IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

# Text Mining Techniques for Information Extraction: Issues and Applications

**Babita Verma[1*], Jyothi Pillai[2]**

[1,2]Department of Information Technology, Bhilai Institute of Technology, Durg, C.G., India

**Corresponding Author: babita.verma@bitdurg.ac.in*

**Available online at: www.ijcseonline.org**

Accepted: 19/Jan/2019, Published: 31/Jan/2019

***Abstract-*** Text mining research area has become very popular among researchers from various disciplines today. Text mining is one of the major areas of research for natural language documents. In this review, a comprehensive introduction and overview of text mining and existing research status is discussed. The major issue in text mining is the discovery of relevant information and patterns that are used to analyze text documents from the huge volume of information available over the internet. A number of tools and numerous methods exist for determining the relevant text and identifying valuable information for future research analysis and decision making.

The correct and effective methods and tools for text mining helps in speed up the extraction of valuable information and it also decreases the effort and time required for the analysis. The paper describes the methods, applications and issues of text mining in various fields of life. These results based on the text mining information from the various cited research articles and publications will be very useful for the researchers working in this research area. In addition, various issues related to text mining are identified that affect the accuracy and relevancy of results.

***Keywords-*** Text mining, Information extraction, Information Retrieval, Applications, NLP.

## I. INTRODUCTION

In the contemporary world text is the most widely recognized means for trading data[1]. Text is the most widely used means of data storage[2]. Text Mining is the revelation of new, already obscure data, by computer, via consequently separating data from various composed assets[3]. The information discovered can be in any of the following structures (i) structured (ii) semi structured and (iii) unstructured.The most common example of structured dataset is data storage in databases. The examples for semi structured and unstructured data sets incorporate emails, full text documents and HTML files and so on. In today's world, large amount of data is usually stored in text database rather than structured database. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents[1]. Text is drawn from a number of sources in research processes and innovative studies. Qualitative approaches for text analysis such as manual coding, discourse analytical methods, or grounded theory have been used by research scholars[4]. These manual strategies, nonetheless, are time and labour consuming and appear to have reached their natural limits with regards to investigating progressively a lot of text material[5,6]. Consequently, research scholars have begun to look for computer-aided, or automated text analysis methods[7,8]. The process of retrieving useful information from unstructured text is highly complex as this process involves specific processing methods and algorithms. As the most probable type of data storage is text, text mining is considered to have a higher worth than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing[1]. This paper mainly focuses on framework of text mining, area of its applications and issues regarding them.

## II. TEXT MINING FRAMEWORK

Text Mining can be pictured as consisting of two phases: (i) Text refining and (ii) Knowledge distillation. Text refining phase converts the free form unstructured text documents into a chosen intermediate form. Knowledge distillation construes patterns or knowledge from the intermediate form. The Intermediate Form (IF) can be semi structured such as the conceptual graph representation or structured such as relational data representation. When a document-based IF is mined, it procures similarities and relationships between documents. Document clustering/visualization and categorization are examples of mining from a document based IF[1,2].

 

Unstructured text documents are converted into an intermediate form (IF) using Text mining. IF can be either document-based or concept-based. In document-based IF each object will symbolize a document whereas in a concept-based IF each object symbolizes a concept in a particular domain. Knowledge distillation from a document-based IF infers patterns or knowledge from documents. A document-based IF can be forecasted onto a concept-based IF by drawing out object information pertinent to a domain. Knowledge distillation from a concept-based IF infers patterns or knowledge from objects or concepts[2]. Text mining framework is shown below in Figure 1.

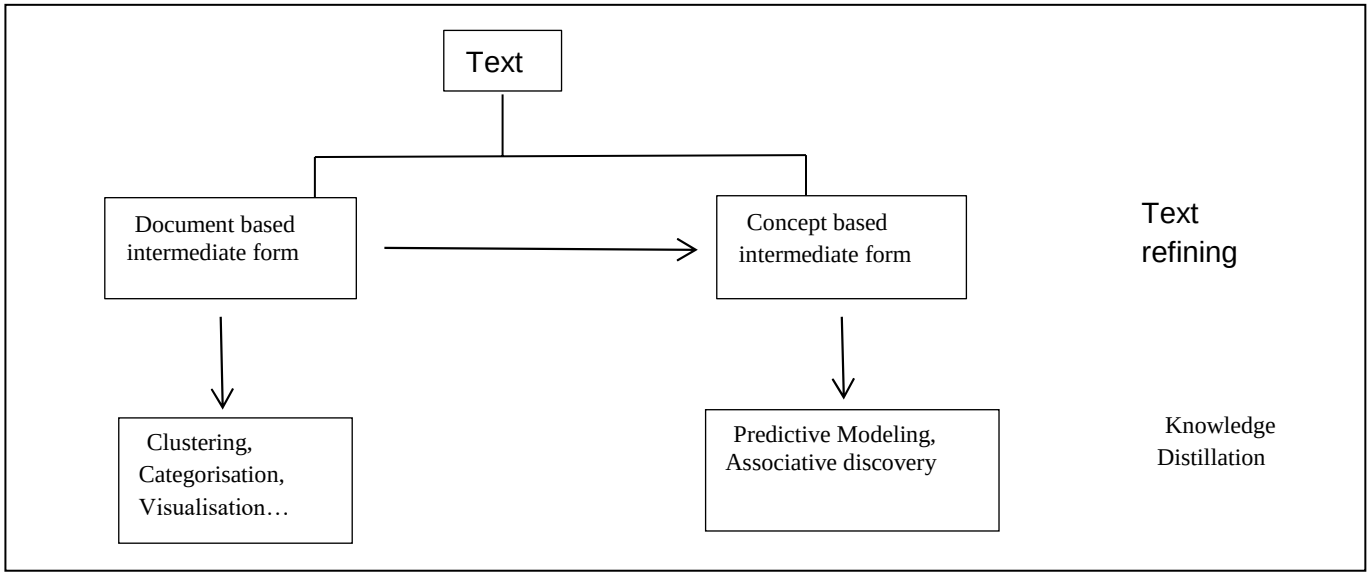


Figure 1: Text mining framework

In today's business environment, information is the most useful asset. Companies obtain unstructured and structured text of various origins on a daily basis. This data is received as raw text which is difficult to analyze without the usage of text analysis tools[2]. The intermediate forms used by some of the companies are listed below in Table 1:

Table 1: Intermediate forms used by the companies

| **Company/ Organization** | **Product/ Application** | **Text Refining Functions** | **Intermediate Form** | **Knowledge Distillation Functions** |
|---|---|---|---|---|
| Cartia | ThemeScape | | Document-based | Clustering Visualization |
| Canis | cMap | | Document-based word histograms | Clustering Visualization |
| IBM/ Synthema | Technology Watch | | Document-based | Clustering Visualization |
| Inxight | VisControls | | Document-based Hyperbolic Tree | Visualization |
| Semio Corp | Semio Map | | Concept-based | Visualization |
| Knowledge Discovery System | Concept Explorer | Info Retrieval | Concept-based | |
| Inxight | Linguist | Info Retrieval, text analysis, summarization | Document-based | |
| IBM | iMiner | Info Retrieval, summarization | Document-based | Clustering Visualization |
| TextWise | DR_LINK CINDOR CHESS | Info Retrieval, Info extraction | Concept-based | |
| Cambio | Data Junction | Info extraction | Concept-based | |
| Megaputer | Text Analysis | Info Retrieval, summarization | Document-based semantic net | Classification |

 

## III. TYPES OF TEXT MINING TOOLS

Text Mining Tools can be classified into three categories as shown

(1) Proprietary Text Mining Tools: These tools are commercial text mining tools owned by a company. To use these tools purchase is required. Although demo/trial versions are available free of cost but have limited functionality.
(2) Open Source Text Mining Tools: These tools are available free of cost along with their source code and one can even contribute in their development.
(3) Online Text Mining Tools: These tools can be run from the website itself. Only a web browser is required. These tools are generally simple and provide limited functionality[9].

The most commonly used tools for text mining are shown in Table 2

Table 2: Tools used for text mining

| Tool | Type | Techniques supported | Features/Uses | Website | Additional Remarks |
|---|---|---|---|---|---|
| Ranks.nl | Online | Keyword Analysis | Page analysis, Article analysis, Multi page analysis | http://www.ranks.nl/ | Website has been put together using Perl, Mysql, Javascript and HTML. Input Supported: Text/URL |
| Text Sentiment Visualizer | Online | Deep neural networks and D3.js | Sentiment analysis | http://sentiment.lucas/ | Input Supported: Text/URL |
| Textalyser | Online | Text analysis, Keyword analysis | Text analysis | http://textalyser.net/ | Input Supported: Text/URL |
| Alceste | Proprietary | Hierarchical descending classification, ascending classification, thematic classification | Textual data analysis, Multilingual analysis, temporal analysis | http://www.image-zafar.com/Logicieluk | OS required- Win XP, VISTA, 7, 8 et Mac OS-X |
| Anderson Analyics odin text | Proprietary | Advanced statistics and other ML techniques | Text analytics | http://odintext.com/# | |
| Ascribe | Proprietary | Hybrid Technology approach, NLP, ML and semi-automated coding tools | Text analytics | http://goascribe.com/ | |
| Basis Technology Rosette | Proprietary | Linguistic analysis, statistical modelling and machine learning | Text analytics, multilingual text analytics | http://www.rosette.c/ | Integrated with curl, Python, PHP, JAVA, C#, nodejs, Ruby |
| Aika | Open source | Machine learning, artificial neuronal networks, frequent | Syllabification | http://www.aika-software.org/ | Aika is implemented as a Java library. |

 

| | | | | | |
|---|---|---|---|---|---|
| | | pattern mining and grammar induction | | | |
| Data Science Toolkit | Open source | Advanced algorithms | Sentiment Analysis, Language Detection, Topic Classification | http://www.datascien/ | |
| Datumbox | Open source | Machine learning, keyword extraction | Text analysis, search engine optimization, social media monitoring, sentiment | http://www.datumbo/ | |

## IV. TEXT MINING TECHNIQUES

Some of the popular techniques for text mining are as follows[9]:

(1) Natural Language Processing and Machine Learning: Most of the tools employ Natural Language Processing (21%) and/or Machine Learning techniques (21%) for mining text.
(2) Statistical Methods: as used for data mining are also applied for text mining. In fact most of the tools use statistical methods (11%) in conjunction with other methods.
(3) Artificial Intelligence (9%): techniques such as neural networks are also employed in many text mining tools.
(4) Classification techniques (8%): are also used to categorize text and documents. These classification techniques must be able to handle unstructured data.
(5) Linguistic Learning (5%), Semantic Analysis (5%), and Predictive Modeling (7%) techniques are also employed for mining text.

These text mining techniques are shown bwlow in Figure 2.

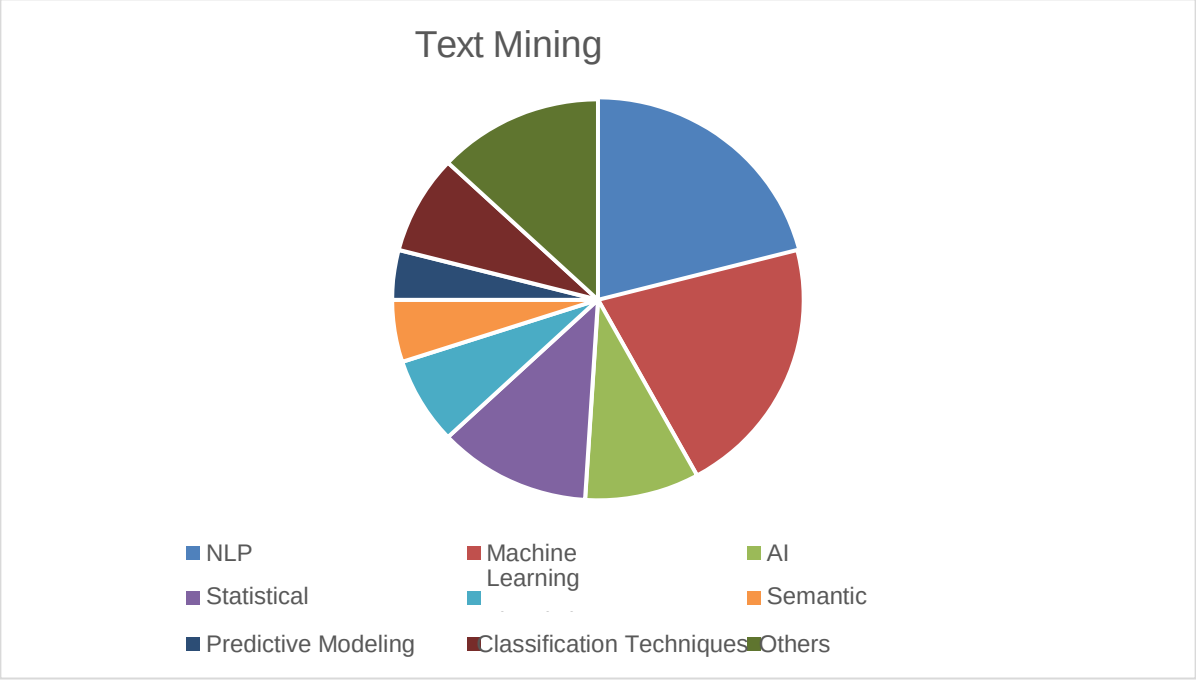


Figure 2 : Popular Text Mining Techniques

 

## V. APPLICATIONS

Text mining is a relatively new space of software engineering, and its utilization has developed as the unstructured information available keeps on expanding dramatically in both significance and amount. In the business world, this translates in being able to reveal insights, patterns and trends in even large volumes of unstructured data. In fact, it's this ability to push aside all of the non-relevant material and provide answers that is leading to its rapid adoption, especially in large organizations. Therefore text mining has wide range of applications, some of which are as follows:

- **Security applications**

Numerous text mining programming bundles are showcased for security applications, particularly observing and examination of online plain text sources like Internet news, online journals, and so on for public safety purposes. It is likewise engaged with the investigation of text encryption/decoding.

- **Digital Libraries**

Numerous text mining techniques and methods are in use to assess the patterns and trends of articles and proceedings from a vast number of repositories. These information sources relate to research and development. Digital library offers a new way of organizing data so that trillions of documents can be made available online. It also supports the processing of data, together with text documents, in audio visual, and image format. Various operations are carried out in the text mining process, such as document selection, enrichment, extracting information and addressing entities among documents and generating instinctive co-referencing and summarization.

- **Sentiment analysis**

Sentiment analysis might include examination of film surveys for assessing how great an audit is for a movie.Such an examination might require a marked informational index or naming of the affectivity of words. Text has been utilized to distinguish feelings in the related area of emotional computing.Text based ways to deal with full of feeling figuring have been utilized on various corpora like understudies assessments, youngsters stories and reports.

- **Academic and Research Field**

Different text mining tools and techniques are used in learning to analyse academic patterns in the particular region, preferences of students in particular areas and work ratios. Using text mining in research, articles and relevant material in various fields in one place can be found and classified.

- **Medical Science**

Medical science and health sectors produce statistical or written information on the history of patients, infections, medications, illness signs and diagnosis etc. The filtering of a relevant and appropriate text from a large biological repository is a major challenge. Text mining tools in the field of biomedicine provide the chance to gain useful information, to integrate it and to draw connections between various diseases, animals and genes. Text mining is used for discovery of biomarkers, the pharmaceutical industry, analytical clinical trade, preclinical safe studies of toxicity, patents competitive intelligence and landscape, genes mapping and targeted identification through different instruments.

- **Social Media**

For the analysis of applications in social media to monitor and analyze plain text online from Internet news, blogs, email etc, text mining software packages will be available. Text mining tools help to identify and analyze the number of social media posts, favorites and supports. This type of analysis reveals the reaction of people to various articles, media and how they spread.

- **Business Intelligence**

Text mining plays an important role in business intelligence, which lets companies and organizations make better decisions from their clients and rivals. This offers a deeper insight into industry and provides information on how customer satisfaction can be enhanced and competitive advantages gained

- **Risk Management**

No matter the industry, Insufficient risk analysis is often a leading cause of failure. This is especially true in the financial industry where adoption of Risk Management Software based on text mining technology can dramatically increase the ability to mitigate risk, enabling complete management of thousands of sources and petabytes of text documents, and providing the ability to link together information and be able to access the right information at the right time.

 

- **Cybercrime Prevention**

The anonymous nature of the internet and the many communication features operated through it contribute to the increased risk of internet-based crimes. Today, text mining intelligence and anti-crime applications are making internet crime prevention easier for any enterprise and law enforcement or intelligence agencies.

- **Contextual Advertising**

Digital advertising is a moderately new and growing field of application for text analytics. Compared to the traditional cookie-based approach, contextual advertising provides better accuracy, completely preserves the user's privacy.

- **Content Enrichment**

While it's true that working with text content still requires a bit of human effort, text analytics techniques make a significant difference when it comes to being able to more effectively manage large volumes of information. Text mining techniques enrich content, providing a scalable layer to tag, organize and summarize the available content that makes it suitable for a variety of purposes.

- **Spam Filtering**

Email is an effective, fast and reasonably cheap way to communicate, but it comes with a dark side: spam. Today, spam is a major issue for internet service providers, increasing their costs for service management and hardware software updating; for users, spam is an entry point for viruses and impacts productivity. Text mining techniques can be implemented to improve the effectiveness of statistical-based filtering methods.

## VI. ISSUES IN TEXT MINING

Many issues occur during the text mining process and effect the efficiency and effectiveness of decision making. Some of the issues in text mining are as follows:

- **Complexities at the intermediate stage**

Complexities can arise at the intermediate stage of text mining. In preprocessing stage various rules and regulations are defined to standardize the text that make text mining process efficient. Before applying pattern analysis on the document there is a need to convert unstructured data into intermediate form but at this stage mining process has its own complications. Sometimes real theme or data mislays its importance due to the modification in the text sequence[11].

- **Multilingual text refinement dependency**

Multilingual text refinement dependency is another big issue in text mining. Only few tools are available that support multiple languages[12].

- **Use of synonyms, polysems and antonyms/ Gramatical laws**

The use of synonyms, polysems and antonyms in the documents create problems (abstruseness) for the text mining tools that take both in the same context. It is difficult to categorize the documents when collection of document is large and generated from diverse fields having the same domain. Abbreviations gives changed meaning in different situation is also a big issue[13].

## VII. CONCLUSION

The analysis of a large volume of text based data to derive useful information is a tedious task. Text mining techniques are utilized to investigate the intriguing and pertinent data viably and effectively from huge amount of unstructured information. This paper presents a brief overview of text mining applications that help to improve the text mining process. Explicit examples and successions are applied to separate valuable data by disposing of superfluous subtleties for prescient examination. Selection and use of right techniques and tools according to the domain help to make the text mining process easy and efficient.

Domain knowledge integration, varying concepts granularity, multilingual text refinement, and natural language processing ambiguity are major issues and challenges that arise during text mining process. In future research work, we will focus to design algorithms which will help to resolve issues presented in this work.

 

## REFERENCES


[1]. Sumathy K.L. & Chidambaram M. Text Mining: Concepts, Applications, Tools and Issues – An Overview, *International Journal of Computer Applications (0975 – 8887)*, 80(4), 29-32, 2013.

[2]. Ah-Hwee Tan. Text Mining: The state of the art and the challenges. Procedings of the pakdd 1999 workshop on knowledge discover from advance data bases. 1999.

[3]. Gupta V. & Lehal Gurpreet S. A Survey of Text Mining Techniques and Applications, *Journal Of Emerging Technologies In Web Intelligence*, 1(1),60-76, 2009.

[4]. Duriau, V.J., Reger, K.R., & Pfarrer, M.D. A content analysis of the content analysis literature in organization studies: research themes, data sources, and methodological refinements. *Organizational Research Methods*, 10 (1), 5–34, 2007.

[5]. Jamiy, F.E., Daif, A., Azouazi, M., & Marzak, A. The potential and challenges of big data – recommendation systems next level application. *arXiv preprint arXiv*:1501.03424. 2015.

[6]. Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G., & Den Hartog, D.N. Text classification for organizational researchers: a tutorial. *Organizational Research Methods*, 21(3), 766–799, 2018.

[7]. Janasik, N., Honkela, T., & Bruun, H. Text mining in qualitative research: application of an unsupervised learning method. *Organizational Research Methods*, 12 (3), 436–460, 2009.

[8]. Wiedemann, G. Opening up to big data: computer- assisted analysis of textual data in social sciences.*Historical Social Research/Historische Sozialforschung*, 38(4), 332–357, 2013.

[9]. Arvinder Kaur & Deepti Chopra . Comparison of Text Mining. 5th Internaional conferenceon Relaibility,info com technology & optimization (Trends & Fture directions), 2016.

[10]. Henriksson A., Zhao J., Dalianis H., & Bostrom H. Ensembles of randomized trees using diverse distributed representations of clinical events, *BMC Medical Informatics and Decision Making*, 16 (2), 69-78, 2016.

[11]. Solanki H. Comparative study of data mining tools and analysis with unified data mining theory,*International Journal of Computer Applications*, 75(16), 2013.

[12]. Kaklauskas A., Seniut M., Amaratunga D., Lill I., Safonov A., Vatin N., Cerkauskas J., Jackute I., Kuzminske A., & Peciure L. Text analytics for android project, *Procedia Economics and Finance*, 18, 610–617, 2014.


 