International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi :https://doi.org/10.32628/ CSEIT2390391

A Survey on Text Mining - Techniques, Application

Ajay Jadhav[1], Pranjal Jagtap[1], Suraj Gurav[1], Shivani Jadhav[1], Nikita Jadhav[1], Afsha Akkalkot[2]
[1]TE Students, Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India
[2]Assistant Professor, Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 May 2023
Published: 30 May 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
338-343

ABSTRACT

Text mining, also known as text data mining or text analytics, is a field of study that focuses on extracting meaningful information and knowledge from textual data. The rapid advancement of digital data acquisition techniques has resulted in an unprecedented volume of data. In fact, over 80 percent of the data generated today comprises unstructured or semi-structured formats. Extracting meaningful patterns and trends from such massive amounts of text data poses a significant challenge. Text mining addresses this challenge by extracting valuable and nontrivial patterns from vast collections of text documents. Various techniques and tools are available for mining text documents and uncovering valuable information to inform decision-making and future processing. Selecting the appropriate text mining technique is crucial as it can significantly enhance the speed and efficiency of retrieving valuable information, reducing the time and effort required. This paper provides a concise analysis and discussion of text mining techniques and their applications. As technology continues to advance, the availability of digital data continues to increase. A substantial portion, approximately 85 percent, of this data exists in unstructured textual form. Consequently, it has become imperative to develop improved techniques and algorithms to effectively extract useful and interesting information from these vast amounts of textual data. This has resulted in the emergence of information extraction and text mining as popular research areas dedicated to uncovering valuable and necessary information from textual data.

Keywords: Classification; Text Mining Algorithm; text mining, text data mining, text analytics, textual data, Knowledge Discovery; Applications; Information Extraction; Information Retrieval; Patterns. information extraction, knowledge discovery, techniques, applications.

I. INTRODUCTION

Text mining, also referred to as the extraction of valuable information from textual data, has emerged as a crucial field of study [1]. Its focus lies in analyzing natural language text to uncover meaningful insights. Unlike structured data stored in databases, text data is unstructured, ambiguous, and challenging to process. Nonetheless, text remains the predominant medium for formal information exchange in today's society. Text mining addresses the communication of factual information and opinions, making it distinct from data mining, which primarily utilizes structured data from databases. Instead, text mining operates in domains characterized by unstructured or semi-structured datasets, such as emails, text documents, and HTML files [2]. Consequently, text mining provides a more effective solution. It involves extracting significant patterns from textual databases, leading to the exploration of valuable knowledge [3]. This multidisciplinary field draws upon information retrieval, data mining, machine learning, statistics, and computational linguistics [3]. Text mining deals with real-language text stored in semi-structured and unstructured formats [4]. Its techniques find widespread application in various domains, including industry, academia, web applications, the internet, and more [5]. Text mining is employed in search engines, customer relationship management systems, email filtering, product analysis, fraud detection, and social media analytics. These applications leverage text mining for tasks such as opinion mining, feature extraction, sentiment analysis, predictive analysis, and trend analysis [6]. The text mining process involves structuring input text data through parsing, incorporating derived linguistic features, eliminating irrelevant ones, and storing it in a database. Patterns are then derived from the structured data, followed by the final step of interpretation and validation of the output.

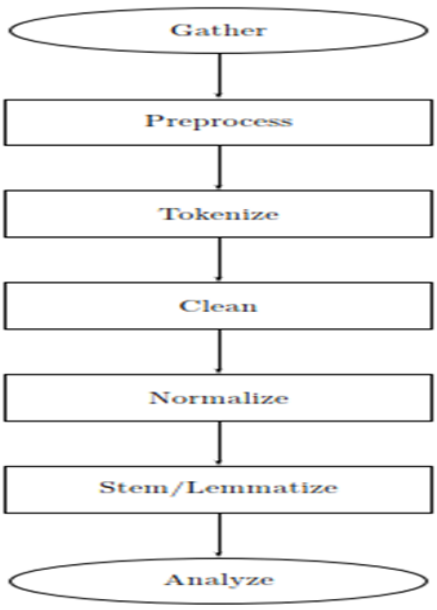

Fig.1: Basic Process of Text Mining

The term —text mining is most commonly used to relate any system that examines huge quantities of real language text and finds lexical or linguistic usage methods in an attempt to extract needful information.

A. Areas Of Text Mining

Text analysis involves data retrieval, information extraction, data mining techniques includes association and link analysis, visualization, and predictive analysis. The aim is, essentially, to turn text (unstructured data) into data (structured format) for analysis, via the use of natural language processing (NLP) methods.

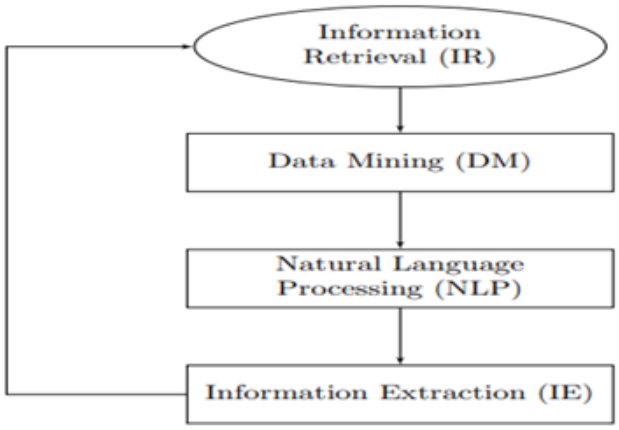

Fig.2: Text mining areas

B. Information Retrieval (IR)

Information retrieval (IR) refers to the process of retrieving documents based on user queries and subsequently extracting or summarizing specific information of interest. Document retrieval involves returning and processing relevant documents to obtain the desired information as requested by the user. Following document retrieval, additional stages such as text summarization or information extraction can be employed to further refine and condense the retrieved content using various techniques. The primary objective of IR systems is to narrow down the pool of documents that are pertinent to a given problem. By leveraging IR in conjunction with text mining, the analysis of vast document collections can be accelerated significantly [8] by reducing the number of documents requiring analysis. This not only enhances efficiency but also aids in focusing on the most relevant information for further exploration and insights.

C. Data Mining (DM)

Data mining is a process that involves the exploration of data to discover patterns and extract valuable information. It encompasses the identification of previously unknown and hidden insights [10]. By utilizing data mining tools, businesses can predict behaviors and future trends, enabling them to make informed decisions based on knowledge. Moreover, data mining tools have the capability to address complex business questions that were traditionally time-consuming to resolve. These tools comb through databases, uncovering hidden and unexpected patterns that may elude domain experts due to their preconceived expectations. The ultimate objective of the data mining process is to extract meaningful information from a dataset and present it in a structured format for further utilization.

D. Natural Language Processing (NLP)

Natural Language Processing (NLP) poses one of the oldest and most complex challenges in the realm of artificial intelligence. It encompasses the study of human language with the objective of enabling computers to comprehend natural languages in a manner similar to humans [8]. NLP research delves into the intricate question of how we derive meaning from sentences or documents. What cues do we employ to understand the subjects, actions, and recipients involved? How do we discern the timing of events or differentiate facts from conjecture or prediction? While words such as nouns, verbs, adverbs, and adjectives [8] serve as the fundamental units of meaning, it is their interrelation within the sentence structure and their contextual significance based on our existing knowledge of the world that truly unveils the essence of a text. In the realm of text mining, NLP assumes a crucial role in the information extraction phase, providing the system with valuable input for analysis.

E. Information Extraction (IE)

Information Extraction (IE) refers to the automated extraction of structured information from machine-readable documents that are unstructured or semi-structured in nature. Typically, this involves the utilization of natural language processing (NLP) techniques to process human language texts. Moreover, the domain of multimedia document processing, encompassing tasks such as automatic annotation and mining information from images, audio, and video, can also be considered as a form of information extraction. A prime example showcasing the practicality and effectiveness of IE is the Google Search Engine. The process of IE entails defining templates that represent the desired form of information, which serve as guides during the extraction process. Notably, IE systems heavily rely on the data produced by NLP systems, further emphasizing the interconnectedness of these two domains.

II. WHAT IS TEXT MINING?

A. The Concept

Text mining, also known as text data mining or text analytics, is a field within data science that involves the process of extracting valuable and meaningful information from textual data. It encompasses various techniques and methodologies aimed at analyzing and interpreting large volumes of unstructured or semi-structured textual data. [6]. Text mining may be characterized as the process of extracting valuable information from text for specific purposes. Textual data, unlike structured data stored in databases, lacks organization, poses ambiguity, and presents challenges in processing. However, text remains the primary medium for formal information exchange in contemporary society. Text mining primarily focuses on analyzing texts that communicate factual information or opinions, driven by the compelling need to automatically extract information from such texts, even if achieving complete success remains challenging. The early stages of text mining relied on manual techniques during the 1980s [11]. However, it soon became evident that manual approaches were labor- intensive, costly, and insufficient to handle the growing volume of information. Subsequently, significant progress has been made in developing automated programs capable of efficiently processing information. The field of text mining encompasses various mathematical, statistical, linguistic, and pattern recognition techniques that enable the automatic analysis of unstructured information and extraction of relevant and high-quality data, ultimately enhancing searchability of the entire text. A text document comprises characters that form words, which can be further combined to create phrases. These syntactic properties represent predefined categories, concepts, senses, or meanings [11]. Text mining involves recognizing, extracting, and utilizing this information. Rather than searching for individual words, text mining enables the exploration of semantic patterns, allowing for searches at a higher level of meaning.

B. Process

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

C. Text Pre-processing

It involves a series of steps that are shown in the below figure.

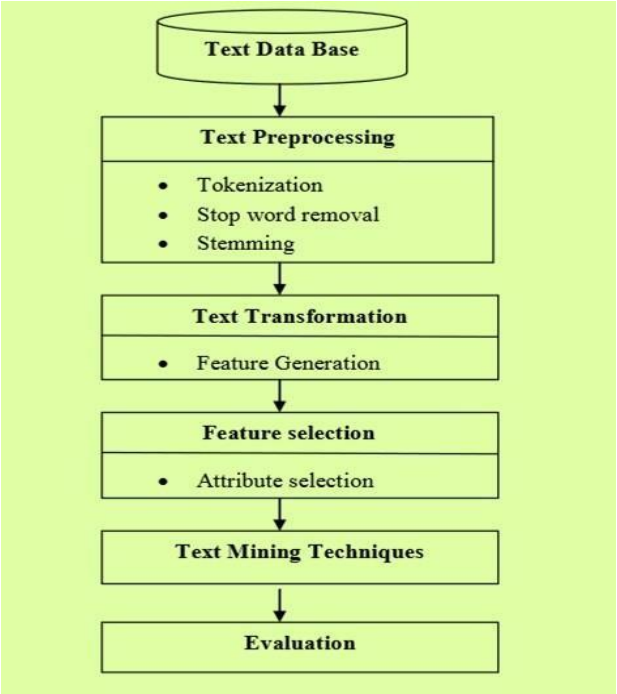

Fig.3: Text preprocessing

D. Text Cleanup

Text cleanup involves the elimination of superfluous or undesirable content, such as removing advertisements from web pages, normalize text converted from binary formats, deal with tables, figures and formulas.

E. Tokenization

Tokenizing can be accomplished by dividing the text into separate tokens based on white spaces and punctuation marks, excluding those that are part of

recognized abbreviations determined in the previous step.

F. Part of Speech Tagging

Part-of-Speech (POS) tagging involves assigning a word class to each token in a text, using the input provided by the tokenized text. Its input is given by the tokenized text. Taggers have to cope with unknown words (OOV problem) and ambiguous word-tag mappings. Rule-based approaches like ENGTWOL [12] rely on dictionaries that contain word forms accompanied by their respective POS labels, as well as morphological and syntactic attributes. These approaches also utilize context-sensitive rules to determine the appropriate labels when applying the tagging process.

G. Text Transformation (Attribute Generation)

A text transformation is a technique that is used to control the capitalization of the text. Two main approaches of document representation are a) Bag of words b) Vector Space.

H. Feature Selection (Attribute Selection)

Feature selection is a significant part of data mining. Feature selection can be defined as the process of reducing the input of processing or finding the essential information sources. The feature selection is also called variable selection. Feature selection also known as variable selection, is the process of selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the datacontain many redundant or irrelevant features.

I. Data Mining

At this view the Text mining process is joined with the traditional Data Mining process. The text mining procedure merges with the conventional process. Classic Data Mining procedures are used in structural database.

J. Evaluate

Check the result for the correctness, after the corrections the result can be miss out or it evaluates the results. Once the result is evaluated, the result abandon.

III.APPLICATION

Text Mining can be applied in many areas [13]. Some of the most common used areas are:

3.1 Web Mining.

In the digital age, the vast expanse of the web contains an abundance of information spanning various subjects such as individuals, companies, products, and more [14]. Web mining, as an application of data mining techniques, plays a pivotal role in uncovering hidden patterns and extracting valuable insights from this wealth of web-based data. Web mining involves the crucial task of identifying relevant terms within large collections of documents, denoted as C, through a mapping process denoted as C → p [14]. The initial step in any web-based text mining endeavor is to gather a substantial number of web pages that pertain to a specific subject of interest. Subsequently, the challenge lies not only in discovering all developments related to the subject but also in distinguishing those instances that carry the intended meaning. Effective web mining techniques enable the extraction of meaningful and relevant information from the vast sea of web content, facilitating knowledge discovery and informed decision-making.

3.2 Clustering

Clustering is an unsupervised process that aims to classify text documents into groups based on various clustering algorithms. It involves grouping together similar descriptions or patterns extracted from

different documents. Clustering can be performed using both top-down and bottom-up approaches. In the field of Natural Language Processing (NLP), a wide range of mining tools and techniques are employed to analyze unstructured text data. Different clustering methods, such as distribution-based, density-based, centroid-based, hierarchical, and k-means clustering [22], are utilized to organize and categorize text documents effectively. These clustering techniques assist in discovering patterns, relationships, and similarities within the text, enabling better understanding and interpretation of the underlying data.

3.3 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 the number of posts, likes and followers on the social media. This kind of findings show the people reaction on different posts, news and howit gets spread around. It shows the behavior of people belonging to specific age group or communities having similarity and differences in views about the specific post [23], [24].

3.4 Resume Filtering

Large corporations and recruitment agencies receive an overwhelming number of resumes from job applicants on a daily basis. Extracting accurate and relevant information from these resumes poses a significant challenge [1]. Resumes do not adhere to a standardized format; they can be presented in various structures, such as structured tables or plain text, written in different languages like Japanese and English, and saved in different file types, including Plain Text, PDF, and Word documents. Furthermore, writing styles can greatly vary. During the initial manual screening of resumes, recruiters typically search for errors, educational qualifications, work history, job titles, frequency of job changes, and other personal details. Acquiring this information correctly becomes crucial as it forms the basis for shortlisting or disregarding resumes. Therefore, the resume selection process plays a pivotal role in the recruitment process, demanding careful attention and consideration.

3.5 Medical and life science

Users often engage in information exchange, seek advice on web-based forums, or request expert services in areas of interest [15]. People have a strong desire to understand their own medical conditions, stay informed about new treatments, and seek second opinions before undergoing treatment. These forums also serve as indicators of unmet medical and psychological needs, which existing healthcare systems may not adequately address [15]. Various forms of communication, such as emails, e-consultations, and online medical advice requests, have been manually evaluated using quantitative or qualitative methods [16]. To assist medical experts and harness the potential of expert forums, it would be beneficial to promptly identify visitors' requests. This would enable directing specific requests to the appropriate experts or even generating semi-automated responses for comprehensive monitoring. Creating a repository of frequently asked questions (FAQs) based on similar patient requests [16] and their corresponding answers could aid in addressing queries even before the involvement of specific experts. Machine-based analysis could assist the public in managing the overwhelming amount of information while enabling medical experts to provide their expertise. Instantly classifying amateur requests in medical expert network forums presents a significant challenge as these requests often involve long and unstructured narratives that blend personal experiences with laboratory data. Extracting accurate and pertinent information from vast biomedical repositories is a daunting task, considering the diverse and complex nature of medical records and the technical vocabulary employed [19][20]. Text mining tools in the biomedical field offer valuable

opportunities for obtaining insights, discovering associations and relationships among different diseases [21]. Text mining finds application in diverse areas such as biomarker discovery, pharmaceutical companies, clinical trade analysis, preclinical safe toxicity report studies, patent competitive intelligence and landscaping, mapping of genetical diseases and exploring the specified identifications by using different techniques [18].

IV.CONCLUSION

In conclusion, text mining plays a crucial role in extracting meaningful information from large volumes of unstructured text data. In order to extract valuable insights from a vast amount of unstructured text data, efficient text mining methods are employed. These methods aim to effectively identify and retrieve relevant information while disregarding irrelevant details. This paper provides a concise overview of text mining techniques that contribute to enhancing the text mining process. By leveraging specific patterns and sequences, meaningful information can be obtained for predictive research purposes. Selecting appropriate methods and tools tailored to the specific domain further enhances the ease and efficiency of the text mining process. Addressing challenges such as domain knowledge integration, concept granularity, multilingual text refinement, and managing ambiguity through natural language processing are crucial considerations in text mining techniques. Particularly in the medical field, the use of accurate text mining tools proves beneficial in assessing the effectiveness of various medical treatments by comparing different diseases and symptoms. Text mining holds significant advantages, particularly in life sciences and healthcare.

V. REFERENCES

[1]. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods. 2011; 2:1-14.

[2]. Vishal Gupta, Gurpreet S. Lehal, 2009. —A Survey of Text Mining Techniques and Applications‖ in Journal of Emerging Technologies in Web Intelligence, Vol. 1 No. 1.

[3]. Shiqun Yin Yuhui Qiu1,ChengwenZhong, 2007. Web Information Extraction and Classification Method .IEEE

[4]. 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.

[5]. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med.2010;7: e1000326

[6]. Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins J, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (Updated March 2011). The Cochrane Collaboration; 2011.

[7]. Navathe, Shamkant B. and ElmasriRamez. —Data Warehousing and Data Mining‖, in —Fundamentals of Database Systems‖, Pearson Education pvtInc, (Singapore, 841-872, 2000).

[8]. Widman LE, Tong DA Arch (Intern Med. 1997), Requests for medical advice from patients and families to health care providers who publish on the World Wide Web. 209-12.

[9]. W. Fan, L. Wallace, S. Rich, and Z. Zhang, —Tapping the power of textmining,‖ Communications of the ACM, (vol. 49, no. 9, pp.76–82, 2006).

[10]. S.-H. Liao, P.-H.Chu, and P.-Y. Hsiao, —Data mining techniques andapplications–a decade review from 2000 to 2011,‖ (Expert Systems withApplications, vol. 39, no. 12, pp. 11 303–11 311, 2012.)