in less number of documents then that term has more discriminative power or weightage. This is called inverse document frequency (IDF) weight. Thus in DTM, term frequency (TF) multiplied by inverse document frequency (IDF) is a structured and numeric representation of each token in the corpus on which various standard data analytics algorithm can be applied. Each tuple in the DTM is a term vector in the vector space model of the corpus. DTM is the numerical representation of the text corpus.

Dimensionality Reduction: As the DTM matrix can have thousands of rows and millions of columns dimensionality reduction is very important for text data mining. In text mining, dimensionality reduction means selecting a subset of words or features from the DTM which can capture better variability or discriminative factors among the documents in the corpus. Some basic dimensionality reduction methods are stemming, stop-words removal, frequency thresholding and information gain analysis. Stemming is keeping only the origin of words. For example "expecting", "expected", "expect" all can be represented with "expect" only. Also English stop words like "the", "is", "am", "can", "also" are removed from the English text document. Very high frequency words and very low frequency words are removed as they have less discriminative power. I nformation gain technique analyses each word to find information gain because of that word and keeps only those words which have high information gain.

Further dimensionality reduction is done two ways, by feature selection or by feature extraction. Feature selection is choosing subset of features from the original set of features.

If there are *n* features in the original corpus, there will be 2^n different subsets to choose from, for the best subset of features. Feature extraction is deriving new features from the original features which are better representative for the original corpus. Latent semantic indexing (LSI) mentioned earlier as efficient information retrieval tool, also helps in dimensionality reduction.

LSI is based on singular value decomposition (SVD) technique which helps in dimensionality reduction. The original large feature-document matrix (M) can be decomposed by SVD method like this: $M = U\sum V^T$, where $\sum$ is the singular values for M. Considering only *k* largest singular values, M can be approximated as $M = U_k\sum_k V_k^T$. Using truncated SVD the latent semantic structure of the corpus can be better represented in reduced *k*-dimensional space. In this reduced *k*-dimensional space similar terms, and similar documents come closer to each other and dissimilar terms and documents become further apart. A new 'key words' query is also mapped in the reduced dimension and then the search proceeds in the reduced dimension. This low rank approximation of the original matrix is called LSI and it produces optimal approximation. Even though SVD is computationally costly, subsequent query searches in the low rank DTM will be faster. Further when document clustering is performed using Cosine distance or Euclidean distance, clustering results improve in the reduced *k*-dimensional space.

After preprocessing of text data i.e. converting unstructured or semi-structured data to structured data standard data analytics algorithm can be applied for text mining as well. Only not all data analytics algorithms are suitable for high-dimension text data analysis.

Document Clustering: Document clustering is an important text mining sub task for unsupervised labeling of documents. Document clustering is performed by measuring Euclidian or Cosine distance between the documents in the vector space model to find out their similarity or dissimilarity with each other. On that basis documents can be grouped into several clusters so that within the same cluster their distance is minimum and inter cluster distance is maximum. K-Means algorithm is an effective clustering algorithm for unsupervised predictive analysis of data. For example, if there is a corpus of newspaper articles, then they can be grouped or clustered into sports, politics, science, cuisine etc. automatically based on the distance of the term vectors in the vector space model.

Document Classification: For document classification or categorization there are many good algorithms in data analytics tool box. Document classification is a supervised machine learning (ML) technique where a model is constructed from training documents which are already labeled. And the new documents are labeled based on this predictive model. Decision tree based supervised machine learning techniques is not suitable for high dimension text documents. Thus dimensionality reduction by important feature selection is very important here. Other classification algorithms like Naïve Bayes classifier, K-nearest neighbor (KNN) and Support Vector Machine (SVM) are some supervised classification algorithms which are also good for high dimension data and can be used for document classification. Accuracy and efficiency of text classification model is important. The accuracy of the model depends on the number of training set data and its fitness.

With so many classification algorithms, the suitability of an algorithm depends on ease of understanding, ease of model building, computation cost and result accuracy measures. Even though there is no single superior algorithm, SVM is good for concept extraction from documents and thus groups related documents more accurately.

Information Extraction (IE): Information extraction is a major part of text mining. After retrieving relevant documents from the web, information extraction (IE) from each document is required for deeper understanding of a document. IE is used to summarize the content of each document. It is the information extraction technique that extracts the explicit or implied semantic metadata in a document and creates a hierarchical structure for this metadata. In IE, unstructured text document is converted into structured database which can be further analyzed by standard data mining algorithms. Identification of proper nouns as named entities (NE), such as person, place, organization, date, time and establishing their classification

and semantic relations is what comprises of knowledge discovery as part of information extraction.

With traditional search engines getting the facts in documents is hard and slow. IR system retrieves whole document where the facts may be somewhere there or not. A human still has to read the lengthy document to search for the facts in IR. Information extraction task is much easier on semantic web than on plain text documents. As mentioned before, following W3C's recommendation all web publishing needs to be semantic web pertaining to a particular ontology so that they can be automatically processed by computers for information extraction and knowledge discovery.

Visualization: Visualization is the last step of text data mining and is very important to understand the complexity of relationship in text data. There are many visual graphing techniques such as charts, histogram, box plot, scatter plot, word cloud etc. The suitability of selecting one depends on the business domain and data types. Text mining can use a lot of visualization to navigate and explore concepts and relations more effectively.

1.2 Accuracy

In supervised classification techniques, a confusion matrix (Table 1), lists a model's predictive capability. The accuracy of the model is given by: correct predictions/total no. of predictions = (TP+TN)/(TP+FP+FN+TN). Here TP = true positive, FP = false positive, FN = false negative and TN =true negative.

Table 1. Confusion Matrix

	Actual Class	
Predicted Class	True Class	False Class
True Class	TP	FP
False Class	FN	TN

Even though highest accuracy of a model can be 100%, practically 70% accurate model is usable for many business domains. For unsupervised cluster analysis there is no right objective measure for accuracy.

Accuracy of information retrieval and information extraction are evaluated with two measures named precision and recall. Precision is the measure of the ratio of relevant documents retrieved, divided by number of documents retrieved by the algorithm i.e. (TP/(TP+FP)). Recall is the measure of the ratio of relevant documents retrieved divided by number of relevant documents out there in the corpus (TP/(TP+FN)). Even though all data mining algorithms aim to increase precision and recall both, that is not possible. As when the precision measure is high the recall measure is low and vice versa. There is another statistical measure called *F*-score to measure the accuracy of the algorithm, which is the harmonic mean of precision and recall.

F-Score = Precision*Recall/(precision+recall)/2

Binary classification models can also be compared against each other with Receiver˝s Operating Characteristic (ROC) curve (Figure 2). ROC is a plot of true positive rate (TPR) or sensitivity of the model along *y* axis vs. false positive rate (FPR) or (1 – specificity) of the model along *x* axis. If TPR increases FPR also increases. All data mining algorithms aim to increase the area under the curve (AUC) of ROC.

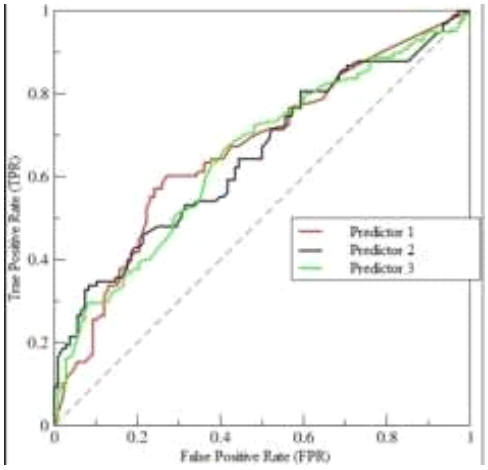

Figure 2: Receiver's Operating Characteristics (ROC)

As we can see, text mining field is vast, involving many sub-areas and sub-tasks. It needs interdisciplinary approach requiring the skills of databases, data mining, statistics, linguistics, computer science, machine learning and artificial intelligence.

2. LITERATURE SURVEY: TEXT MINING APPLICATIONS

Text mining has many interesting and diverse range of applications for business intelligence (BI) extraction. Businesses collect a lot of data about their current and potential customers, customer behavior, product trends, suppliers, business partners and competitors' behavior. This huge data is in the form of structured, semi structure or unstructured format collected from databases, emails, blogs and social communication forums. This data need to be mined effectively for BI extractions. Data cleansing and integration of data in a unified format is the essential preprocessing steps for extracting useful information from it. Below is a list of interesting and important application of text mining and big data analytics found in the literature.

Customer Relationship Management (CRM): Businesses depend on their customers. Thus product design, marketing and customer service should be such that facilitates customer satisfaction, customer retention and new customer development. All these functions are collectively known as customer relationship management. To achieve these, companies must generate knowledge from social media text data [13].

In reference [14], the authors have reviewed several data mining functions used for CRM. These functions are mainly association rules mining, clustering, classification, regression, sequence, prediction and visualization. Association rules determine the coexistence of items in sales database that sells together. Association rule mining is also called market basket analysis which helps in cross selling of products.

Classification and clustering are used for market segmentation so that the whole customer population can be segmented into several groups to treat them differently as per need such as new customer identificationion, one-to-one marketing, and loyalty program etc. Regression model technique is also used for CRM. Based on existing data a regression model is built which will minimize error in predicting and forecasting future market analysis. Artificial neural network (ANN) is also used for predicting more accurately. Visualization of data is very important in CRM

in understanding complex relationship between data.

In reference [15-16] they have studied sentiment analysis and opinion mining from text and how it can be used for forecasting purposes. Sentiment analysis from social community forum blogs can reveal customers' opinion or emotion polarity about products and services. Sentiment analysis can be done to classify text opinion as positive, negative or neutral. Businesses can act accordingly to alter product design or can adopt a different marketing strategy. In the above reference they have used an automatic java crawling program first to download all the blogs from *Sina* (Chinese) sports community forum posting. Then after cleansing the text, they have determined the sentiment value of each word/term as positive (+1), negative (-1) or neutral (0) from a sentiment dictionary lookup. Aggregated value of all the words gives an integer value for each text. The sign of the number gives polarity (+ve or–ve) of sentiment and magnitude gives intensity of the sentiment. Then they have used *K*-means algorithm to cluster the blogs and SVM classification for prediction.

Spam Filtering: Unsolicited bulk emails (UBE) or unsolicited commercial emails (UCE) are called spam. Spammers collect recipients' email address from public sources or generate them synthetically to spread their business advertisement fast, without incurring any cost. Ninety percent of the emails we receive every day are spam emails. Even if a very small percent of these recipients response to the spam advertisement, spammers can make a profit. Spam is a major problem in the era of email communication both for recipients and for the internet service providers. Spam mails are unwanted, consume space, bandwidth and spread virus as well. Spam email can be blocked at the senders' side before they spread through the network wasting valuable communication resources. Spam can also be blocked or flagged as spam at the recipients' side.

Spam detection is a classification problem to classify an email as spam or not and block them accordingly. Wilfried [17], investigated the Spam detection problem in detail. Document features are extracted from email header, body and sender's email address. Naïve Bayes (NB) classifier is proven to be good for email classification. NB classifier computes a probability score to test an email for it to be a spam or not, depending on a model built from training data set. NB can be used for large dataset without sacrificing on speed and accuracy. Even then Latent Semantic Indexing (LSI) is used for dimensionality reduction in the pre-processing stage here also. LSI is effective in concept grouping and proves to significantly improve the efficiency of spam detection problem.

Fraud Detection Applications: Fraud detection is very important for industry and government to protect businesses [18-19]. Modern day fraudsters are using sophisticated and organized techniques to deceive businesses. E-commerce, retail, credit card, insurance, telecommunication often suffer from illegal transactions and claims. A company loses 5% of their yearly revenue because of fraud. That amounts to 3.5 trillion dollars globally [18]. Manual auditing of papers to detect some anomaly and fraud usually takes one to two years after the fraud has taken place. Manual investigation is also cost-prohibitive when businesses deal with millions of customers/parties and millions of transactions. Automated fraud detection system can build predictive model for anticipating probability of fraudulent behavior, anomaly detection and intrusion detection at low cost. This automated system can prevent fraud early and minimize loss.

One thing to note here is that in fraud detection false negatives (FN) are more damaging than false positives (FP). Clifton [20], has presented a comprehensive study of all data mining algorithms that have been applied so far for fraud detection. Supervised techniques such as decision trees, information gain, regression, neural network, Bayesian network and support vector machine (SVM) can be applied when there are previous examples of fraud instances. Some other techniques are genetic algorithm, association rules and expert system. Unsupervised technique such as *K*-means clustering algorithm can be used first to label the insurance/claim data before applying supervised techniques.

So far till 2012, only standard data mining algorithms were used to process structured data for fraud detection. But so much additional information from different sources like, email, blog, internet, social networking, public records, phone records, bank documents, investigator's interview etc. can be integrated for the benefit of fraud detection. Continual text mining can be applied for automated fraud detection at low cost. A well-integrated profile generation of fraudster can improve prediction accuracy, detect anomaly early and prevent fraud better.

Health and Wellness: Our healthcare system can be improved tremendously if automated text mining technique is adopted. Human doctors have some limitation as they cannot keep up to date with all relevant medical journals and come to a precise decision instantly. Text mining will facilitate information synthesis from various sources like medical journals, patient's admission note, lab reports etc. New knowledge can be discovered from this integrated information that will enhance clinical decision. Text mining tools will provide preventive care as well as treatment decision. Automated diagnosis will reduce cost and error and can act as a complimentary aid to a physical doctor.

National center for text mining, NaCTeM [21] is the first publicly funded text mining center founded in UK to make text mining an effective tool in the domain of biology, science and social science. They have developed several tools that extract and identify biological named entities, semantic relations between them and summarize the concept of the documents. These tools using natural language processing and advanced search techniques can also link the named entities to external knowledge sources, can associate symptom with diseases, associate gene with disease, and associate disease with disease. Business's like Pfizer is using NaCTeM¨s text mining tools.

3. EXPERIMENTAL RESULTS

In this section I have conducted some experiments that show text summarization results by different techniques. Text summarization is a very important and interesting application of text mining that applies machine learning techniques. Automatic text summarization can be used to determine topic of the document, document length, word frequency, grammatical and syntactic correctness, vocabulary and reading ease. Automatic text summarization is very essential as it reduces reading time to glimpse through the content of the text or to judge the quality of the text or to understand the text category and topic. Text summarization needs natural language processing capabilities by machines.

Figure 3: Word Cloud for this Document

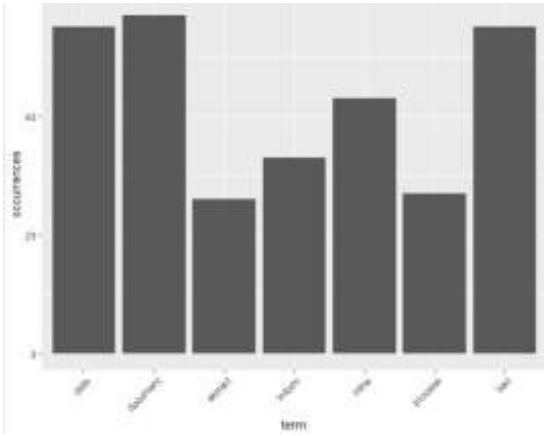

Figure 4: Histogram of Word Count for this Document

In Figure 3, the summary of this research paper is shown in a word cloud based on frequency counts of words. Higher frequency words are larger in font than the lower frequency words. It helps in understanding very quickly that the paper is about text data mining without reading the whole paper. Figure 4 shows the corresponding histogram in traditional data visualization technique. The word cloud in Figure 3 is more effective visualization tool for text data.

In text summarization, automatically finding the readability score of a document is very essential. Readability score of a text is calculated from word count, paragraph count, sentence count, average word length, syllables per word, words per sentence and paragraph length etc. Among many other readability scores one measure is called Flesch Reading ease score developed by Flesch in 1948. Another measure is Flesch –Kincaid Grade Level score. Figure 5, generated by Microsoft Word shows a few text summarization and readability score measures including Flesch Reading Ease and Flesch-Kincaid Grade Level for this research paper. This research paper is showing a score of 34.3 and 12.8 respectively for the above two measures. Other research papers on the same topic have almost similar score.

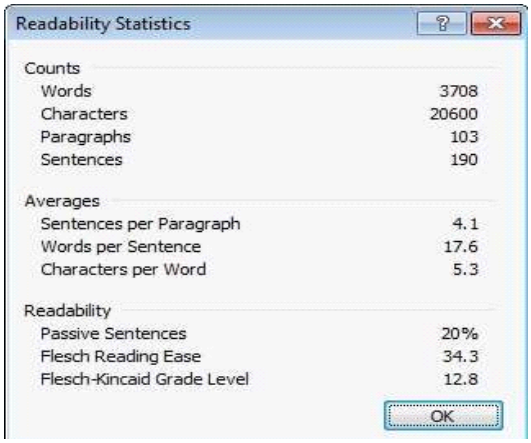

Figure 5: Readability Score of this Document

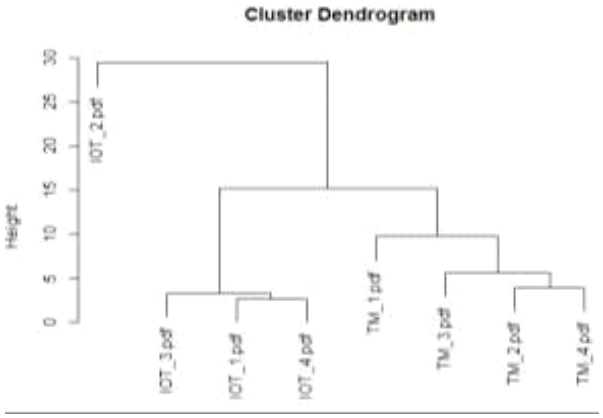

Figure 6: Document Clustering

Flesch Reading Ease measure varies from 0 to 100, 0 being hardest and 100 being easiest reading. This score is slightly affected by domain specific terminology and language instead of giving true measure of complexity of text. Thus technical and scientific papers will have slightly lower Flesch Reading Ease score. But it is very effective if large scale text documents on similar topics need to be compared automatically. For example students' essay writing on similar topic can be graded initially by computers and then again can be rechecked by a teacher manually. Here a computers prediction can complement a teachers' evaluation. Many other analytics can be performed on readability score such as authors can be linked with readability score of documents etc.

Clustering, an unsupervised data mining technique also is very essential in automatic text summarization where the concept or topic of the document can be known without reading the document manually. In Figure 6, document clustering experiment result is shown. In this experiment there were four research papers on text mining. This research paper is one of them. There were four other documents on the topic of Internet of Things (IOT). In this text corpus I have named the internet of things research papers as "IOT_1.pdf", "IOT_2.pdf" …etc. Similarly I have named the text mining research papers as "TM_1.pdf", "TM_2.pdf" …etc. These eight documents were not labeled beforehand. The unsupervised clustering algorithm written in *R* programming language uses Manhattan distance between documents and has produced this clustering result as shown in Figure 6. It was reasonably accurate to produce this result automatically. For very dissimilar topics like sports, politics and cooking the result must have been more accurate.

Document classification is a supervised learning technique where a model is trained with previously labeled data. If in the text corpus I have labeled the internet of things research papers with "IOT' and text mining research papers with "TM" and build a classification model then subsequent un-labeled test data can be classified automatically. Classification models for text data are many such as K Nearest Neighbors (KNN), Naïve Bayes classifier, support vector machine (SVM) classier etc. SVM is the most accurate text data classifier even though it is computationally more costly. SVM is a powerful technique for high dimension data classification, regression and outlier detection. In my experiment I have compared KNN algorithm with SVM algorithm for a larger corpus of 100 research papers on two different topics i.e. i) text mining (50 documents) and ii) IOT (50 documents). The test results with various measures of accuracy mentioned before are shown in Table 2 and Table 3.

Table 2. SVM Classification

Predicted Class	**Actual Class**		
	True Class	**False Class**	
True Class	**TP=45**	**FP=15**	**Precision=75%**
False Class	**FN=5**	**TN=35**	
	Sensitivity/Recall=90%		**Accuracy=80%**

Table 3. KNN Classification

Predicted Class	**Actual Class**		
	True Class	**False Class**	
True Class	**TP=39**	**FP=19**	**Precision=67%**
False Class	**FN=11**	**TN=31**	
	Sensitivity/Recall=78%		**Accuracy=70%**

4. CONCLUSION

In this research paper I explored the emerging and exciting text mining field. There are many text mining applications that can generate profit by extracting business intelligence out of text data. There are also many challenges in text mining which are explored in this paper. Businesses have just started using text mining tools to get a competitive edge. In order these tools to be more effective the challenges of text mining need to be researched and solved. I have also shown results of some of the text summarization experiments using open source *R* programming language which serves as a tutorial for the researchers in the text mining field. It indicates how computers can solve many of the text mining problems so easily, and quickly what human cannot do manually. Design, simulation, implementation and testing of text mining sub tasks will motivate students to do research on this emerging field. This paper was written out of my own interest in text mining research and the graduate level course that I am teaching on big data analytics. This paper is basically a survey paper and text mining algorithms have not been numerically compared against each other, which I am going to present in my future publications.

5. REFERENCES

[1] G. Miner et al. 2012. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Elsevier.

[2] Anil Maheswari. 2017. Data Analytics. McGraw Hill Education (India) Private Limited.

[3] Chakrabarti Soumen. 2002. Mining the Web: Analysis of Hypertext and Semi-Structured Data. Morgan Kaufmann, San Francisco.

[4] Manning et al. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

[5] Nisbet et al. 2009. Handbook of Statistical Analysis and Data Mining Applications. Elsevier, Burlington, MA.

[6] S. Valenti et al. 2003. An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Information, Vol. 2.

[7] Escudeiro et al. 2011. Semi-Automatic Grading of Students' Answers Written in Free Text. The Electronic Journal of e-Learning Vol. 9.

[8] Michael W. Berry et al. 2007. Survey of Text Mining: Clustering, Classification and Retrieval. Springer, Second Edition.

[9] Croft Bruce et al. 2009. Search Engines: Information Retrieval in Practice. Addison-Wesley, Boston, MA.

[10] Manning et al. 2008. Introduction to Information Retrieval. Cambridge University Press, New York.

[11] Radha Guha. 2013. Impact of Semantic Web and Cloud Computing Platform on Software Engineering. Software Engineering Frameworks for the Cloud Computing Paradigm, Computer Communications and Networks, Springer-Verlag-London.

[12] Vignesh Prajapati. 2013. Big Data Analytics with R and Hadoop. PACKT Publishing.

[13] Sergio-Orenga Rogla. 2016. Social Customer Relationship Management: Taking Advantage of Web2.0 and Big Data Technologies. Springer Plus. DOI 10.1186/s40064-016-3128-y.

[14] E.W.T Ngai et al.. 2009. Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. A transaction on Elsevier Journal on Expert System and its Applications, 2592-2602.

[15] Nan Li et al. 2012. Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast. Elsevier.

[16] G. Vinidhini et al. 2012. Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research on Computer Science and Software Engineering, Vol. 2.

[17] Wilfried N. Gansterer et al. 2007. Spam Filtering Based on Latent Semantic Indexing. Springer.

[18] Greg Handersion et al. 2007. SAS, an Enterprise Approach to Fraud Detection and Prevention in Government Programs. SAS.

[19] ACFE 2016: A report to the Nation on Occupational Fraud and Abuses. ACFE.

[20] Clifton Phua et al. 2004. A Comprehensive Survey of Data Mining-based Fraud Detection Research.

[21] NaCTeM. 2016. Providing Text Mining Services to UK, www.nactem.ac.uk.