IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114199

TEXT MINNING: TECHNIQUES, APPLICATIONS AND ISSUES

Poonam Balaji Parnale, Sheetal A Wadhai

Department of Computer Engineering, Universal College of Engineering And Research, Pune

Abstract: Rapid advancements in digital data collecting techniques have resulted in massive data volumes. Unstructured or semi-structured data makes up more than 80% of today's data. Finding acceptable patterns and trends to interpret text documents from huge amounts of data is a major challenge. The process of extracting interesting 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 prediction and decision-making. The suitable and appropriate text mining technique is used to increase the speed and reduce the time and effort necessary to retrieve relevant data. Text mining techniques and their applications in various spheres of life are briefly discussed and analysed in this work. Furthermore, text mining difficulties that affect the accuracy and relevancy of results are identified.

Keywords: Information Extraction; Patterns; Classification; Knowledge Discovery; Applications;

I. INTRODUCTION

Every day, the size of data grows at an exponential rate. Electronic data storage is used by almost all types of institutions, organisations, and corporate industries. In the form of digital libraries, archives, and other textual material such as blogs, social media networks, and e-mails, a massive volume of text is streaming across the internet [1]. Determining appropriate patterns and trends to extract valuable knowledge from this vast volume of data is a difficult task [2]. Textual data is difficult to mine using traditional data mining methods since extracting information takes time and effort. Every day, the size of data grows at an exponential rate. Electronic data storage is used by almost all types of institutions, organisations, and corporate industries. In the form of digital libraries, archives, and other textual material such as blogs, social media networks, and e-mails, a massive volume of text is streaming across the internet [1]. Determining appropriate patterns and trends to extract valuable knowledge from this vast volume of data is a difficult task [2]. Textual data is difficult to mine using traditional data mining methods since extracting information takes time and effort. [5]. Text mining is used for opinion mining, feature extraction, sentiment, prediction, and trend analysis in domains such as search engines, customer relationship management systems, filter emails, product text mining process are as follows: (Figure 2) Obtaining unstructured data from a variety of source

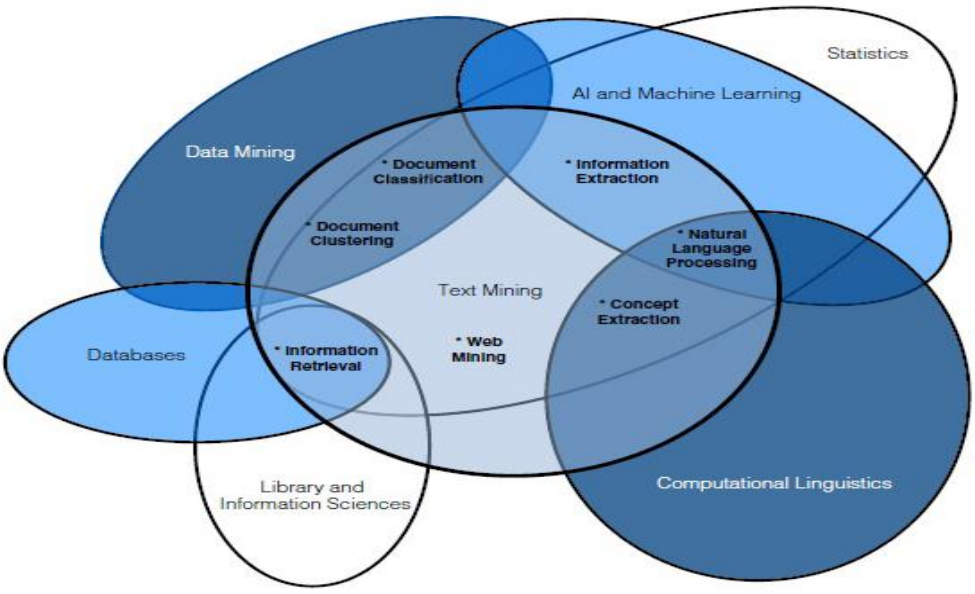

Figure 1 shows the interplay of text mining with other fields [4].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114199

Available in a variety of file types, including plain text, web pages, and PDF files. Anomalies are detected and removed using pre-processing and cleansing techniques. The cleaning procedure ensures that the true essence of the text is captured, and it is used to remove stop words, stemming (the process of determining the root of a word), and indexing the data [7]. Automatic processing is used to audit and further clean the data set using processing and regulating procedures. Management Information System implements pattern analysis (MIS). The above-mentioned methods are utilised to extract meaningful and relevant data for quick and effective decision-making and trend analysis [8]

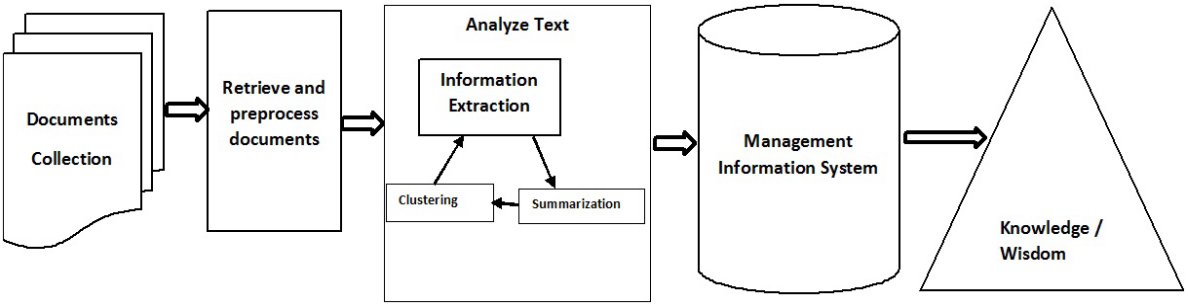

Fig. 2. Text mining process [5]

Extraction of useful information from a large number of documents is a time-consuming and exhausting operation. The assortment of The right text mining technique can cut down on the time and effort it takes to identify important patterns for analysis and decision-making. The goal of this research is to examine several text mining approaches that aid in performing text analytics from big amounts of data effectively and efficiently. The problems that develop throughout the text mining process are also identified. This paper is divided into several sections. Section II discusses previous work. Different text mining approaches are presented in section III. The application areas of text mining techniques are discussed in Section IV. Issues and challenges in the field of text mining are highlighted in section V. The outcomes are summarised in Section VI.

II. REVIEW OF LITERATURE

Gathering, extracting, pre-processing, text modification, feature extraction, pattern selection, and evaluation are all processes in the text mining process, according to [5]. In addition, the application of various frequently used text mining techniques, such as clustering, classification, and decision tree categorization, in various domains is examined. [8] drew attention to the problems with text mining applications and approaches. They addressed how dealing with unstructured text is more difficult using typical mining tools and procedures than dealing with organised or tabular data. They demonstrated how text mining may be used in biology, corporate intelligence, and national security. Text mining difficulties have been decreased thanks to advances in natural language processing and entity recognition algorithms. However, there are several concerns that require care.

[9] integrated a framework for named entity recognition, text classification, hypothesis development and testing, relationship and synonym extraction, and extract abbreviations into the MEDLINE biomedical database. This new framework aids in the removal of irrelevant details and the extraction of useful data. [10] used text mining patterns to examine the text and found that term-based techniques fail to correctly analyse synonyms and polysemy. Furthermore, a prototype model was created for pattern specification in terms of weighting patterns based on their distribution. This method aids in improving the text mining process' efficiency. [11] described a text mining-based crime detection system that used a relation finding algorithm to associate terms with abbreviations.

[12] proposed a top-down and bottom-up strategy to text mining on the web. They use the k-mean clustering technique for bottom up partitioning to join related text documents. To find information about specific issues, the TF-IDF (Term Frequency- Inverse Document Frequency) algorithm was employed to find similarity within the document. [13] provided an overview of text mining applications, techniques, and difficulties. They talked about how papers can be structured, semi-structured, or unstructured, and how extracting relevant information is a time-consuming process. They presented a conceptual mining paradigm that may be viewed as text refinement and knowledge distillation processes. Depending on the domain, the intermediate form of entity representation mining is used.

[14] described novel and effective pattern discovery methods. They made use of the pattern of evolution and discovery. strategies for improving the efficiency of finding relevant and acceptable data To assess the effectiveness of the recommended technique, they used BM25 and vector support machine based filtering on data from the router corpus volume 1 and the text retrieval conference. [15] conducted a number of classification studies on text using multi-word

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114199

characteristics. They proposed a method for extracting multi-word characteristics from the data set that was done by hand. They divide text into linear and nonlinear polynomial forms in order to classify and extract multi-word text, which helps the extracted data be more successful.

III. THE REFLECTIVE PROCESS

There are several ext mining approaches that can be used to analyse text patterns and the mining process [16]. The Venn diagram illustrating the interrelationships among text mining approaches and their essential functionality is shown in Figure 3. Information retrieval (keyword search / querying and indexing), document clustering (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) are all examples of document classification [6].

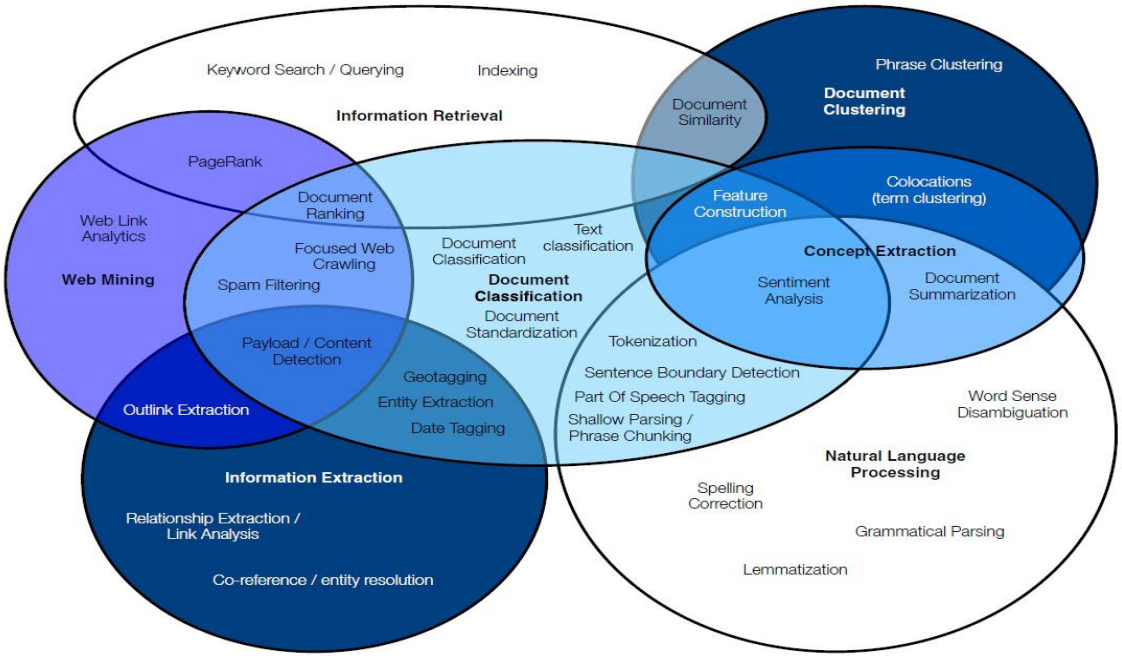

Figure 3 shows the interrelationships between various text mining approaches and their basic functions [6].

A. INFORMATION EXTRACTION

Information Extraction (IE) is a method for extracting useful data from enormous amounts of text. Domain specialists define domain-specific properties and relationships [17]. Specific characteristics and entities are extracted from the document using IE systems, and their relationships are established [18]. The extracted corpus is saved in a database to be processed later. To assess and evaluate the relevance of results on the extracted data, the precision and recall procedure is employed. To undertake the information extraction procedure and obtain more relevant results, in-depth and complete knowledge about the relevant field is necessary [19].

B. INFORMATION RETRIEVAL

The practise of collecting relevant and associated patterns based on a set of words is known as information retrieval (IR). or a phrase Text mining and information retrieval for textual data have a close relationship. Different algorithms are utilised in IR systems to track user behaviour and search relevant material accordingly [19]. Google and Yahoo search engines are increasingly adopting information retrieval systems to extract relevant documents based on a Web word.

To detect trends and produce more significant results, some search engines employ query-based algorithms. These search engines provide users with more relevant and relevant information that meets their needs [8].

C. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is the process of automatically processing and analysing unstructured textual data.

It performs a variety of analyses, including Named Entity Recognition (NER) for abbreviation and synonym extraction to discover links between them [10]. NER extracts all instances of a particular object from a collection of documents. These entities and instances allow for the identification of relationships and other information in order to

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114199

achieve their core notion. This technique, on the other hand, lacks a complete dictionary list for all named things needed for identification [9], [10]. To achieve satisfactory results, complex query-based algorithms must be applied. A single entity in the actual world has multiple titles, such as TV and Television. A multi-word moniker is sometimes used to distinguish a series of consecutive words. Using classification techniques, you may define boundaries and address overlapping situation. Approaches to NER usually fall into one of four categories: lexical, rule, statistical, or a combination of these. The relevance level of NER systems has ranged from 75 to 85 percent. The co-referencing approach is often used in NLP to extract synonyms and abbreviations from textual input. Natural Languages (NL) are complicated by the fact that text taken from various sources does not contain similar terms or abbreviations.

It is necessary to recognise such problems and establish standards for their consistent detection [21]. For example, NER and co-referencing techniques construct a logical relationship to extract and identify a person's position in an organisation (use a person's name once, then use pronoun instead of name repeatedly) [22].

D. CLUSTERING

Clustering is an unsupervised technique that uses several clustering algorithms to categorise text sources into groups.

Similar terms or patterns are clustered in clusters and extracted from various documents. Clustering is done in two ways: top-down and bottom-up. For the study of unstructured text, NLP employs a variety of mining tools and approaches. Hierarchical, distribution, density, centroid, and k-mean are some of the clustering strategies [22].

E. TEXT SUMMERIZATION

The act of collecting and constructing short representations of original text materials is known as text summarization [23]. For summarising, pre-processing and processing processes are applied to the raw text. Pre-processing procedures include tokenization, stop word removal, and stemming. Lexicon lists are created during the text summarising processing stage. Automatic text summarising used to be done based on the presence of a specific word or phrase in a document. Additional text mining approaches were then used with the regular text mining procedure to improve the relevance and accuracy of the results [11]. The weighted heuristics approach extracts information from text documents by following certain rules. Text summerization can use properties including sentence length, fixed phrase, paragraph, thematic word, and upper case word identification. Text summarising techniques can be used simultaneously on numerous texts. The nature and theme of the text documents influence the quality and type of classifiers [24].

IV. APPLICATIONS OF TEXT MINNING

A. DIGITAL LIBRARIES

To find patterns and trends in journals and proceedings from a large number of sources, a variety of text mining approaches and tools are used. These information sources are beneficial in the field of research and development. For academics, libraries are a valuable source of knowledge, and digital libraries are working to increase the value of their collections. It provides a revolutionary approach of arranging information in such a way that trillions of documents can be accessed online. It offers a revolutionary method of organising data and enables internet access to millions of documents. The Green-stone international digital library, which supports several languages and multilingual interfaces, offers a flexible way for extracting documents in a variety of formats, including Microsoft Word, PDF, Postscript, HTML, and others. E-mail communications and programming languages [11]. Along with text documents, it also allows document extraction in the form of audiovisual and image formats. Various operations are conducted in the text mining process, such as document selection, enrichment, extracting information, handling entities among documents, and generating instinctive co-referencing and summarising [25]. Text mining tools such as GATE, Net Owl, and Aylien are extensively employed in digital libraries.

B. ACADEMIC AND RESEARCH FIELD

Various text mining tools and approaches are employed in the field of education to examine educational trends in a certain region, student interest in a specific field, and employment ratios [24].Text mining is used in the research sector to identify and classify research papers and relevant content from many fields all in one location. The use of k-means clustering and other algorithms aids in the identification of relevant information's properties. Students' success in various disciplines can be accessible, as well as how different traits influence subject selection [11], [26].

C. LIFE SCIENCE

Patients' records, diseases, medicines, symptoms and treatments of diseases, and many other topics generate a vast volume of textual and numerical data in the life science and health care industries. Filtering out an acceptable and relevant literature from a massive biological collection to make a judgement is difficult [25]. Medical records contain a wide range of sophisticated, long, and technical terminology that makes knowledge discovery extremely challenging [27]. Text Mining methods in the biomedical area allow researchers to extract important data, associate it with other data, and infer

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114199

relationships between diseases, species, and genes. The use of proper text mining technologies in the medical field aids in evaluating the effectiveness of medical therapies by comparing different diseases, symptoms, and treatment courses [28]. Using diverse approaches, text mining is used in biomarker discovery, pharmaceutical industry, clinical trade analysis, preclinical safe toxicity investigations, patent competitive intelligence and landscaping, mapping of genes illnesses, and investigating targeted identifications [20].

D. SOCIAL MEDIA

For monitoring social media applications, text mining software packages are available to monitor and analyse plain text from the internet, such as news, blogs, and email. Text mining technologies aid in the identification and analysis of the amount of posts, likes, and followers on social media networks. This type of study demonstrates how individuals reacted to various articles, news, and how it propagated. It depicts the behaviour of people belonging to a given age group or community, with similarities and differences in their perspectives on the same topic [29], [30].

E. BUSINESS INTELLIGENCE

Text mining is an important part of business intelligence since it allows companies to study their consumers and competitors in order to make better decisions. It gives a better understanding of company and information on how to improve customer satisfaction and acquire competitive advantages [31]. Text mining systems such as IBM text analytics, Rapid miner, and GATE assist in making organisational decisions by generating warnings regarding excellent and bad performance, market transition, and remedial actions. It is also beneficial to the telecommunications industry, business and commerce applications, and customer chain management systems [32].

V. ISSUES IN TEXT MINNING FIELD

Many challenges arise during the text mining process, affecting decision-making efficiency and efficacy. At the intermediate stage of text mining, complications can develop. Various rules and restrictions are set in the preprocessing stage to standardise the text and make the text mining process more efficient. There is a requirement to convert unstructured data into intermediate form before applying pattern analysis to the document, however the mining process has its own challenges at this stage. Due to changes in the text sequence, true themes or data might sometimes lose their significance [27]. Another significant issue is the multilingual text refinement reliance, which causes issues. There are only a few tools that handle many languages [33]. To accommodate multilingual text, various algorithms and techniques are utilised independently. Because Because many technologies do not support them, many significant documents remain outside the text mining process. These difficulties wreak havoc on the knowledge finding and decision-making processes. Because contemporary text mining algorithms and tools rarely support multilingual materials, real advantage is difficult to achieve [34].

Domain knowledge integration is significant because it conducts specific operations on a corpus and achieves specific results. expected outcomes In this case, domain knowledge from which the document corpus will be extracted must be combined with computer capabilities from which the information will be obtained. Experts from many areas must collaborate to extract more effective, precise, and accurate outcomes, according to the field's requirements [22], [27]. The employment of synonyms, polysems, and antonyms in texts causes challenges (abstruseness) for text mining methods that combine the two. When a collection of documents is huge and comes from many fields with the same domain, categorising them can be tough. Abbreviations can have varied meanings depending on the situation [35]. Various granularity notions alter the context of text depending on the situation and domain expertise. There is a need to define field-specific rules that will be utilised as a standard in the field and can be included as a plug-in in text mining tools. Developing and deploying plug-ins in each field independently takes a lot of time and work. It will be necessary to have in-depth knowledge of the relevant area in order to design plug-ins [34], [36]. Natural languages have many complexities that cause problems with text refining methods and entity relationship recognition. Words with the same spelling but different meanings, such as fly. and fly. While one is a verb and the other is a noun, text mining technologies believe them to be similar. In the field of text mining, grammatical norms according to nature and context are still an unresolved question [36].

VI. CONCLUSION

The availability of a large volume of text-based data must be investigated in order to extract useful information. Text mining techniques are used to extract interesting and relevant information from enormous amounts of unstructured data in an effective and efficient manner. This paper gives a quick review of text mining approaches that can help improve the process. For predictive analysis, specific patterns and sequences are used to retrieve useful information by removing extraneous details. The correct techniques and tools, chosen and used according to the domain, make text mining simple and efficient. During the text mining process, key concerns and challenges such as domain knowledge integration, changing idea granularity, multilingual text refining, and natural language processing ambiguity develop. In future

DOI: 10.17148/IJARCCE.2022.114199

studies, We shall concentrate our efforts on developing algorithms that will aid in the resolution of the difficulties raised in this paper.

REFERENCES

[1] R. Sagayam, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012, A overview of text mining: Retrieval, extraction, and indexing strategies.
[2] N. Padhy, D. Mishra, R. Panigrahi, and others, "Data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.
[3] "Tapping the Power of Text Mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006. W. Fan, L. Wallace, S. Rich, and Z. Zhang.
Text mining: predictive approaches for evaluating unstructured material, S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. 2010 by Springer Science and Business Media.
[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, December 2011. 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, 2013, pp. 90–102.
[7] "Computer-assisted keyword and document set discovery from unstructured text," by G. King, P. Lam, and M. Roberts. http://j. mp/1qdVqhx a copy Download Paper, vol. 456, 2014, BibTex Tagged XML Download Citation
[8] N. Zhong, Y. Li, and S.-T. Wu, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30–44, 2012.
"Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014. [9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol.
[10] "Improved approach for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.
[11] "Data-intensive applications, challenges, methodologies, 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, vol. 59, no. 1, 2013, pp. 159–171.
[13] "Text mining: Concepts, applications, methods, and issues-an overview," International Journal of Computer Applications, vol. 80, no. 4, 2013.
[14] J. Korra and P. J. Joby, "Accessing accurate documents by mining auxiliary document information," Advances in Computing and Communication Engineering (ICACCE), 2015. IEEE, 2015, pp. 634–638. Second International Conference on
[15] "A research with multi-word feature and text categorization," 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 study of text mining approaches and applications," Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, 2009, pp. 60–76.
[17] R. Agrawal and M. Batra, "A comprehensive study on text mining approaches," ISSN: 2231–2307, International Journal of Soft Computing and Engineering (IJSCE), 2013.
[18] "A review of text mining techniques connected with various application areas," International Journal of Science and Research (IJSR), vol. 4, no. 2, pp. 2461–2466, 2015. D. S. Dang and P. H. Ahmad, "A review of text mining techniques associated with various application areas," IJSR, vol. 4, no. 2,
[19] R. Steinberger, "A survey of strategies for making highly multilingual text mining applications easier to design," Language Resources and Evaluation, vol. 46, no. 2, pp. 155–176, 2012.
[20] "A summary of current work in biological text mining," Briefings in Bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.
[21] E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernandez, "Searching research papers using clustering and text mining," in IEEE International Conference on Electronics, Communications, and Computing (CONIELECOMP), 2013, pp. 78–81.
[22] B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," International Journal of Science, Engineering, and Technology, vol. 3, no. 3, pp. 69–71, 2015.
[23] B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis based document summarization," vol. 14, no. 4, p. 145, International Journal of Computer Science and Information Security, 2016.
"Text summarization extraction system (tses) employing extracted keywords," R. Al-Hashemi. 164–168 in Int. Arab J. e-Technol., vol. 1, no. 4, 2010.
[25] International Journal on Digital Libraries, vol. 4, no. 1, pp. 56–59, 2004. I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library."

DOI: 10.17148/IJARCCE.2022.114199

[26] "Data mining approach for higher education system," European Journal of Scientific Research, vol. 43, no. 1, pp. 24–29, 2010.

[27] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostrom, "Ensembles of Musical Instruments," BMC Medical Informatics and Decision Making, vol. 16, no. 2, p. 69, 2016. randomised trees employing heterogeneous distributed representations of clinical events

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

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

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

[31] "Impact and use of F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha," 256–264, European Journal of Scientific Research, vol. 47, no. 2, 2010.

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

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

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

[35] "Text analytics for android project," Procedia Economics and Finance, vol. 18, pp. 610–617, A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure. "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, vol. 3, 2013, pp. 3–5. www.

