International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-12, Special Issue-1, March-2024
https://doi.org/10.55524/CSISTW.2024.12.1.49
Computing for Sustainable Innovation: Shaping Tomorrow's World (CSISTW-2024) on 16th-17th February, 2024
Organized by B.N. College of Engineering and Technology, Lucknow, India

Text Mining: Techniques, Applications and Issues

Subhash Chandra Gupta[1], and Noopur Goyal[2]

[1]Research Scholor, Department of Computer Applications, V.B.S. Purvanchal University , Jaunpur, India
[2]Professor, Department of Computer Applications, V.B.S. Purvanchal University , Jaunpur, India

Correspondence should be addressed to Subhash Chandra Gupta; csubhashgupta@gmail.com

ABSTRACT- Day-to-day transactions made on digital world, produces a large volume of data which contains some valuable information. But they are not in visible but present in hidden form. These information are extracted by applying mining techniques considering the type of data source. Since most of the data sources are either unstructured or semi-structured nature and available in the form of text such as web pages, CSV files, PDFs, text files and rich text format files. Text mining techniques is used to discover the hidden information, called patters/ trends, from these unstructured text documents. Text mining has become an exciting field of research, and enhances its extraction capability with NLP, AI and machine learning algorithms. In this research paper, the general concept of text mining is discussed along with the procedure of text mining. A focus is also given on the techniques used for text mining and highlighted the application area of text mining. In last of the paper, a brief discussion is made on the challenges and issues of text mining too.

KEYWORDS-Text mining, Text Summarization, Classification, Information Extraction, Information Retrieval, Knowledge Discovery, Applications.

I. INTRODUCTION

It is impossible to imagine a life without internet since it makes life easy for the people for different operations- online shopping is one of them. Making an online order of a product is much easier in comparison to choose correct product for shopping on the net. Most of the people study the reviews of buyers to buy a product; but it is not an easy task to study each review, which are in thousands of numbers. Text mining techniques has made it easy for the buyers; using text summarization method, thousands of reviews may be concised just in few sentences[1].It is not a single example where such type of information extraction process are needed there are so manyarea where text mining becomes a necessity to extract or find hidden information from textual contents.Data mining is the concept of applying different techniques to find valuable information/ patterns/ trends from the storage of large volumes of data. Most of the data source applied for data mining are structured documents such as spreadsheet, database files etc. or in available in highly structured format[2], [3]. In the era of Internet, maximum number of commercial and noncommercial organizations, institutions and industries preserve their information/ data in digital form. Every day, million megabytes of data are added in the digital world through day-to-day transactions of these organizations. A large amount of data are stored in unstructured textual form and circulates on the internet in the form of blogs, emails, SMS, newspapers, journals, product reviews, articles, repositories, digital libaries. Approximately three-fourth of world data is available in unstructured text format[3].Identifying the patterns/trends, hidden in these textual documents, become a very difficult task. Simple data mining techniques are unable to extract valuable information from the ocean of data and takes large amount of time and efforts too. In such scenario, a new mining technique called text mining / text data mining is used[4]–[6]. Text mining is the process of extracting valuable information or identifying patterns and new insights from text documents available in the ocean of digital world. Text mining can be named in a variety of words such as text data mining, knowledge discovery, analysis of intelligent text, extracting valuable information from unstructured text etc.[3]. It is also considered as the extension of data mining process from unstructured text databases and discovered unidentified valuable information, which is not extracted earlier. Knowledge extraction[4], [7], [8] is not an easy task. Several techniques like text summarization, classification, information retrieval, clustering etc. are applied to get valuable information from textual content. First, the objective of the given text mining task is defined to meet the financial/ social/ scientific goals of mining. The data source may be available in structured or unstructured way in the form of text files available in different formats such as plain text files, web pages, PDF files etc. These text files are collected, and preprocessed by detecting and removing anomalies. Data cleaning is also required to remove stop words and get the real essence of the text.In the next step, patterns are identified and analyzed by applying management information techniques. The final stage, the extraction of valuable or hidden information, is applied for effective decision-making and trend analysis[7][8][9].Extracting the valuable/ hidden trends from a large number and different types of text document is a very tedious and tiresome task. Time and effort may be reduced by selecting suitable text mining techniques for analysis and decision-making. Text extraction techniques mostly processed text documents available in natural language of different sources and used in a number of application areas[10][11].In the research paper it is focused

on different text mining techniques, its application in various area and challenges and issues.

II. LITERATURE REVIEW

R. Boorugu et al[1] has developed a text mining model using natural language processing techniques, to extract information (product reviews) form the messages of customers and summarize them. In this paper they made a brief discussion about the prominent method s of text summarization and text mining. S. Dang et al[3] has made a discussion about the concept of text mining and steps perform to extract information from unstructured text documents. They have also mentioned the major techniques related to text mining and focus on its application area and challenges. H. Hassani et al[4] has emphasize the role of text mining in big data analytics to extract information from semi-structured or unstructured data sources. They have also made a review of a number of papers related to the role of text mining in analysis of data available on net. [5][6]–[8] and [9] has made discussion about the theoretical aspect of text mining and its process in their papers. The techniques such as NLP, information extraction, information retrieval and text summarization, are discussed in detail. They have also indicated the challenges and issues before the text mining in coming in future.

III. DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

In some occasion data mining and text mining are referred interchangeably to a process to extract valuable information from different sources and converted into a knowledge[7], [10]. But it is not true; data mining is a broad term in comparison to text mining and it refers as an umbrella to all mining techniques such as classification,association, clustering, pattern analysis and many more too. Text mining technique is one of them. Text mining and data mining are technically different things. The term data mining is used for the extraction process when data sources are structured and available in a well format (structured or semi - structured format) such as tables, spreadsheets, CSV files, ERP, CRM, database files and accounting information [4]. These data files contains data mostly in numbers or available in such a format which can be easily converted into numeric forms. Data mining techniques are applied to process these billions of numbers and extracted information, usually in statistical nature, called knowledge / patterns or emerging trends[3]. While text mining is the process of extracting information from unstructured or semi-structured data sources which are usually contains textual material. More than two-third portion of whole available digital data are in text format and required mining techniques such as text mining to extract information. Text mining, which can be said as the extension of data mining too, uses machine learning, AI and Natural language processing to automatically collect statistical and meaningful information for the processing[12]. Text mining is the techniques which relates computer languages, data mining and knowledge management. In simple words one can say that if patterns are belongs from structured databases (non-textual) then standard data mining techniques are applied otherwise a new approach, called text mining, is used to extract these valuable information from unstructured data source (textual data source)[5].

IV. STEPS BEHIND TEXT MINING

Text mining is a multidisciplinary approach consists of techniques belongs from data mining, machine learning, Natural Language processing, information retrieval, statistics and computational linguistics[2]. The following steps are performed in text mining-

Collection of unstructured data:

Unstructured data files are collected from different sources available in different formats such as plain text files, web pages, PDF files etc.

Preprocessing of unstructured data:

For the detection and removal of anomalies, unstructured data are preprocessed and cleaned. Cleaning is applied to capture the real essence of the text and removal of redundancies, independent words, and inconsistencies and stop words. The following stages are done during this phase-

- ***Tokenization***

The document is seen as string and split into units or tokens by recognizing words[7].

- ***Removal of stop words***

Constant words such as a, an, but, and, of, the etc., are removed during this step[5] [10] [13].

- ***Stemming***

The words having similar meaning are called stem. In this step these type of words are recognized[10], [14]. This approach defines the base of a specific word.

Identify the patterns: Patterns are identified by applying auditing and cleaning of data by automatic processing.
Analysis of patterns: Pattern analysis is implemented by Management Information System (MIS).Extraction of valuable information: the final step is applied to extract valuable or hidden information for effective decision-making and trend analysis.Steps of text mining is shown in following figure fig 1.

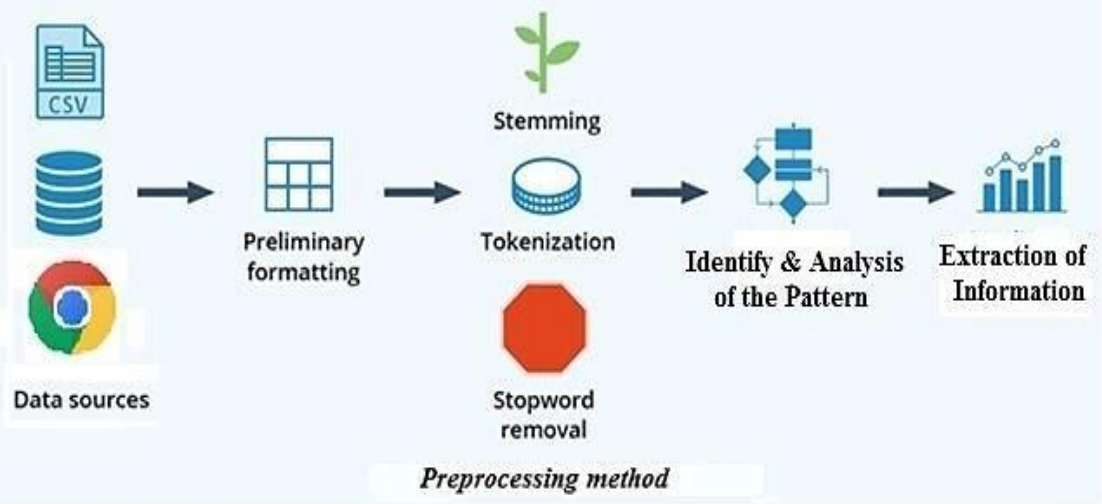

Figure 1: different phase of text mining process

V. TEXT MINING TECHNIQUES

Text Mining is an interdisciplinary approach, applied to analyze patterns hidden in text and their extraction process, includes information retrieval, information extraction, text classification, text summarization, document classification.

A. Information Extraction

Information extraction[3], [9] is the process applying to extract automatically useful information form non-structured files such as web pages, PDFs and textdocuments. It can also be describe the transformation of unstructured information in structured representation for selected information drawn from the text. The goal of IE to find the key information from textual contents, therefore in it, the inputs and its extraction targets are defined clearly.[3][7].The inputs may be an unstructured files written in natural language or semi-structured documents available on the web. IE uses NLP to extract information from textual contents. To find the targeted information for extraction, IE apply a sub-tool called named entity recognition (NER)[15]. It recognizes the category of entities either as a location (loc), persons or organizations (org). Once this task is completed, an IE utility extracts the related information about named entity and constructs a machine readable documents for further processing. IE finds meaning by way of other subtasks which included relationship extraction, co-reference resolution, language and vocabulary analysis and sometimes audio extraction.Information extraction is applied in large capital companies and business organization for invoice automation, KYC automation, and in financial investigation and healthcare systems[16].

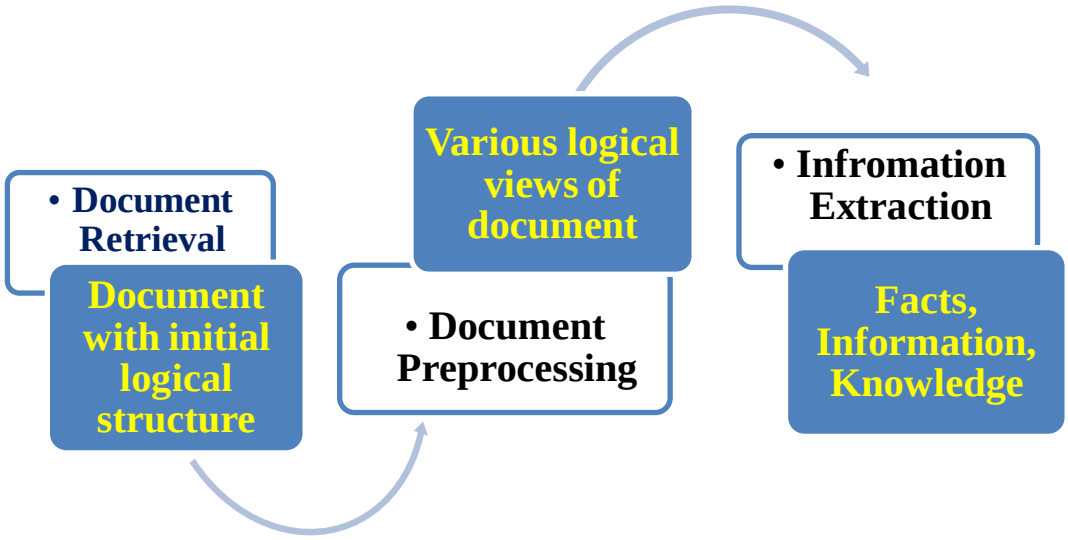

Figure 2: Steps of Information Extraction (Activity – White boxes, Outputs- Blue boxes)

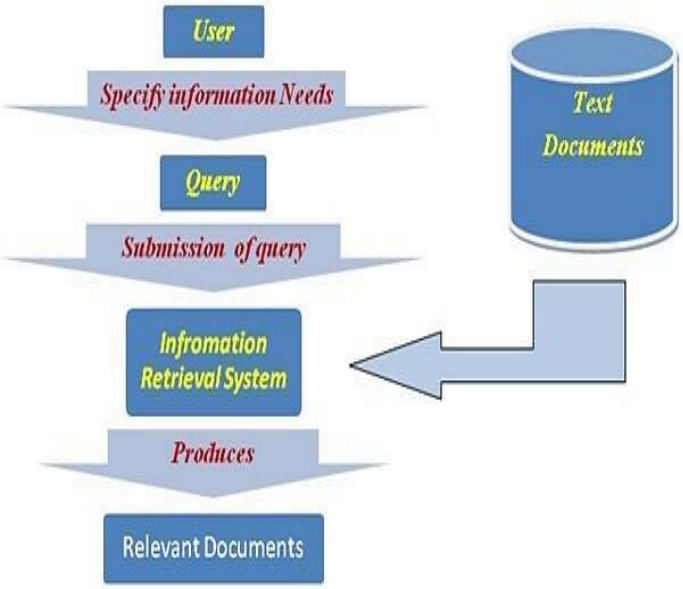

Figure 3: Steps of Information Retrieval

B. Information Retrieval

Information retrieval is applied when desired information is based on a specific field of interest or query and stored as textual contents in document files[3]. The example of Information retrieval (IR) processes are search engine, such as Google or Yahoo search engines, which recognize those web files which are related to or contains a set of given words[7]. The search engine provides the more relevant and useful information to the user, which satisfy their needs. IR is also considered as the process of extension of document retrieval, where returned documents are useful for users to extract crucial information from it. Document retrieval process is followed by text summarization stage based on a query imposed by the user or obtained from information extraction stage. Text mining and information retrieval is closely related for textual data. The major difference between Information extraction and information retrieval is that information retrieval is related with the task to automatically extract structured information form machine readable semi-structured and /or unstructured documents; while information retrieval is related with the finding of those materials/ documents of an unstructured nature (in textual nature) that fulfill the needs of information, from large collection of text files stored on computers[5], [8], [10].In Figure-3 the detail process of information retrieval is given.

C. Natural Language Processing (NLP)

Natural language processing (NLP) is a branch of artificial intelligence to provide computers the ability to understand textual content written and spoken in the same way as human beings can. It combines statistical, text mining, machine learning and deep learning models with computational linguistics (rule based modeling of human language) [10]. Along with these technologies, a computer is enable to process text or voice of human language and understand the full meaning of it, with the intent and sentiment of write or speaker. NLP also develop the ability in computer to respond the spoken commands, translate the given text from one language to another language and summarize the large volume of text even in real time [7].

D. Document Clustering

Document clustering is an unsupervised learning process which put a text document in the closer cluster/ group by applying clustering algorithms. A group, called a cluster, contains object which have similar patterns and dissimilar from the objects of other groups. A cluster of documents is a group containing those text documents from where similar patterns are extracted using different clustering techniques such as distribution, density, hierarchical, centroid and k-mean[8]. The objective of document clustering is to maximize the inter-cluster distance (distance measures) while minimizes the intra-clustering (similarity measures) distance amongst the documents. The application areas of document clustering are many more such as Finding Similar Documents, Organizing Large Document Collections, Duplicate Content Detection, Recommendation System and Search Optimization[10].

E. Text Summarization

by applying a software, Automatic text summarization is the process of shortening a text material obtained from one or more than one documents and producing a summary containing major points of the original documents[17]. The automatic summarization techniques enable a human to read the content of files, understand it and produce a concise material containing major points of the documents in very short time. The research processes, in automatic text summarization, have been done in past decades, but still a large scope of research to develop a system that is able to summarize the text like human.Text summarization is a mining process that helps the user to summarize the contents belongs from a single or multiple text documents and provides important and relevant information. Summarization of contents refers to the reduction of the length and/or details of a document without losing important information and maintains its key points and overall meaning[3], [17]. In general a human first read the

content of a document and develops his understanding about it based on some key points and then writes a summarized view which contains the relevant and important information according the key points. Textsummarization software performs the same operation using computer on a broad level.Text summarization is done either through Extractive approaches or Abstractive approaches[7]. In extractive approach one summarizes his text using simple algorithms called frequency count method which store all important words and their counting / occurrences in documents. The final summarized content contains the words/ sentences which has high frequency. Abstractive approach is more advanced in comparison to the extractive approach and uses sentence extraction tool. The sentence extraction tool extracts important sentence from the text document and statistically weight those sentences. Further it uses heuristic approach to get the final summarized view of document[12].

F. Document classification

Document classification is related with the categorization of digital documents available on the web. Today, it is difficult to categorize manually all digital documents available in the digital world due to their huge size. So a technique, which is based on text mining, machine learning and artificial intelligence, is used to categorize these documents automatically called document classification[18]. Document classification techniques assigns categories or classes to digital documents, which makes them easier to search, filter, manage and analyze.

Document/Text classification is a supervised machine learning method which placed the document in specific labels according to its content. According the content available in document it assigns one or more class labels[10]. There are a number of application of document classification in different areas like media art galleries, web pages, library books, emails such as email routing, spam detection, sentiment analysis and many more. Google's Gmail Spam detector, categorization of email considering the content of it, is one of the best examples of document classification[7].

G. Web Mining

The estimation of the size of web document available on WWW is a very difficult task since its size is increasing rapidly every day. An estimate came in the last decade, says that it is about 350 million pages and everyday one million new pages are adding in it. These web pages are depository of unexplored information since extraction of information is a tedious work due to its textual nature. Web mining is the activity of mining of specific information or patterns present in the large collection of web pages available on internet. Web mining is the integration of different mining technology such as data mining, text mining, computational linguistics, machine learning and AI[3], [4]. Web mining produces its outputs in terms of web pages which include textual, graphical and visual data. Considering the target to analyze, web mining is categorize into web structure mining, web content mining and web usage mining.

VI. TEXT MINING APPLICATIONS

In their day-to-day transactions, Organizations produce a large volume of digital data in structured / semi-structured and/or unstructured documents containing valuable information required necessarily to increase their future aspect. These information / patterns are hidden within textual contents of web pages, text files and emails. An organization is incapable to directly extract information from it. Text mining and its supporting techniques provide the ability to extract these patterns and convert it into structured format. The major text mining application areas are IT sector, banks, research, healthcare systems, political analysis and in media[3], [4], [16]. Some text mining application are given below:

Detection of spam mails

spam filtration is the process of detecting unsolicited emails for an individual recipients. A machine learning model based on text mining is used to solve this problem to categorize the email as a spam or nonspam. It is also an important objective of spam detection is to extract the way by which an email is classified as spam and used it for the purpose of categorize the mail in future [18][4][10].

A. Customer Relationship Management

The responsibility of a customer relationship manager is provide the response of a client's query, messages or complaints as soon as possible. Using the text mining technique, response time may be reduced by analyzing their queries [4], [7].

B. Competitive Intelligence

Competitive intelligence which is also referred as corporate intelligence, refers to provide the ability to computer to gather, analyze and use the information about customer's choice, competitors and market trends which will contribute in the company's strategies in future[7].

C. Classification of NEWS as Text

text mining is also helpful to search a specific news headlines / web pages/ pdfs based on a key points. Performing the same task manually will be hectic and time consuming process. Google News service is an example of it[7].

D. Classification of Scientific Documents

To find the article related to the scientific interest, text mining techniques like clustering, regression analysis and classification are applied[3], [5]. By using regression analysis, classification rules are generated to cope up with this problem.

E. Sentiment Classification

Sentiment analysis is an example of application of the use of NLP and text mining to sense the mood, emotions or anger of people on a specific event or product by studying the social media messages[19]. Text mining extract the information (state of mind) hidden in messages and classify them as positive, negative or neutral sentiments. Sentiment analysis is, a process of automatically extracting opinion of people from their messages, also known as Opinion Mining[4], [8], [10].

VII. ISSUES IN TEXT MINING FIELD

Text mining is a powerful tool to extract patters/ trends/ information from textual contents of documents, but it also have many problems or challenges. Ambiguity in words of text, availability of noise (irrelevant words), language barriers and the availability of large volume of text documents are of a few problems associated with text mining[16]. Ambiguity of words, is related with the different meaning of same word in different sentences in text depending on the context. Ambiguity of words may be either semantic ambiguity, syntactic ambiguity, lexical ambiguity, or pragmatic ambiguity. To overcome this problem, text mining may be applied with NLP along with fuzzy set theory and knowledge regarding the context to lexical semantics[7][16].Noise in the text may be like typo error, spelling mistakes and use of non-standard abbreviations. To manage noise in text, use of spell-checking application and regular expression make clean and normalize data before applying text mining[18], . Availability of text document in different language is another big challenge to extract information from multilingual datasets during text mining. The use of NLP libraries such as the spaCy library that support more than 50 language support and provide pre-trained models, reduces the language barrier problem and address the issue in a better way[4], [10], [16].In digital world, text documents are available in large volume which makes it more tedious and challenging task to extract information promptly from it during text mining. The use of Parallel processing and distributed computing may reduce the challenges of large volume of text documents during processing and analyzing[4], [10]. It distribute the processing load of text mining on many nodes. Some other issues of text mining are the processing cost and integration of domain knowledge.

VIII. CONCLUSION

Text mining is an extension of data mining process to extract information from unstructured or semi- structured text document. Data mining is applied for the extraction of information from structured files while text mining is for semi-structured or unstructured documents. Text mining is also an intra-disciplinary approach works as an umbrella for different techniques such as information retrieval, information extraction, document classification, and clustering and sentiment analysis, which help to increase the processing capability of text mining. The major application are of text mining are spam detection, sentiment analysis, fraudulent detection, customer relationship management, text summarization, review analysis, R&D work in the fields of science, technology and healthcare.

- ***Issues and challenges of Text mining***

Along with the opportunities provided by text mining, there are some issues and challenges with this techniques. Some prominent issues of text mining are of language barrier, availability of multi-lingual text documents, presence of ambiguity and noise in documents, large volume of available data source and cost of extraction processes. These issues may be overcome by using suitable techniques and supports during text mining.

- ***Limitation and future scope of paper***

The paper provides an emphasis only on the concepts of text mining, approaches, techniques used during mining, its application areas, challenges and issues. A number of researches in this area has been made in the past by researchers but still there exists some scope of further researches in future. The future scope of the paper is to be built a model on real data to apply the text mining process and analyze its performance.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020, pp. 352–356, 2020, doi: 10.1109/ICIRCA48905.2020.9183355.

[2] A. T. Sadiq, "English Text Summarization Using Statistical , Linguistics , Heuristics and Machine Learning Techniques," vol. 40, pp. 1–13.

[3] S. Dang and P. H. Ahmad, "Text Mining : Techniques and its Application Text Mining View project Text Mining: Techniques and its Application," IJETI Int. J. Eng. Technol. Innov., vol. 1, no. March, 2014.

[4] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," Big Data Cogn. Comput., vol. 4, no. 1, pp. 1–34, 2020, doi: 10.3390/bdcc4010001.

[5] J. A. Atkinson A, "Text mining: Principles and applications," Rev. Fac. Ing., vol. 7, no. Enero-Junio, pp. 57–62, 2000.

[6] N. Zanini and V. Dhawan, "Text Mining: An introduction to theory and someapplications," Res. Matters A Cambridge Assess. Publ., vol. 19, pp. 38–44, 2015.

[7] N. Goel, "A study of text mining techniques: Applications and Issues," Pramana Res. J., vol. 8, no. 12, pp. 307–316, 2020.

[8] Y. Zhang, M. Chen, and L. Liu, "A review on text mining," Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2015-Novem, pp. 681–685, 2015, doi: 10.1109/ICSESS.2015.7339149.

[9] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, pp. 414–418, 2016, doi: 10.14569/ijacsa.2016.071153.

[10] A. Kołcz, "Text Mining for Spam Filtering," Encycl. Mach. Learn. Data Min., pp. 972–977, 2017, doi: https://doi.org/10.1007/978-0-387-30164-8_828.

[11] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," R D Manag., vol. 50, no. 3, pp. 329–351, 2020, doi: 10.1111/radm.12408.

[12] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," SN Comput. Sci., vol. 4, no. 1, 2023, doi: 10.1007/s42979-022-01446-w.

[13] J. Benchimol, S. Kazinnik, and Y. Saadon, "Text mining methodologies with R: An application to central bank texts," Mach. Learn. with Appl., vol. 8, no. March, p. 100286, 2022, doi: 10.1016/j.mlwa.2022.100286.

[14] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," Alexandria Eng. J., vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.

[15] S. M. Patel, "Extractive Based Automatic Text

Summarization," J. Comput., vol. 12, no. 6, pp. 550–563, 2017, doi: 10.17706/jcp.12.6.550-563.
[16] D. Madigan, "Text Mining : An Overview," 2014.
[17] N. Andhale and L. A. Bewoor, "An overview of text summarization techniques," Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016, 2017, doi: 10.1109/ICCUBEA.2016.7860024.
[18] M. T. Verma* and D. N. Singh Gill, "Email Spams via Text Mining using Machine Learning Techniques," Int. J. Innov. Technol. Explor. Eng., vol. 4, no. 9, pp. 2535–2539, 2020, doi: 10.35940/ijitee.d1915.029420.
[19] J. Benchimol, S. Kazinnik, and Y. Saadon, "Text mining methodologies with R: An application to central bank texts," Mach. Learn. with Appl., vol. 8, p. 100286, 2022, doi: 10.1016/j.mlwa.2022.100286.

AUTHOR'S PROFILE

Subhash Chandra Gupta is a research scholar in the Department of Computer Applications of V.B.S. Purvanchal University, Jaunpur. He has received his M.C.A. degree from Indira Gandhi National Open University, New Delhi. His research area is Data mining. He has published some articles in international journals and conferences. His interests include data mining, python, data structure and algorithms.

Dr. Noopur Goel is currently working as a Professor in the Department of Computer Applications, V. B. S. P. U. Jaunpur, and have more than 20 years of teaching experience. She has achieved her Ph.D. (Computer Science) from BHU, Varanasi. Her research interests include Software Testing, Software Reuse. Dr. Noopur Goel is the member of many international affiliations. She has successfully completed her Administrative responsibilities. She has authored of many research articles/books related to Software Testing, Software Reuse.