Text Mining: Techniques, Applications and Issues

Ramzan Talib*, Muhammad Kashif Hanif†, Shaeela Ayesha‡, and Fakeeha Fatima§

Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyze the text documents from massive volume of data is a big issue. Text mining is a process of extracting interesting and non-trivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. This paper briefly discuss and analyze the text mining techniques and their applications in diverse fields of life. Moreover, the issues in the field of text mining that affect the accuracy and relevance of results are identified.**

***Keywords*—Classification; Knowledge Discovery; Applications; Information Extraction; Patterns**

I. Introduction

The size of data is increasing at exponential rates day by day. Almost all type of institutions, organizations, and business industries are storing their data electronically. A huge amount of text is flowing over the internet in the form of digital libraries, repositories, and other textual information such as blogs, social media network and e-mails [1]. It is challenging task to determine appropriate patterns and trends to extract valuable knowledge from this large volume of data [2]. Traditional data mining tools are incapable to handle textual data since it requires time and effort to extract information.

Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [3]. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics [3]. Figure 1 shows the Venn diagram of text mining and its interaction with other fields. Several text mining techniques like summarization, classification, clustering etc., can be applied to extract knowledge. Text mining deals with natural language text which is stored in semi-structured and unstructured format [4]. Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields [5]. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [6].

Generic process of text mining performs the following steps (Figure 2)

- Collecting unstructured data from different sources available in different file formats such as plain text, web pages, pdf files etc.
- Pre-processing and cleansing operations are performed to detect and remove anomalies. Cleansing process make sure to capture the real essence of text available and is performed to remove stop words stemming (process of identifying the root of certain word) and indexing the data [7].
- Processing and controlling operations are applied to audit and further clean the data set by automatic processing.
- Pattern analysis is implemented by Management Information System (MIS).
- Information processed in the above steps are used to extract valuable and relevant information for effective and timely decision making and trend analysis [8].

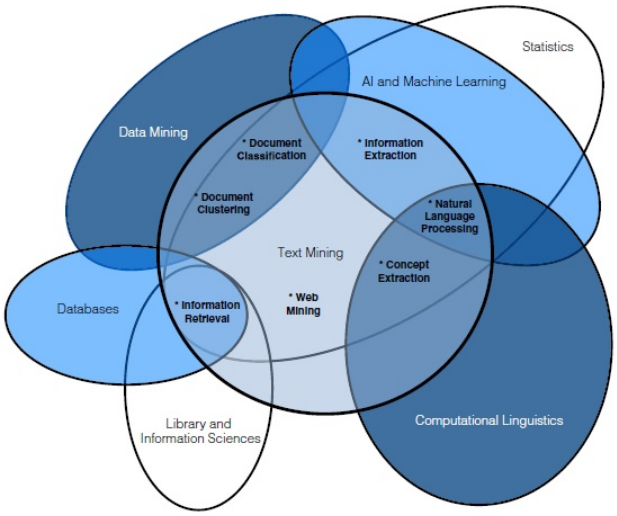

Fig. 1. Venn diagram of text mining interaction with other fields [4]

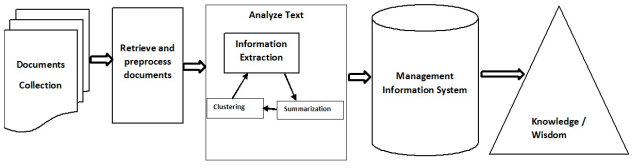

Fig. 2. Text mining process [5]

Extraction of valuable information from a corpus of different document is a tedious and tiresome task. The selection of

appropriate technique for mining text reduce the time and effort to find the relevant patterns for analysis and decision making. The objective of this paper is to analyze different text mining techniques which help to perform text analytics effectively and efficiently from large amount of data. Moreover, the issues that arise during text mining process are identified.

This paper is organized in different sections. Previous work is discussed in section II. In section III, different techniques of text mining are explained. Section IV presents the application areas of text mining techniques. In section V, issues and challenges in text mining field are highlighted. Section VI concludes the outcomes.

II. Review of Literature

[5] described that gathering, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In addition, different widely used text mining techniques, i.e., clustering, categorization, decision tree categorization, and their application in diverse fields are surveyed. [8] highlighted the issues in text mining applications and techniques. They discussed that dealing with unstructured text is difficult as compared to structured or tabular data using traditional mining tools and techniques. They have shown the applications of text mining process in bioinformatics, business intelligence and national security system. Natural language processing and entity recognition techniques has reduced the issues that occur during text mining process. However, there exist issues which need attention.

[9] explored MEDLINE biomedical database by integrating a framework for named entity recognition, classification of text, hypothesis generation and testing, relationship and synonym extraction, extract abbreviations. This new framework helps to eliminate unnecessary details and extract valuable information. [10] analyzed the text using text mining patterns and showed term based approaches cannot analyze synonyms and polysemy properly. Moreover, a prototype model was designed for specification of patterns in terms of assigning weight according to their distribution. This approach helps to enhance the efficiency of text mining process. [11] presented a crime detection system using text mining tools and relation discovery algorithm was designed to correlate the term with abbreviation.

[12] presented a top down and bottom up approach for web based text mining process. To combine the similar text documents, they apply k-mean clustering technique for bottom up partitioning. To find out the similarity within the document TF-IDF (Term Frequency- Inverse Document Frequency) algorithm has been used to find information regarding specific subjects. [13] gave an overview of applications, tools and issues arises to mine the text. They discussed that documents may be structured, semi structured or unstructured and extracting useful information is a tiresome task. They presented a generic framework for concept based mining which can be visualized as text refinement and knowledge distillation phases. The intermediate form of entity representation mining depends on specific domain.

[14] presented innovative and efficient pattern discovery techniques. They used the pattern evolving and discovering techniques to enhance the effectiveness of discovering relevant and appropriate information. They performed BM25 and vector support machine based filtering on router corpus volume 1 and text retrieval conference data to estimate the effectiveness of the suggested technique. [15] performed various experiments of classification using multi-word features on the text. They proposed a hand-crafted method to extract multi-word features from the data set. To classify and extract multi-word text they divide text into linear and nonlinear polynomial form in support of vector machine that improve the effectiveness of the extracted data.

III. The Reflective Process

Different text mining techniques are available that are applied for analyzing the text patterns and their mining process [16]. Figure 3 shows the Venn diagram for the inter-relationship among text mining techniques and their core functionality. Document classification (text classification, document standardization), information retrieval (keyword search / querying and indexing), document clustering (phrase clustering), natural language processing (spelling correction, lemmatization, grammatical parsing, and word sense disambiguation), information extraction (relationship extraction / link analysis), and web mining (web link analysis) [6].

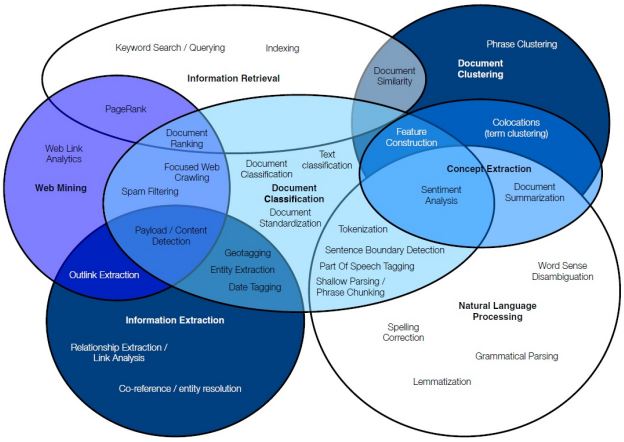

Fig. 3. Inter-relationship among different text mining techniques and their core functionalities [6]

A. Information Extraction

Information Extraction (IE) is a technique that extract meaningful information from large amount of text. Domain experts specify the attributes and relation according to the domain [17]. IE systems are used to extract specific attributes and entities from the document and establish their relationship [18]. The extracted corpus is stored into database for further processing. Precision and recall process is used to check and evaluate the relevance of results on the extracted data. In-depth and complete information about the relevant field is required to perform information extraction process to attain more relevant results [19].

B. Information Retrieval

Information Retrieval (IR) is a process of extracting relevant and associated patterns according to a given set of words

or phrases. There is a close relationship in text mining and information retrieval for textual data. In IR systems, different algorithms are used to track the user's behavior and search relevant data accordingly [19]. Google and Yahoo search engines are using information retrieval system more frequently to extract relevant documents according to a phrase on Web. These search engines use query based algorithms to track the trends and attain more significant results. These search engines provide user more relevant and appropriate information that satisfy them according to their needs [8].

C. Natural Language Processing

Natural language processing (NLP) concerns to the automatic processing and analysis of unstructured textual information. It perform different types of analysis such as Named Entity Recognition (NER) for abbreviation and their synonyms extraction to find the relationships among them [10]. NER identify all the instances of specified object from a group of documents. These entities and their instances allow the identification of relationship and other information to attain their key concept. However, this technique lacks complete dictionary list for all named entities used for identification [9], [10]. Complex query based algorithms need to be used to attain acceptable results. In real world, a single entity has numerous terms like TV and Television. Sometimes, a group of successive words have a multi-word names to identify the boundaries and resolve overlapping issues by using classification technique. Approaches to deal with NER usually fall into four categories: lexicon, rule, statistical based or mixture of these approached. NER systems have achieved the relevance level from 75 to 85 percent [20].

To extract synonym and abbreviation from textual data, co-referencing technique is frequently in use for NLP. Natural Languages (NL) have lot of complexities as a text extracted from different sources don't have identical words or abbreviation. There is a need to detect such issues and make rules for their uniform identification [21]. For example, NER and co-referencing approaches establish a logical relationship to extract and identify the role of person in an organization (use the name of a person at once and then use pronoun instead of name again and again) [22].

D. Clustering

Clustering is an unsupervised process to classify the text documents in groups by applying different clustering algorithms. In a cluster, similar terms or patterns are grouped extracted from various documents. Clustering is performed in top-down and bottom up manner. In NLP, various types of mining tools and techniques are applied for the analysis on unstructured text. Different techniques of clustering are hierarchical, distribution, density, centroid, and k-mean [22].

E. Text Summarization

Text summarization is a process of collecting and producing concise representation of original text documents [23]. Pre-processing and processing operations are performed on the raw text for summarization. Tokenization, stop word removal, and stemming methods are applied for pre-processing. Lexicon lists are generated at processing stage of text summarization. In past, automatic text summarization was performed on the basis of occurrence a certain word or phrase in document. Later on, additional methods of text mining were introduced with standard text mining process to improve the relevance and accuracy of results [11].

To summarize the text documents, weighted heuristics method extract features by following specific rules. Sentence length, fixed phrase, paragraph, thematic word, and upper case word identification features can be implemented and analyzed for text summerization. Text summarization techniques can be applied on multiple documents at the same time. Quality and type of classifiers depend on nature and theme of the text documents [24].

IV. Application of Text Mining

A. Digital Libraries

Numerous text mining techniques and tools are in use to ascertain the patterns and trends from journals and proceedings from immense amount of repositories. These sources of information help in the field of research and development. Libraries are a great source of information for the researchers and digital libraries are endeavoring to the significance of their collection. It provides a novel method of organizing information in such a way that make it possible to available trillions of documents online. It provides a novel way to organize information and make it possible to access millions of documents online. Green-stone international digital library that support multiple languages and multilingual interfaces provide a springy method for extracting documents that handle multiple formats, i.e., Microsoft word, pdf, postscript, HTML, scripting languages and e-mail messages [11]. It also supports the document extraction in the form of audio visual and image format along with text documents. In text mining process various operation are performed like documents selection, enrichment, extracting information and tackling entities among the documents and generating instinctive co-referencing and summarization [25]. GATE, Net Owl and Aylien are frequently used tools for text mining in digital libraries.

B. Academic and Research Field

In education field, various text mining tools and techniques are used to analyze the educational trends in specific region, student's interest in specific field and employment ratio [24]. Use of text mining in research field help to find and classify research papers and relevant material of different fields at one place. The use of k-means clustering and other techniques help to identify the attributes of relevant information. Students performance in different subjects can be accessed and how different attributes effect the selection of subjects [11], [26].

C. Life Science

Life science and health care industries are generating large amount of textual and numerical data regarding patients record, diseases, medicines, symptoms and treatments of diseases and many more. It is a big challenge to filter out an appropriate and relevant text to take a decision from a large biological repository [25]. The medical records contain varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process very difficult [27]. Text

mining tools in biomedical field provides an opportunity to extract valuable information, their association and inferring relationship among various diseases, species, and genes. Use of an appropriate text mining tools in medical field help to evaluate the effectiveness of medical treatments that show effectiveness by comparing different diseases, symptoms and their course of treatments [28]. Text mining use in biomarker discovery, pharmaceutical industry, clinical trade analysis, pre-clinical safe toxicity studies, patent competitive intelligence and landscaping, mapping of genes diseases and exploring the targeted identifications by using various tools [20].

D. Social Media

Text mining software packages are available for analyzing social media applications to monitor and analyze the online plain text from internet news, blogs, email etc. Text mining tools help to identify and analyze number of posts, likes and followers on the social media network. This kind of analysis show the people reaction on different posts, news and how it spread around. It shows the behavior of people belong to specific age group or communities having similarity and variation in views about the same post [29], [30].

E. Business Intelligence

Text mining plays a significant role in business intelligence that help organizations and enterprises to analyze their customers and competitors to take better decisions. It provides a deeper insight about business and give information how to improve the customer satisfaction and gain competitive advantages [31]. The text mining tools like IBM text analytics, Rapid miner, GATE help to take decisions about the organization that generate alerts about good and bad performance, market changeover that help to take remedial actions. It also helps in telecommunication industry, business and commerce applications and customer chain management system [32].

V. Issues in Text Mining Field

Many issues occur during the text mining process and effect the efficiency and effectiveness of decision making. Complexities can arise at the intermediate stage of text mining. In pre-processing 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. Sometime real theme or data mislay its importance due to the modification in the text sequence [27]. Another major issue is a multilingual text refinement dependency that create problems. Only few tools are available that support multiple languages [33]. Various algorithms and techniques are used independently to support multilingual text. Because numerous important documents persist outside the text mining process because various tools dont support them. These issues create a lots of problems in knowledge discovery and decision making process. Infect real benefit is difficult to attain by using the existing text mining techniques and tools because its rarely support multilingual documents [34].

Integration of domain knowledge is an important area as it performs specific operations on specified corpus and attain desired outcomes. In this situations domain knowledge from which document corpus to be extracted need to integrate with the computing abilities from which information have to be attained. According to the requirements of the field, experts are needed to work collaboratively from diverse domains to extract more effective, precise and accurate results [22], [27].

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 [35]. Varying concepts of granularity change the context of text according to the condition and domain knowledge. There is need to describe rules according to the field that will be used as a standard in the area and can be embedded in text mining tools as a plug-in. It entails lots of effort and time to develop and deploy plug-ins in all fields separately. To develop plug-ins in depth and proper knowledge about the specific domain will be required [34], [36]. Natural languages have lots of complications in itself that create problem in text refinement methods and the identification of entity relationship. Words having same spelling but give diverse meaning, for example, fly and fly. Text mining tools considered both as similar while one is verb and other is noun. Grammatical rules according to the nature and context is still an open issue in the field of text mining [36].

VI. Conclusion

The availability of huge volume of text based data need to be examined to extract valuable information. Text mining techniques are used to analyze the interesting and relevant information effectively and efficiently from large amount of unstructured data. This paper presents a brief overview of text mining techniques that help to improve the text mining process. Specific patterns and sequences are applied in order to extract useful information by eliminating irrelevant details for predictive analysis. 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] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, *International Journal of Computational Engineering Research*, vol. 2, no. 5, 2012.

[2] N. Padhy, D. Mishra, R. Panigrahi *et al.*, "The survey of data mining applications and feature scope," *arXiv preprint arXiv:1211.5723*, 2012.

[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, no. 9, pp. 76–82, 2006.

[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*. Springer Science and Business Media, 2010.

[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[6] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," *Computers in Human Behavior*, vol. 29, no. 1, pp. 90–102, 2013.

[7] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," *Copy at http://j.mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper*, vol. 456, 2014.

[8] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44, 2012.

[9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravičius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," *Journal of biomedical semantics*, vol. 5, no. 1, p. 1, 2014.

[10] B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," *International Journal of Research in Engineering and Technology*, vol. 2, no. 1, pp. 2321–2328, 2013.

[11] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[12] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," *International Journal of Computer Applications*, pp. 159–171, 2013.

[13] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," *International Journal of Computer Applications*, vol. 80, no. 4, 2013.

[14] P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on*. IEEE, 2015, pp. 634–638.

[15] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in *Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan*, vol. 51, 2007, p. 45.

[16] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[17] R. Agrawal and M. Batra, "A detailed study on text mining techniques," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp. 2231–2307, 2013.

[18] D. S. Dang and P. H. Ahmad, "A review of text mining techniques associated with various application areas," *International Journal of Science and Research (IJSR)*, vol. 4, no. 2, pp. 2461–2466, 2015.

[19] R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," *Language Resources and Evaluation*, vol. 46, no. 2, pp. 155–176, 2012.

[20] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[21] E. A. Calvillo, A. Padilla, J. Muñoz, J. Ponce, and J. T. Fernandez, "Searching research papers using clustering and text mining," in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on*. IEEE, 2013, pp. 78–81.

[22] B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," *IJSEAT*, vol. 3, no. 3, pp. 69–71, 2015.

[23] B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis based document summarization," *International Journal of Computer Science and Information Security*, vol. 14, no. 4, p. 145, 2016.

[24] R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." *Int. Arab J. e-Technol.*, vol. 1, no. 4, pp. 164–168, 2010.

[25] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library," *International Journal on Digital Libraries*, vol. 4, no. 1, pp. 56–59, 2004.

[26] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," *Europen Journal of Scientific Research*, vol. 43, no. 1, pp. 24–29, 2010.

[27] A. Henriksson, J. Zhao, H. Dalianis, and H. Boström, "Ensembles of randomized trees using diverse distributed representations of clinical events," *BMC Medical Informatics and Decision Making*, vol. 16, no. 2, p. 69, 2016.

[28] I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach," *Expert Systems with Applications*, vol. 44, pp. 386–399, 2016.

[29] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[30] Y. Zhao, "Analysing twitter data with text mining and social network analysis," in *Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013)*, 2013, p. 23.

[31] F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha, "Impact and usage of internet in education in pakistan," *European Journal of Scientific Research*, vol. 47, no. 2, pp. 256–264, 2010.

[32] R. Sharda and M. Henry, "Information extraction from interviews to obtain tacit knowledge: A text mining application," *AMCIS 2009 Proceedings*, p. 283, 2009.

[33] H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," *International Journal of Computer Applications*, vol. 75, no. 16, 2013.

[34] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information:-extended abstract," *OTT06*, vol. 1, p. 55, 2007.

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

[36] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in *Proceedings of the World Congress on Engineering*, vol. 3, 2013, pp. 3–5.