AJLLSE GLOBAL RESEARCH NETWORK

AMERICAN Journal of Language, Literacy and Learning in STEM Education
Volume 02, Issue 04, 2024 ISSN (E): 2993-2769

Text Mining Application of Computational Linguistics

Nadia Hameed Hassoon
Department of Media, College of Arts, University of Babylon, Iraq

Abstract: One of the foremost requirements of modern man is undoubtedly the possibility to access the desired information with minimal effort. And it is an undeniable fact that the amount of storing information digitally is increasing minute by minute.

As people are moving to a digital world of information or knowledge. Whenever one wishes to collect some information, the first task is to identify the relevant points from the vastness of digitally stored data. The upsurge and flourishing of Text Mining paved the way for a new beginning in the area of Information Extraction (IE) and Information Retrieval (IR). As the term suggests, it is to mine relevant information from the text document. The text document could be un-structured or semistructured. There exist different approaches and methods for text mining and most of these techniques are computational linguistics . The entitled paper attempts to provide a deliberation on the existing various concepts, techniques, pre-processing steps, applications and issues of text mining.

Keywords: Text Mining, Computational Linguistics, NLP.

1. 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 . Figure 1 shows the Venn diagram of text mining and its interaction with other fields. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics. 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 [3]. Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields . 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 .

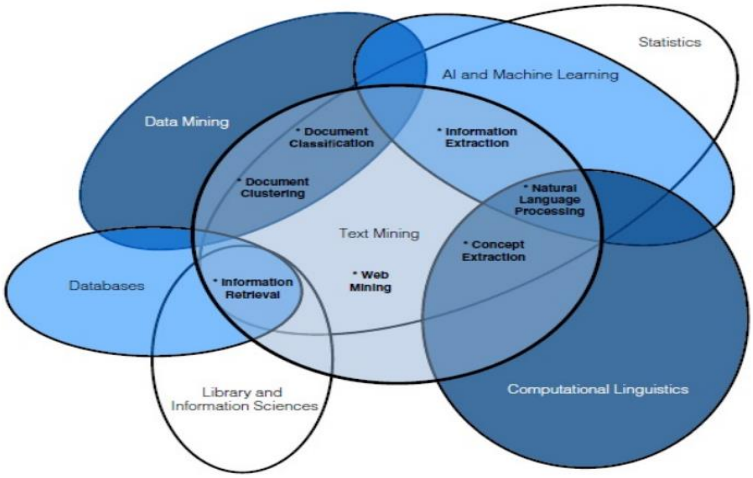

Fig. 1. Venn diagram of text mining interaction with other fields [4]

2. Problem Statement

For businesses, the large amount of data generated every day represents both an opportunity and a challenge. On the one side, data helps companies get smart insights on people's opinions about a product or service. Think about all the potential ideas that you could get from analyzing emails, product reviews, social media posts, customer feedback, support tickets, etc. On the other side how to process all this data. And that's where text mining plays a major role.

Text mining is being able to analyze complex and large sets of data in a simple, fast and effective way. At the same time, companies are taking advantage of this powerful tool to reduce some of their manual and repetitive tasks, saving their teams precious time and allowing customer support agents to focus on what they do best.

3. Generic process of text mining performs the following steps:

- 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 [4].
- 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 [5].

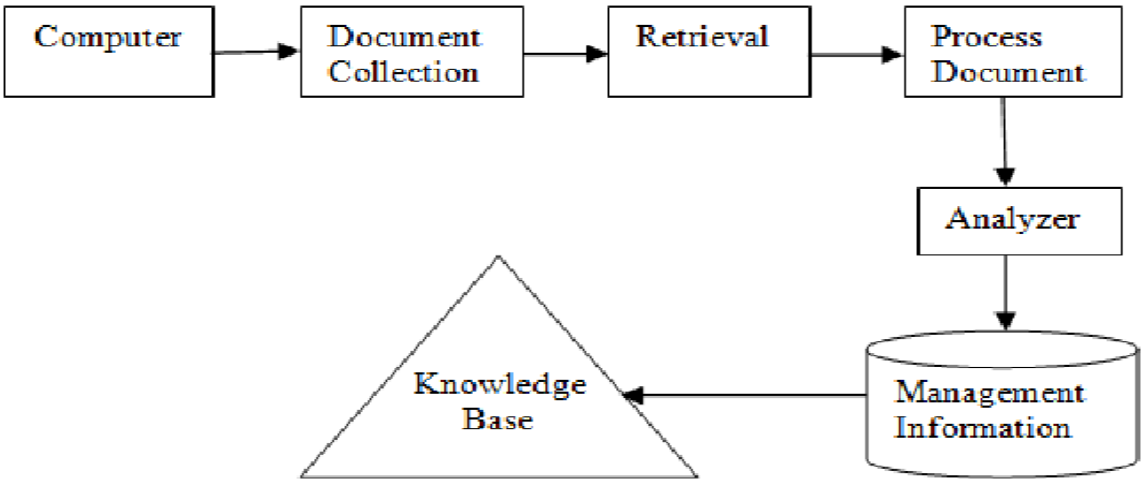

Fig. 2. Text Mining processing

4. The Reflective Process

Different text mining techniques are available that are applied for analyzing the text patterns and their mining process [6]. Figure 1 shows the Venn diagram for the interrelationship 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 , grammatical parsing, and word sense disambiguation), information extraction (relationship extraction / link analysis), and web mining (web link analysis) [7].

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 [8]. IE systems are used to extract specific attributes and entities from the document and establish their relationship . 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 [9].

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 [9]. 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 [5].

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 . 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 .

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 [11]. 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) [12].

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 [12].

E. Text Summarization

Text summarization is a process of collecting and producing concise representation of original text documents [13]. 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.

5. Filtering, Lemmatization and Stemming:

In order to reduce the size of the dictionary and thus the dimensionality of the description of documents within the collection, the set of words describing the documents can be reduced by filtering and lemmatization or stemming methods

Filtering methods remove words from the dictionary and thus from the document. The idea of stop word filtering is to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc. Furthermore, words that occur extremely often can be said to be of little information content to distinguish between documents, and also words that occur very seldom are likely to be of no particular statistical relevance and can be removed from the dictionary [14].

Lemmatization methods try to map verb forms to the infinite tense and nouns to the singular form. However, in order to achieve this, the word form has to be known, i.e. the part of speech of every word in the text document has to be assigned. Since this tagging process is usually quite time consuming and still error-prone, in practice frequently stemming methods are applied.

Stemming methods try to build the basic forms of words, i.e. strip the plural 's' from nouns, the 'ing' from verbs, or other affixes. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem. A well-known rule based stemming algorithm has been originally proposed by Porter [15]. He defined a set of production rules to iteratively transform (English) words into their stems.

6. Linguistic Preprocessing

Often text mining methods may be applied without further preprocessing. Sometimes, however, additional linguistic preprocessing may be used to enhance the available information about terms. For this, the following approaches are frequently applied [16].

Part-of-speech tagging (POS) determines the part of speech tag, e.g. noun, verb, adjective, etc. for each term.

Text chunking aims at grouping adjacent words in a sentence. An example of a chunk is the noun phrase "the current account deficit"

Word Sense Disambiguation (WSD) tries to resolve the ambiguity in the meaning of single words or phrases. An example is 'bank' which may have – among others – the senses 'financial institution' or the 'border of a river or lake'. Thus, instead of terms the specific meanings could be stored in the vector space representation. This leads to a bigger dictionary but considers the semantic of a term in the representation.

Parsing produces a full parse tree of a sentence. From the parse, we can find the relation of each word in the sentence to all the others, and typically also its function in the sentence (e.g. subject, object, etc.)

7. 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 [17]. 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.

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 [18]. 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 [17] .

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 [19]. The medical records contain varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process very difficult .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 .Text mining use in biomarker discovery, pharmaceutical industry, clinical trade analysis, preclinical 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 [21].

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 [22]. 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 .

8. Issues in Text Mining Field

There is growing demand from the various parts of industry, business, marketing, forecasting, language computing, internet related areas and academic community to enhance the research and developments in the field of text mining. The application areas of text mining like email filtering, search engines, fraud detection, customer` relationship management system, social media analysis, opinion mining, sentiment analysis and prediction, customer trend analysis and etc. [23]. For analyzing the final text using different text mining techniques like clustering, classification, summarization and etc. [24]. The text mining could be utilized effectively in the areas of opinion mining, sentiment, predictive, trend analysis, fraud detection, social media analytic etc. The main challenges of text mining could be seen as following:

1. one of the daunting tasks is to handle Multilingual documents/websites [25].
2. The second important challenge is handling Multimedia documents such as audio/video files.
3. To elicit data from the morphologically complex data.
4. The natural language is not free from ambiguity problem [25].
5. The issue of natural language ambiguity resolving [26].
6. The issue over mining hidden items from social network sites.
7. The difficulty on working over typologically complex languages such as agglutinative, inflectional etc. [26].

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. S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
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. E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernan- ˜ dez, "Searching research papers using clustering and text mining," in Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on. IEEE, 2013, pp. 78–81.
12. 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.
13. 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
14. Frakes, W. B. & Baeza-Yates, R. (1992). Information Retrieval: Data Structures & Algorithms. New Jersey: Prentice Hall.
15. Porter, M. (1980). An algorithm for suffix stripping. Program, 130–137.
16. Manning, C. D. & Schütze, H. (2001). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
17. 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.
18. R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." Int. Arab J. e-Technol., vol. 1, no. 4, pp. 164– 168, 2010.
19. 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.
20. 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.
21. 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.
22. 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.

23. Mrs. B. Meena Preethi, and Dr.P. Radha. A Survey Paper on Text Mining-Techniques, Applications and Issues IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP: 46-51.

24. Paramjit Kaur, and NeenaMadan Techniques of Text Mining: A Survey International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), ISSN (Print): 2279-0047. ISSN (Online): 2279-0055, PP. 265-267.

25. K.L.Sumathy, and M.Chidambaram Text Mining: Concepts, Applications, Tools and Issues An Overview International Journal of Computer Applications (0975 8887), Volume 80 No.4, PP. 29-32, October 2013.

26. Sonali Vijay Gaikwad, Archana Chaugule, and Pramod Patil Text Mining Methods and Techniques International Journal of Computer Applications (0975 8887), Volume 85 No 17, PP. 42-45, January 2014.