Proceedings of the SMART–2021, IEEE Conference ID: 52563
10th International Conference on System Modeling & Advancement in Research Trends, 10th–11th December, 2021
Faculty of Engineering & Computing Sciences, Teerthanker Mahaveer University, Moradabad, India

2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART) | 978-1-6654-3970-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMART52563.2021.9676285

A Review on Text Mining: Techniques, Applications and Issues

Preeti
Assistant Professor, University Institute of Computing, Chandigarh University, Gharuan
Email: preetigrovr@gmail.com

***Abstract:*— Due to the rapid rise of digital data collection techniques, a vast amount of data has become available. Unstructured and unsaturated data account for more than 85 percent of present data. Finding acceptable trends and patterns to interpret text documents from huge amounts of data is a major challenge. The process of extracting valuable and nontrivial patterns from large amounts of text documents is known as text mining. There are a variety of strategies and tools for mining text for useful information for future forecasting and decision-making. To boost the speed and reduce the time and effort required to obtain important data, an appropriate and proper text mining methodology is applied. This paper discusses and analyses text mining techniques and their applicability in numerous areas of life. Furthermore, challenges in the field of text mining have been discovered, which have an impact on the accuracy and relevance of the results.**

Keywords: Text Mining Process, Techniques, Summarization, Applications.

1. Introduction

Every day, data grows at an exponential rate. Electronic data storage is used by almost all types of institutions, organizations, and corporate businesses. In the form of digital libraries, archives, and other textual material including blogs, social media networks, and e-mails, a tremendous volume of text is streaming across the internet [1]. It's tough to find suitable patterns and trends to obtain relevant insights from such a massive amount of data [2]. Textual data is difficult to mine using traditional data mining methods since extracting information takes time and effort.

"Text mining is a method for extracting interesting and noteworthy patterns from textual data sources in order to explore knowledge. Information retrieval, data mining, machine learning, statistics, and computational linguistics are all used in text mining" [3]. "Text mining techniques like summarization, classification, clustering, and others can be used to retrieve knowledge. Text mining is concerned with natural language text in semi-structured and unstructured formats"[4]. Text mining techniques are employed in a wide range of fields, including industry, academia, web applications, the internet, and others [5]. "In fields such as search engines, customer relationship management systems, filter emails, product suggestion analysis, criminal identification, and social media analytics, text mining is used for opinion mining, feature extraction, sentiment, prediction, and trend analysis" [6].

II. Text Mining Process

The various steps of text mining process are as follows:

- Collecting unstructured data from a variety of sources in a number of document formats, like plain text, web pages, pdf records and others.
- The cleansing approach is used to eliminate stop words, as well as stemming (the procedure of discovering the base of a word) and indexing the data, in order to capture the true substance of text [7].
- Automatic processing is used to audit and cleanse the data set using processing and regulatory approaches.
- The preceding methods' data is used to extract relevant and meaningful data for quick and effective judgment and market analysis [8].

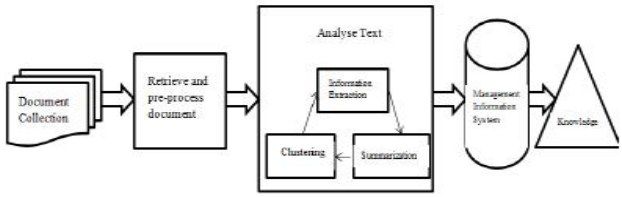

Fig.1: Text Mining Process

III. Text Mining Techniques

There are a variety of text mining approaches that can be used to analyze text patterns and the mining process[9]. We'll go through each one of these technologies and their importance in text mining in the sections that follow. The types of scenarios in which each technology could be beneficial to people are also described.

A. Information Extraction (IE)

A beginning stage for PCs to inspect unstructured content is to utilize data extraction. In text, information extraction software recognizes essential terms and relationships. The software indicates the relationships between all of the recognized persons, locations, and times in order to deliver useful information to the user. When working with vast amounts of text, this technology can be extremely useful. The information to be "mined"

in conventional data mining is assumed to be in the form of a relational database. Tragically, for some applications, electronic data is just reachable as free characteristic language reports as opposed to organized data sets. "The database created by an IE module can be handed to the KDD module for advanced knowledge mining since IE overcomes the difficulty of translating a corpus of textual documents into a more organized database"[10].

B. *Categorization*

Categorization automatically adds one or more categories to a free text document. Because it is relied on input output samples to classify new texts, categorization is a supervised learning method. Text documents are assigned predefined classes based on their content. Pre-processing, indexing, dimensionally reduction, and classification are all steps in the usual text categorization process[11][12]. The purpose of categorization is to train a classifier using known instances, and then automatically categorise unexpected examples. Nave Bayesian classifiers, Nearest Neighbour classifiers, Decision Trees, and Support Vector Machines are examples of statistical classification approaches. Categorization's main purpose is to divide a batch of text into a predetermined number of groups.

C. *Clustering*

The partition of a set of items or data into a sequence of related and recognized subclasses is known as clustering. The most typical application of clustering is to create a group of related papers and files. Clustering provides the benefit of separating the document or text files into several subtopics, reducing the likelihood of significant documents being lost in the search [13]. The clustering method separates entries in a dataset into groups, with themes that are similar within each cluster but differ between clusters. The main purpose of cluster analysis is to locate a group that has some value in respect to the topic at hand. It is not always possible to get the intended result [13].

Clustering can be divided into two categories:

1. Hierarchical
2. Nonhierarchical

A Dendogram, or cluster hierarchy, is a tree-like representation of clusters formed using hierarchical clustering. The shared parents span the point partitioned by the relation cluster, and each parent node cluster has child clusters. The hierarchical clustering method makes it easier to find data at various levels of granularity.

On the other hand, the non-hierarchical approach divides a data collection of x objects into y groups with overlapping boundaries. The approaches are additionally divided into subcategories, which encompass equally unique stamping methods and less frequent stamping methods.

D. *Visualization*

In text mining, visualization approaches can aid in the finding of essential information by improving and simplifying the process. Text flags identify individual papers or groupings of documents, whereas density colors indicate document categorization. Massive volumes of text are organized into a visual hierarchy using visual text mining. Zooming and scaling are two ways for the user to interact with the document. Authorities can utilize information visualization to find information about crimes or to identify terrorist networks. The various steps involved in the visualization process are depicted as follow:

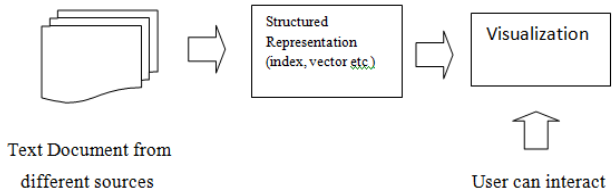

Fig. 2: Visualization [14]

The purpose of data visualization is broken down into three steps:

i. The data preparation process entails choosing on and gathering original visualization data, as well as forming an original data space.
ii. Data analysis and extraction is the process of reviewing and extracting visualization data from source data in order to create a visualization data space.
iii. A mapping algorithm is used in the mapping step of the visualization process to map the visualization data space to the visualization goal.

E. *Summarization*

The act of collecting and constructing short representations of original text materials is known as text summarization [15]. Text summarizing software analyses and summarizes the enormous text content in less time than it takes the user to read the first paragraph. The raw text is subjected to pre-processing and processing operations in order to summarize it. Tokenization, stop word removal, and stemming procedures are utilized for pre-processing. During the text summarizing processing stage, lexicon lists are created.

The steps in the summarizing process are as follows:

(1) Pre-processing transforms the original text into a structured representation.
(2) An algorithm is used to convert the text structure into a summary structure at the next level of processing.
(3) During the innovation process, the final summary is derived from the summary structure.

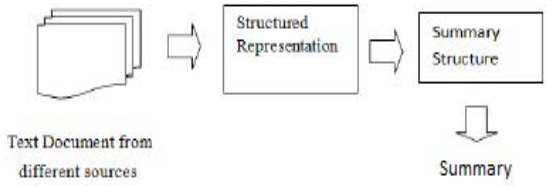

Fig. 3: Summarization [14]

IV. Applications

A. Digital Libraries

To find patterns and trends in journals and proceedings from a wide range of sources, a number of text mining approaches and technologies are applied. For academics, libraries are a valuable source of knowledge, and digital libraries are working to increase the value of their collections. It offers a unique approach to manage information and gain access to number of papers on the internet. "Greenstone international digital libraries provides a quick means to extract documents from a wide range of formats, including MS Word, PDF, postscripts, HTML, script languages, and email messages, as well as a multilingual interface and several languages" [16]. Various operations are conducted in the text mining process, like document selection, enrichment, information extraction, dealing with entities between documents, and generating instinctive co-referencing and summarization.

B. Web Mining

Nowadays, the internet has a wealth of information on people, companies, goods, and other topics of potential interest [17]. "Web mining is a popular use of data mining techniques for uncovering hidden and unknown patterns on the internet. Web mining is a crucial activity for recognizing terms implied in big document collections, such as C, which can be represented by a mapping, *i.e.* C→ p"[17]. In any Web-based text mining project, the first step is to collect a wide range of web pages that include subject observation. As a result, the problem becomes not just locating all subject advances, but also separating those that are of interest.

C. Research Field and Academics

"In the subject of education, several text mining tools and approaches are used to investigate educational patterns in a specific region, student interest in a given field, and employment ratios" [18]. Text mining is used in the research sector to identify and classify research papers and relevant content from various domains in one location. K-means clustering and other algorithms aid in the identification of relevant information's properties. It is possible to obtain information about a student's success in several disciplines, as well as how certain factors influence subject selection. [16] [19]

D. Social Media

Text mining software can monitor or analyze the text from a range of sources, including social media, blogs, the internet, news, and email. Text mining techniques are particularly beneficial for determining or evaluating the overall amount of social media followers, posts, and likes. This type of study demonstrates how individuals reacted to various postings and news, as well as how it propagated. It displays the attitude of people in a particular age group or community, as well as the similarities and variances in their viewpoints on the same subject [20][21].

E. Business Intelligence

Text mining is used by businesses and organizations to evaluate their customers and competitors in order to make effective decisions. It is more important since it gives a clear picture of the firm, increases client satisfaction, and gives you a competitive advantage. Text mining techniques like IBM, text analysis, and fast mine, as well as Gate, assist in drawing conclusions about the institute by highlighting both positive and negative execution.

F. Resume Filtering

Every day, large firms receive thousands and lakhs of resumes from job searchers. Resumes provide extremely accurate information, and sifting through them is a major undertaking. Resumes are available in a variety of forms. Rather having a constrained domain (e.g. basic text or graphs), several languages (for example Spanish and Russian) and file formats are produced (example Word, EXCEL etc.). Furthermore, the writing style can be altered in a variety of ways. The recruiter checks for defects, qualifications, the employee's background, variations in past employment, and personal information during the initial physical scan of the resume. The first step in ignoring resumes will be to collect this information precisely. As a result, selecting resumes is a crucial step in the selection process.

G. Medical and Life Science

"Users frequently share information about topics of interest with others, make requests on online forums, and seek professional assistance. Everyone wants to learn more about specific conditions (what they have), learn about new treatments, and get a second opinion before starting therapy. Furthermore, these forums mention seismographs for medical and/or psychological needs that are currently unmet by current health-care systems"[22]. Quantitative and qualitative methodologies have been used to weigh e-mails, e-consultations, and network-based requests for medical guidance [23]. As a result, specific inquiries might be sent to an expert or even addressed semi-automatically, allowing for total surveillance. Finding an accurate and significant text to make an informed selection from a vast biological collection is a difficult task [24]. Medical records contain a wide range of content, including sophisticated, long, and specialized terminology, making information discovery challenging [25]. Researchers can acquire relevant information on diseases, their linkages, and relationships using text mining methods in the biomedical area [26]. "Text mining approaches are utilized in biomarker development, pharmaceutical businesses, clinical trade analysis, preclinical safe toxicity report research, patent competitive intelligence and landscaping, mapping of genetic illnesses, and exploring the specified identifications" [27].

V. Issues In Text Mining

Several difficulties arise throughout the text mining process, affecting decision-making quality and productivity. Complications can arise during the intermediary stages of text mining. Many restrictions and limits are added to the text during the preprocessing step in order to standardize it and make the text mining procedure more effective. Before using pattern analysis, unstructured data must first be translated into an intermediate format, but this phase of the text mining process has its own set of problems. Due to a change in the text's sequencing[28], the true subject of data may be misplaced numerous times[28]. To accommodate multilingual text, various algorithms and techniques are utilized independently. Because a variety of technologies do not support crucial documents, they remain outside the text mining process. These difficulties wreak havoc on the knowledge finding and decision-making processes. In fact, contemporary text mining approaches rarely support multilingual resources, making real-world advantage difficult to realize [29].

Domain knowledge integration is a crucial field since it conducts certain operations on a corpus and achieves the necessary results. In this scenario, domain expertise, which will be used to extract the document corpus, must be combined with computer capabilities, which will be used to extract the information. Experts from a variety of fields must work together to provide more productive, accurate, and precise results that meet the field's criteria[30], [28]. Because text mining technologies consider all of these in identical scenarios, the texts might cause confusion and complication. It becomes difficult to differentiate the textual content as a large number of papers are evaluated that refer to different classes and still have the identical domain.

To design the plug-ins in depth and with the necessary understanding [31], [32], a specific domain will be required. Text refining approaches are hampered by natural language, which identifies entity relationships and generates problems. For instance, the words tear and tear have the same spelling but have different meanings. Both terms are treated equally by text mining algorithms, although one is a verb and the other is a noun. In the discipline of text mining, grammatical rules based on nature and context are still a work in progress [33].

VI. Conclusion

To extract useful information, a massive amount of text-based data must be reviewed. Text mining techniques are used to retrieve relevant and informative facts from enormous amounts of unstructured data in an effective and efficient manner. This paper provides a high-level overview of the text mining process and strategies that can be used in a variety of applications, including web mining, medical, social media etc. Text mining is the process of evaluating text documents and extracting key phrases, concepts, and other information in order to prepare the text for future analysis using data mining techniques. For predictive analysis, particular sequences and patterns are used to retrieve valuable information by removing extraneous features. The text mining process is made simple and efficient by using the appropriate techniques and tools, which are chosen and applied as per the domain. The research also covered the most complicated aspect of text mining system development.

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," 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] 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.

[10] Thilagavathi, D & Antony, Selvadoss & Thanamani, "An impression on performance metrics for scheduling problem in grid computing environment" international journal of research in computer applications and robotics issn 2320-7345. 2. 40-45, 2014.

[11] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[12] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[13] S. S. Tandel, A. Jamadar and S. Dudugu, "A Survey on Text Mining Techniques," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 1022-1026, 2019.

[14] Sonali Vijay Gaikwad, Archana Chaugule and Pramod Patil. "Text Mining Methods and Techniques". *International Journal of Computer Applications* 85(17):42-45, January 2014.

[15] 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.

[16] 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.

[17] Shiqun Yin Yuhui Qiu1,Chengwen Zhong, 2007. Web Information Extraction and Classification Method .IEEE

[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] 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.

[20] 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.
[21] 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.
[22] 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.)
[23] Daniel Waegel. ―The Development of Text-Mining Tools and Algorithms.Ursinus College, 2006.
[24] Johannes C. Scholtes. ―Text-Mining: The next step in search technology‖, DESI-III Workshop Barcelona, 2009.
[25] Umefjord G, Hamberg K, Malker H, Petersson G FamPract, 2006. The use of an Internet-based Ask the Doctor Service involving family physicians: evaluation by a web survey, 159-66.
[26] Y. Zhao, ―Analysing twitter data with text mining and social networkanalysis,‖ in Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013), 2013, p. 23.
[27] Johannes C. ScholtesA. Voutilainen.―A syntax-based part of speech analyser‖.In Proc. of the Seventh Conference of the European Chapter of the Association for Computation al Linguistics, pages 157–164, Dublin. Association for Computational Linguistics, 1995
[28] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostr¨om, "Ensembles of randomized trees using diverse distributed representations of clinical events," BMC Medical Informatics and Decision Making, vol. 16, no. 2, p. 69, 2016.
[29] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information:-extended abstract," OTT06, vol. 1, p. 55, 2007.
[30] 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.
[31] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif. 2007. "Automatic extraction of synonymy information:-extended abstract," *OTT06*, vol. 1, p. 55.
[32] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri. 2013. "Immune based feature selection for opinion mining," *Proceedings of the World Congress on Engineering*, vol. 3, pp. 3–5.
[33] 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.