International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Techniques, Applications and Issues of Text Mining

B UDAY JASWANTH REDDY , C H A RAGHU RAMI REDDY, B KAVYA

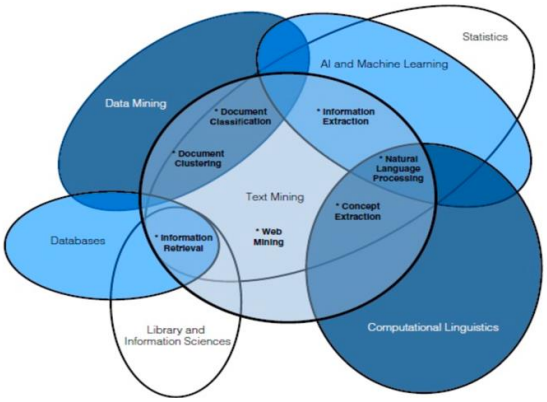
ABSTRACT

The Tremendous advancements in digital data collecting techniques have resulted in massive large datasets. unstructured data makes up well over 80% of today's data. The identification of wonderful ways and characteristics to learn text - based files from a massive amount of datasets are a major issue. Text mining /Data analytics is the process for identifying interesting and difficult problem patterns in vast quantities of text information. For mining textual material and uncovering useful facts for making predictions and decision making, there are various ways and methods available. Choosing an acceptable and wonderful text/word - based analysis approach improves speed and saves both time required to obtain useful data from huge amount of unstructured data . This research will highlight & study all primary sources carefully and helps in understand few methods/technique's for text mining.

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

A. INTRODUCTION

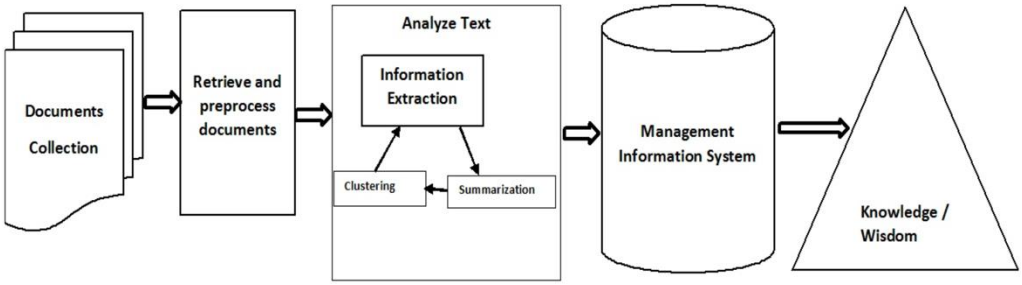
Data is rapidly expanding every day. all sorts of agencies, and corporate world use electronic data storage. A tremendous amount of text is pouring across the web of electronic databases, journals, and some other content such as weblog, Facebook, and e-mails. Locating useful trends to extract useful data from such a huge amount of information is a challenging. Data is difficult to mine with normal mining techniques since knowledge discovery takes mental energy.



Text mining is a technique for discovering knowledge by identifying meaningful and informative trends from text data sources. Data mining incorporates information gathering, data analysis, deep learning, statistics, and sentiment analysis. A graphic depicts text analytics and its relationships with various fields.
To extract knowledge, data gathering tools such as summary, categories, and gathering can be used. Text mining is the study of natural language text in quasi and unstructured formats. mining techniques are used in many fields, including Large industry, academia, application forms, the online, and others-uses.
In sectors such as search engines, customer relationship management systems, filter emails, product suggestion analysis, fraud detection, and social media analytics, text analytics is used for information extraction, extraction of features, prediction, and market analysis.

The basic mining method performs following steps, which are available in a variety of file types such as text format, web sites, pdf documents, and so on.

• To discover and eliminate irregularities, pre-processing and cleansing activities are conducted. The cleaning procedure ensures that the real substance of the text accessible is captured and is conducted to eliminate words stemming
• Automatic processing is used to inspect and clean the data set using processing and monitoring operations.
• Management Information System performs pattern analysis (MIS).
• The information derived from the preceding procedures is utilised to extract useful & relevant data for effective decision making spontaneously



Extraction of essential database queries of numerous documents is a difficult and time-consuming task. The time it takes to identify important similarities for judgement is reduced when a suitable text mining technique is used. The purpose of this research is to look at a few mining algorithms that can help with text classification on large datasets. Also noted are issues that arise throughout the mining process.

B. REVIEW OF ARTICLE

Object recognition, design choice, and research are all steps in the extraction procedure. In addition, the use of text mining techniques including such grouping, categorization, and classification tree categorization in diverse fields is examined. highlights the difficulties regarding text mining methods and techniques. They highlighted about handling the unstructured data is more challenging than working with structured or collections of data when using standard mining tools and procedures. They showed how feature extraction may be applied in fields such as biology, database management, and national defense.Text mining difficulties have been decreased thanks to advances in nlp and entity recognition tools. However, there are many

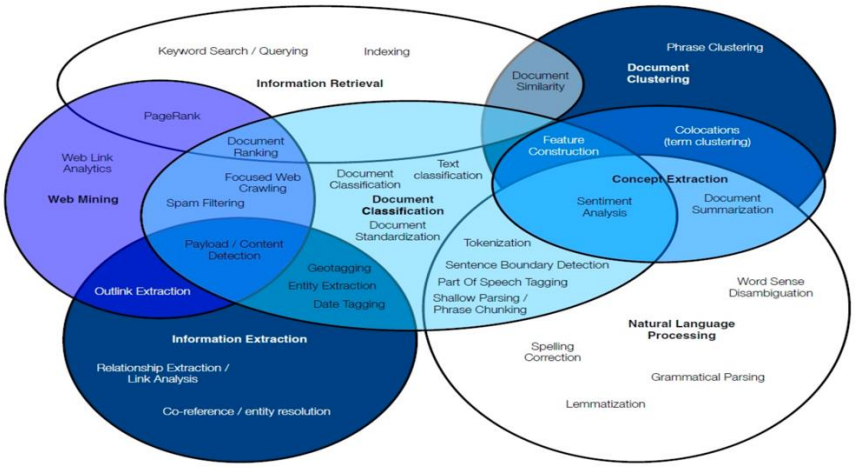
concerns that require care. In the biomedical database, created a system for object recognition, text classification, hypothesis generation and verification, connection and word retrieval, and abbreviation extraction. This new framework aids in the removal of irrelevant details and the extraction of useful data. using text mining patterns to examine the text and discovered that term-based techniques fail to correctly analyse synonyms and vocabulary. A concept design for classification tasks in terms of ranking sequences depending on their distribution also was built. This method helps improve the text mining process' efficiency. described a text mining-based crime detection system that used a link discovering algorithm to associate terms with spellings.

For web-based text mining, a top-down and bottom-up method was provided. To combine related documents, they employ the k-mean clustering algorithm for bottom up splitting. The TF-IDF (Term Frequency- Inverse Document Frequency) method was used to find information about certain concerns within the document. mining applications, approaches, and challenges were discussed. They discussed about how papers can be structured, semi-structured, or unstructured, and how extracting relevant information is a time-consuming process. They came up with a conceptual mining paradigm that can be compared to textual enhancement and information compression. The intermediate form of object description extraction is used based on the topic.

provided novel and effective pattern discovery methods They improved the effectiveness of finding same and appropriate data by using pattern developing and finding strategies. To measure the efficiency of the recommended technique On data from the network dataset publication and the text retrieval conference, they employed BM25 and vector support machine based filtering. Conducted a number of classification studies on text using multi-word characteristics. They suggested a method for manually extracting multi-word features from a data source. To categorise and retrieve multi-word text, they divide it into nonlinearity polynomial forms, which makes the extracted data more successful.

C. REFLECTIVE PROCESS

To investigate structures and the mining operation, several mining approaches can be used. The interconnections of mining methods and its basic functions are shown in a Venn diagram. Information to the right people (relationship extraction / link analysis), text grouping (clustering), natural language processing (spelling correction, tokenisation, grammar processing, and sentiment classification), and network analysis (web link analysis)



A. Data Extraction

Information extraction is applied to automatically extract useful information from unstructured and semi structured text documents such as PDFs, webpages and text files. Experts define properties and connections unique to their area. Specific characteristics and organisations are extracted from the document using IE systems, and their relationships are established. The extracted corpus is saved in a database to be processed afterwards. The precision - recall procedure is used to analyse the value of results on the extracted data. In
order to carry out the information extraction technique and achieve more relevant results, detailed and comprehensive knowledge of the relevant subject is necessary.

B. Data Retrieval

Information retrieval is a process that returns the information that is relevant for a specific query or field of interest. Text information retrieval and text mining are tightly related. IR systems use numerous algorithms to monitor user actions and accurately locate relevant material. Microsoft and Bing search engines are increasingly using information retrieval methods to find valuable content based on the a search term. Some search engines compare and provide more relevant results using query-based algorithms. These search engines give customers more timely information that meets their requirements.

C. Natural Language Processing

This uses a unstructured text analysis and processing approach. It uses Named Entity Recognition to perform a variety of analyses, include abbreviation and synonym retrieval to uncover relationships between them (NER). NER extracts all possibilities of a particular object from a set of retrieved documents. In order to realise their primary notion, these independent and examples allow the spotting of connections and other data. This technique, on the other hand, lacks a comprehensive vocabulary list for all named entities that must be recognised. To achieve satisfactory results, complex query-based algorithms must be applied. A single entity might have multiple names in the real world, such as Broadcast and Media.

Using classification algorithms, a set of repeated words may contain multi-word domains to recognise the borders and address over-lapping difficulties. Lexical, rule-based, statistical, or a combination of these approaches to NER can be found. The significance of NER systems has increased from 75% to 85%. The co-referencing method is often used in natural language processing to extract synonyms and abbreviations from textual input. Natural Languages is complicated by the fact that content from different sources does not

use the same vocabulary or acronyms (NL). Such problems must be identified, and processes for detecting them must be devised. For example, in order to extract and identify a person's function in an organisation, NER and co-referencing techniques construct a logical relationship.

D. Clustering

Clusters is an unstructured strategy for classifying text resources into groupings that employs several classification algorithm. Similar phrases are grouped together and extracted from documents. There are two methods for clustering: top-down and bottom-up. Natural language processing uses a number of text mining techniques & approaches to explore unstructured text. Clustering techniques include hierarchical, distributions, concentration, centre, and k-mean clustering, among others.

E. Text Summarisation

Text summarisation is a mining tool that assists in determining if a article meets the user's needs and is useful in obtaining more relevant information. Text summarising software decrease the length and complexity of a file/document while maintaining its major facts & real message. To summarise a book, a human must first read the complete content and have a thorough comprehension before writing a summary that highlights the essential elements. Because a computer lacks the ability to communicate like a person, it must rely on other means. Most of the pre-processing approaches include tokenisation, stop word deletion, and stemming. Language lists are established during textual summaries step. Text summarisation formerly required the existence of a specific phrase of words in the document.. To enhance the validity and validity of the results, new text analysis methodologies were combined with the standard text mining process. The balanced heuristics approach uses specific criteria to extract information from text documents. Article size, fixed sentence, paragraph, thematic term, and capital letter word recognition can all be incorporated and studied in text. Text summarisation is the act of collecting and generating brief summaries of source texts.Textual summary techniques can be applied repeatedly on a variety of texts. The quality and kind of classifiers are affected by the nature or theme of a text documents.

D. APPLICATION OF TEXT MINING

A. Online Libraries

Huge range of text mining methodologies and techniques are used to detect relationships and correlations in journals from a huge number of data and sources. In the area of research and development, many information sources are beneficial. Libraries are an important source of information for academics, and library services are aiming to improve the quality of their resources. It offers a new method to information organisation that allows trillions of documents to be viewed online. It provides a novel data organisation strategy as well as internet access to thousands of documents. The GreenStone international online library offers a versatile technique to extract materials in a variety of forms, like Word documents, PDF files, Postscripts, HTML and more. It can also extract documents in multimedia and graphic formats in addition to text documents. Text mining involves a number of activities, including doc selection,IR, handling entities between articles, and developing natural summarising.
In digital libraries, text mining programmes like GATE, Net Owl, and Aylien are often utilised.

B. Research Field and Academic

For education, various mining technologies and different ways are used to investigate fresh learning in a certain place, students engagement in a particular fields, and career.
In the research world, text analytics is used to find & categorize research articles and useful content across various fields inside one place. The use of algorithms such as k-means clustering and others aids in the identification of significant properties. It is possible to look at how students perform on various topics and how various variables influence subject selection.

C. Online Media

Text processing software's available to track and evaluate normal language from web, such as newspapers, weblog, and email, for monitoring social networking apps. The volume of tweets, likes, and friends on an online community can be identified and analysed using text mining tools. This form of research reveals how people reacted to various news, as well as how it travelled to all. this depicts the conduct of individuals of certain age community or group who hold similar and opposing viewpoints on the same issue.

D .Business Intelligence

Text analytics is an important aspect of business information because it allows firms to better understand their customers and competitors. It provides a better understanding of company as well as information on how to boost satisfaction of customer and gain advantage. Mining technologies like IBM text processing, Quick miner, and Gate help people make better decisions by generating patterns of good and bad performances, market transitions, and remedial actions. The telecom sector, commercial and trade applications, and client chain control systems all benefit from it.

E. CONCLUSION

In order to obtain valuable information, the existence of a vast quantity of data should be studied. Data mining tools are used to effectively and efficiently retrieve interesting and important information from massive amounts of uncertain data. This article provides a overview of textual mining methods that can aid in the process improvement. Particular behaviour and sequences are employed in predictive analysis to recover most useful information by deleting features. Text mining is way-forward and useful when the accuarate techniques and tools are picked and employed according to the data used . Key problems & challenges that develop during mining process include domain specific integration, changing concept resolution, multiple languages textual refining, and nlp is difficulty. We will concentrate more on future research.

REFERENCES

1. [1] R. Sagayam, A survey of text mining: Retrieval, extraction and in- dexing techniques, *International Journal of Computational Engineering Research*, vol. 2, no. 5, 2012.
2. [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. [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. [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. [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. [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. [7] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," *Copy at http://j. mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper*, vol. 456, 2014.
8. [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. [9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicˇius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," *Journal of biomedical semantics*, vol. 5, no. 1, p. 1, 2014.
10. [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. [11] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sci- ences*, vol. 275, pp. 314–347, 2014.
12. [12] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," *International Journal of Computer Applications*, pp. 159–171, 2013.
13. [13] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applica- tions, tools and issues-an overview," *International Journal of Computer Applications*, vol. 80, no. 4, 2013.
14. [14] P. J. Joby and J. Korra, "Accessing accurate documents by min- ing auxiliary document information," in *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on*. IEEE, 2015, pp. 634–638.
15. [15] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in *Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan*, vol. 51, 2007, p. 45.
16. [16] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.
17. [17] R. Agrawal and M. Batra, "A detailed study on text mining techniques," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp. 2231–2307, 2013.
18. [18] D. S. Dang and P. H. Ahmad, "A review of text mining techniques associated with various application areas," *International Journal of Science and Research (IJSR)*, vol. 4, no. 2, pp. 2461–2466, 2015.
19. [19] R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," *Language Resources and Evaluation*, vol. 46, no. 2, pp. 155–176, 2012.

20. [20] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.
21. [21] E. A. Calvillo, A. Padilla, J. Mun˜oz, 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.
22. [22] B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," *IJSEAT*, vol. 3, no. 3, pp. 69–71, 2015.
23. [23] B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis based document summarization," *International Journal of Computer Science and Information Security*, vol. 14, no. 4, p. 145, 2016.
24. [24] R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." *Int. Arab J. e-Technol.*, vol. 1, no. 4, pp. 164– 168, 2010.
25. [25] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library," *International Journal on Digital Libraries*, vol. 4, no. 1, pp. 56–59, 2004.
26. [26] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," *Europen Journal of Scientific Research*, vol. 43, no. 1, pp. 24–29, 2010.
27. [27] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostro¨m, "Ensembles of randomized trees using diverse distributed representations of clinical events," *BMC Medical Informatics and Decision Making*, vol. 16, no. 2, p. 69, 2016.
28. 28] I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach," *Expert Systems with Applications*, vol. 44, pp. 386–399, 2016.
29. [29] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and com- putational biology*, vol. 3, no. 02, pp. 185–205, 2005.
30. [30] Y. Zhao, "Analysing twitter data with text mining and social network analysis," in *Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013)*, 2013, p. 23.
31. [31] F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha, "Impact and usage of internet in education in pakistan," *European Journal of Scientific Research*, vol. 47, no. 2, pp. 256–264, 2010.
32. [32] R. Sharda and M. Henry, "Information extraction from interviews to obtain tacit knowledge: A text mining application," *AMCIS 2009 Proceedings*, p. 283, 2009.
33. [33] H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," *International Journal of Computer Applications*, vol. 75, no. 16, 2013.
34. [34] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information:-extended abstract," *OTT06*, vol. 1, p. 55, 2007.
35. [35] A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure, "Text analytics for android project," *Procedia Economics and Finance*, vol. 18, pp. 610–617, 2014.
36. [36] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in *Proceedings of the World Congress on Engineering*, vol. 3, 2013, pp. 3–5.