Application of Text Mining in Effective Document Analysis: Advantages, Challenges, Techniques and Tools

Amreen Kausar Gorvankolla
PG Student
Department of Information Technology
RV College of Engineering and Technology
Bengaluru, 560031

Rekha B.S
Assistant professor
Department of Information Science
RV College of Engineering and Technology
Bengaluru, 560031

Abstract: **Age of Big Data is leading to a drastic increase of digital information. Immeasurable amount of data and information are produced on regular basis through financial, education and social means. Manual analysis and extraction of useful information from such a huge data is a challenging task. Without the technological help, understanding the text contents and correlations is something which cannot be carried out solely by human mind. There is a need for text mining to extract the exact information which the user requires.**

Potentially valuable business insights and high-quality structured data can be derived by organizations from unstructured text using Text mining. Collection and extraction of additional information pertaining to customers from the unstructured data can be either to enrich customer master data, or in production of new customer insights or in determining product and services sentiments. For all these, text mining plays a vital role.

This survey paper provides information and brief idea on text mining, its advantages, applications and various text mining techniques that can be used for effective and efficient document analysis that in turn will provide information to build product roadmaps and make better decisions about their activities.

Keywords: Big Data, text mining, decision making, web crawlers, Information Extraction (IE), Information Retrieval (IR), Categorization, Summarization, Visualization, and Clustering, Natural Language Processing (NLP)

I. INTRODUCTION

Every business and every life on this planet will be revolutionized by the advent of Big Data. Big data is such a term that often describes tera, peta and exabytes of huge data that is increasing exponentially day by day. The value of big data does not depend on how much data one has, but it depends on what business insights we derive from it. Big data can be collected from any source and analyzed that in turn will help in cost reduction, time reduction, new product development, optimized offerings, and making decision smartly. When big data is combined with high-powered analytics, following business-related tasks can be established such as:

- Determination of failures and their root causes, issues and defects in near-real time.
- Based on the customer's buying habits coupons can be generated during the sale.
- Entire risk portfolios can be recalculated very quickly.
- Fraudulent behaviours can be easily detected before an organization is affected.

According to the survey conducted by Forbes, prediction is made that every second for every human being on this planet nearly 1.7 megabytes of information will be created by the year 2020. Also, the survey conducted by IBM Big Data and Analytics Hub reveals that only 23% of the organizations that have been assessed have enterprise-wide Big Data strategy. One of the statistics by TIME reveal that more than 30 petabytes of data solely generated by Facebook is stored, accessed and analyzed. Recently, National Centre for Text Mining has been established and it is the first publicly funded text mining clearinghouse in the world.

At this point it is clear that text mining helps in better understanding of components and ideas present in a text/phrase than a web crawler or search engine, like Google. Web crawlers just scan through the texts to find out the keywords but the importance of these words or their specific situation cannot be interpreted. The utilization of cutting edge ventures like text mining can help in overcoming such problems to a greater extent. Text mining empowers in distinguishing the patterns and connections among the words and makes it very simple in identifying the patters that can in one way or other be a great degree of troublesome or tedious to find. Seeing these statistics one can infer that there is a huge requirement for text mining techniques to dig deep into big data and get some significant insights out of it. The process of

Text mining is given by following steps:

- Information collection from unstructured data
- Conversion of collected data to structured form
- Pattern identification from structured data
- Pattern analysis
- Extraction and storage [1]

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

The above steps can be summarized in the following Figure 1.

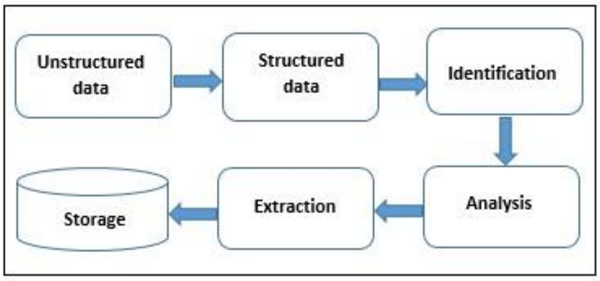


Figure 1: The steps involved in Text Mining

The paper is organized into different sections as follows; Challenges of text mining in section II, applications of text mining in section III, section IV consists of related work, section V contains the methods, text mining techniques in section VI, Tools in section VII and finally conclusion in VIII.

II. CHALLENGES OF TEXT MINING

Challenges and complexities can arise in any stage of text mining. Natural language processing is one of the major challenge in text mining. Multiple words can have same meaning or one word may give multiple meaning leading to ambiguity [2]. Although many researches have been done on eliminating the ambiguity but still it remains an open issue because the solution can’t be generalized as it is not easy to answer what semantics the user is actually looking for. Another challenge in text mining is the structure of the document. Various classification and categorization methods are being used to overcome structuring problem of the document [3].

III. APPLICATION OF TEXT MINING

Text mining plays a vital role in document analysis and provides various for analyzing the documents in today’s business world.

1 Digital libraries:
A large amount of patterns and trends can be derived from the journals and other documents present in the digital library that helps in research and development. Digital library is a significant source of information and collection of trillions of documents online. Text mining helps in accessing of these trillions of documents as quick as possible. Document extraction can be done in the form of audio, image format along with the text [4].

2 Healthcare and Life science:
Healthcare domain produce large amount of information related to patients records, diseases, medicines, treatment etc. Extraction of relevant text from large and complex medical terms is a challenging task. Text mining in medical field helps in evaluation and effectiveness of medical treatments that shows effective comparison between different diseases, symptoms and their course of treatments [4].

3 Sentiment Analysis:
Sentiment analysis deals with categorization of opinions expressed in textual documents. It is a way of identifying and extracting subjective information from unstructured data.

4 Business Intelligence:
Organizations and enterprises get a great platform to analyse and make better decisions about their competitors through text mining. Decision making in turn helps in analysis good and bad performance.

5 Analysing the open ended survey responses:
Usually in a survey, users are not restricted with a set of predefined statements. The views can be written by the users in any format and order. It aims at discovering a certain set of words or terms which the respondents commonly use to describe whether a product or service (under investigation) is valuable or defective, suggests misconceptions that are common or confusion regarding the items in the study.

6 Automatic filtering of undesirable messages, emails etc.:
Helps in automatic processing and filtering of undesirable messages and emails and retains only legitimate one. It also helps in routing the messages and emails to appropriate destination.

7 Competitor investigation by crawling through web sites:
Text mining quickly gathers the available documents from the website and lists down all the important terms and features that are described. Efficiency of delivering valuable business intelligence can be easily improved using these capabilities.

IV. RELATED WORK

Many papers have been surveyed that provide better understanding of the text mining techniques required for document analysis. It is summarized in the below Table 1.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Table 1: Literature survey of various papers and algorithms

Sl.No	Author Name(s)	Techniques/Tools discussed	Description
1	Shilpa Dang, et.al	IE, IR, Categorization, Clustering, Summarization	Several text mining techniques have been discussed and a comparison of every technique is made [1].
2	Sonali Vijay Gaikwad , et.al	IE, Visualization, Categorization, Clustering, Summarization, Term Based, Phrase Based, Concept Based, Pattern taxonomy	The paper presents techniques, methods and challenging issue in text mining. It also says that phrase based approach performs better and is less ambiguous [2].
3	Mrs. Sayantani Ghosh, et.al	IE, IR, NLP, Data Mining (DM), Variants of Clustering algorithm	The paper discusses text Mining Algorithms like Classification Algorithm, Association Algorithm, and Clustering Algorithm along with merits and demerits of the algorithms [3].
4	Ramzan Talib .etal	IE, IR, NLP, Clustering, Text Summarization	Discusses about various text mining techniques, applications and challenges [4].
5	Arvinder Kaur .etal	Proprietary Tools, Open Source Tools	The paper discusses about various text mining tools that can be applied for document analysis [5].
6	Abhilasha Singh Rathor, Dr. Pankaj Garg	IE, NLP, Topic tracking, Summarization, Categorization, Clustering, Association rule mining	The paper provides complete explanation about different text mining techniques like Topic Tracking, Opinion Mining, and Information Extraction [6].
7	Ms.J.Sathya Priya , et.al	Text clustering, text mining, Feature Extraction	Discusses about the clustering technique used for mining the text documents [7].
8	Ning Zhong, et.al	Pattern taxonomy, Closed taxonomy patterns, D-Pattern mining algorithm	The paper proposes an effective pattern discovery technique to overcome misinterpretation and low-frequency problem in text mining [8].
9	K.L.Sumathy, et.al	IR, Classification	This paper gives an overview of concepts, applications, issues and tools used for text mining like clustering, associative modelling etc. [9].
10	Sonali Vijay Gaikwad, et.al	Term Based Method (TBM), Phrase Based Method (PBM), Concept Based Method (CBM), Pattern Taxonomy Method (PTM), SPMining	The paper makes a conclusion that pattern based method outperform better than other methods [10].
11	Megha Rathi, et.al	Support Vector Machine (SVM), Decision Tree, Feature Selection, Classification and Prediction	The paper makes a comparison of different text mining techniques required for spam mail detection and concludes that tree like classifiers outperforms other algorithms [11].
12	Ronen Feldman, et.al	Deviation-Based approach, IR, Statistical Significance approach	The paper discusses the term level text mining of documents that using the concept of tagging the document [12].
13	Shivani Sharma, et.al	Naive Bayes and Bayesian Belief Networks, Neural Networks, Memory Based Reasoning, Decision Tree Based Methods	Discusses about the text classification that is used to find patterns and trends in twitter research [13].
14	YunYun Yang et.al	ClearForest, Goldfire Innovator, Quosa, RefViz	The paper provides an overview of text mining and visualization tools is presented in this paper to provide a comparison of text mining capabilities, perceived strengths, potential limitations, applicable data sources, and output of results, as applied to chemical, biological and patent information [14].
15	Aaron M. Cohen1, et.al	Information Storage and Retrieval, Text-Mining, Evidence-Based Medicine	The paper provides a comparison and explanation on various automated text mining tools [15].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

V. METHODS USED IN TEXT MINING

There are several text mining methods that are used for Text Mining and analysis. The following basic Text Mining methods have been discussed:

1. Term Based Method (TBM):
   In a document, term is referred to as a word with semantic meaning. Using this method, documents are analysed on basis of terms where in terms are picked and labelled. Machine learning and information retrieval widely adopts this method for text mining. The method has both advantages and disadvantages. Advantage is that it provides the most efficient computation of terms, whereas the disadvantage is that one term have multiple meanings or multiple terms have the same meaning. Hence it becomes challenging to derive the exact semantic meaning of the term that can answer what the user actually wants [2].

2. Phrase Based Method (PBM):
   Phrase is a collection of terms. Phrase Based Method analyses the document based on phrases which has more semantic meaning with less ambiguity.

3. Concept Based Method (CBM):
   Sentence and document level analysis is done using concept based method. It is based on statistical analysis where the term frequency is used for every word or phrase. Two terms can have same frequency, but the term which has the appropriate semantics should be given more importance. It contains three main components, where in semantic structure of the sentence is extracted by the first component, second component plots the conceptual ontological graph (COG) and the third component extracts the top concepts using the first two components [3]. Overall it helps in differentiating meaningful and less important words.
4. Pattern Taxonomy Method (PTM):
   It uses patterns to analyze the documents. Patterns can be extracted using different techniques like frequent item set, association rule, sequential pattern mining etc.

VI. TEXT MINING TECHNIQUES

Selection of appropriate text mining technique depends on the application and the requirement of user. There are a variety of techniques that are discussed in this section.

1. Information Extraction (IE):
   Extraction of meaningful information from a large volume of data deals with Information Extraction. It helps in extraction of attributes, entities from different documents and identifies the relation between them. The attributes can be then stored in a database for further analysis [4]. It mainly deals with the semantic information from the text like name of a person, location, organization name etc. [1].

2. Information Retrieval (IR):
   The best example for information based retrieval is Google and Yahoo search engines that search the documents on World Wide Web based on set of words. They provide most relevant and appropriate information according to the users' needs [4].

3. Categorization:
   In this technique, set of documents are sorted automatically into from a predefined set into different categories. It contributes to several applications, including scientific articles indexing, patents filing, spam filtering, document genre identification, authorship attribution, survey coding, and grading of automated essay. The main goal of categorization is to train classifier to classify the unknown terms on the basis of known terms.

4. Clustering:
   It helps in grouping of documents with similar content.

   It results in the formation of clusters, where each cluster contains one or more documents. The content of the documents within a document in a cluster is similar and the content of the documents between clusters is dissimilar [2]. Clustering is different from categorization because in clustering, the documents are clustered dynamically unlike using the predefined topics.

5. Summarization:
   Collection and representation of text documents comes under text summarization. It summarizes the details of the whole lengthy document into first paragraph of the document. It retains the important points and meaning of the documents even though the document size is reduced. It helps the user whether to read complete lengthy document or not [2].

6. Visualization:
   To represent different documents and the similarity text among them, different density colours are used. It arranges the data into visual hierarchy. The user is provided with the functionality of zooming in and out. It is very useful in identification of terrorist network and information about crimes [2].

The text mining techniques can differ from each other depending on the need i.e. conversion from structured to unstructured, decreasing the length of text etc., for example: IR deals with retrieval of data from unstructured text and IE deals with extraction of data from structured text. Summarization decreases the length of the documents by keeps the meaning of the document intact. Clustering identifies different patterns and forms them into different groups; Categorization divides the data into predefined topics. Overall, the techniques can be summarized as below in the Table 2 [1].

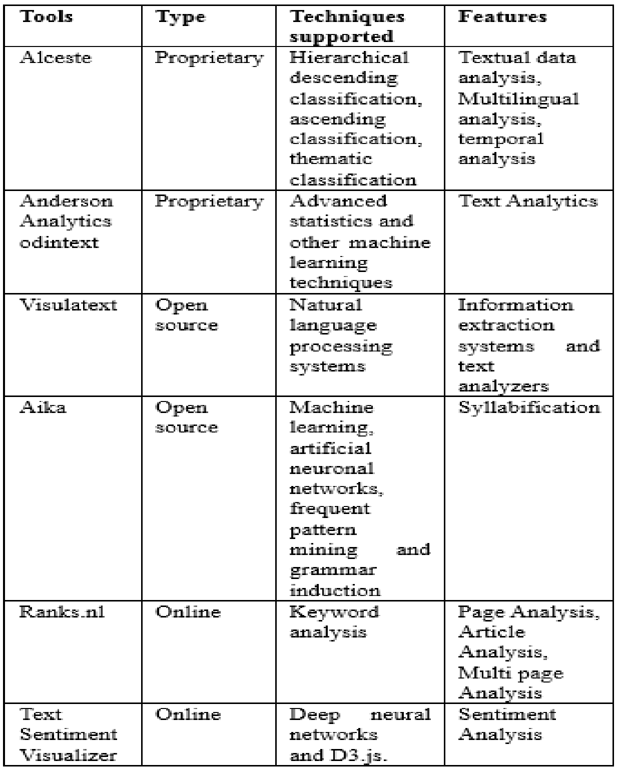
(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Table 2: Comparison of Text mining techniques

Technique	Characteristics	Tools
IR	Retrieves the data from unstructured text	Intelligent Miner, Text Analyst
IE	Extracts data from structured database	Text Finder, Clear Forest Text
Summarization	Reduces the length of document by keeping its meaning intact	Sentence Ext Tool
Categorization	Categorizes the document based on predefined topics	Intelligent Miner
Clustering	Groups the documents into different clusters based on the similarities in the content of the document	Carrot, Rapid Miner

VII. TEXT MINING TOOLS

Some of the popular and open source text mining tools have been discussed in this section. Text mining tools are classified into three categories, namely: *Proprietary tools, Open source tools and online tools.* Proprietary tools have to be purchased and are owned by the company. Open source tools are freely available, can access the source code and modify it for further development. Online tools run on the website and usually require just a browser. These online tools are provided with limited functionality. Some of the tools have been listed in the below Figure 2 [5]



Tools	Type	Techniques supported	Features
Alceste	Proprietary	Hierarchical descending classification, ascending classification, thematic classification	Textual data analysis, Multilingual analysis, temporal analysis
Anderson Analytics odintext	Proprietary	Advanced statistics and other machine learning techniques	Text Analytics
Visulatext	Open source	Natural language processing systems	Information extraction systems and text analyzers
Aika	Open source	Machine learning, artificial neuronal networks, frequent pattern mining and grammar induction	Syllabification
Ranks.nl	Online	Keyword analysis	Page Analysis, Article Analysis, Multi page Analysis
Text Sentiment Visualizer	Online	Deep neural networks and D3.js.	Sentiment Analysis

Figure 2: Tools used for Text Mining

VIII. CONCLUSION AND FUTURE WORK

Text mining is a technique that helps wide variety of users to find useful information from large amount of digital text documents on web or databases. It mainly deals with extracting non trivial information from unstructured data and is often known as Text Data Mining. In order to make the process of text mining easy and efficient, appropriate selection and usage of techniques and tools related to domain have to be done [4]. There are several techniques as discussed above and the application of appropriate technique depends on the developer or the researcher who uses it [6].

REFERENCES

[1] Shilpa Dang .etal, “Text Mining: Techniques and its Applications”, IJETI, Vol. 1 Issue 4, November 2014, ISSN (Online): 2348-0866.

[2] Sonali Vijay Gaikwad .etal, “Text Mining Mtehods and Techniques”, International Journal of Computer Applications (0975 – 8887), Vol 85 – No 17, January 2014.

[3] Mrs. Satyantani Ghosh .etal, “A tutorial review on Text Mining Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012.

[4] Ramzan Talib .etal, “Text Mining: Techniques, Applications and Issues”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[5] Arvinder Kaur .etal, “Comparison of Text Mining Tools”, IEEE, ISSN: 978-1-5090-1489-7, 2016.

[6] Abhilasha Singh Rathore .etal, “Analysis on Text Mining Techniques”, IJARCSSE, Volume 6, Issue 2, February 2016.

[7] Ms.J.Sathya Priya , et.al, “Clustering Technique in Data Mining for Text Documents”, International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012

[8] Ning Zhong ,et.al, “Effective Pattern Discovery for Text Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, January 2012

[9] K.L.Sumathy, et.al, “Text Mining: Concepts, Applications, Tools and Issues – An Overview”, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

[10] Sonali Vijay Gaikwad, et.al, “Performance Comparison For Text Mining Methods: Review”, International Journal of Advanced Engineering Research and Studies, December 2014

[11] Megha Rathi, et.al “Spam Mail Detection through Data Mining – A Comparative Performance Analysis”, I.J. Modern Education and Computer Science, December 2013

[12] Ronen Feldman, et.al, “Text Mining at the Term Level”, International Journal of Computer Science and Information Technologies, 2012

[13] Shivani Sharma, et.al, “Review on Text Mining Algorithms” , International Journal of Computer Applications Volume 134 – No.8, January 2016

[14] YunYun Yang et.al, “Text mining and visualization tools – Impressions of emerging capabilities”, Elsevier Ltd, 2008

[15] Aaron M. Cohen1, et.al, “Evidence-Based Medicine, the Essential Role of Systematic Reviews, and the Need for Automated Text Mining Tools”, IHI, November 2010

(This work is licensed under a Creative Commons Attribution 4.0 International License.)