Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 4, August : 2023

A REVIEW ON TEXT MINING TECHNIQUES

Dr.S.Rizwana[1],Assistant Professor and Head, Department of Computer Science ,Erode Arts and Science College,Erode.
Mrs. N.Sasikala[2], Assistant Professor and Head, Department of Computer Science ,Erode Arts and Science College,Erode.

Abstract
The massive quantity of information or data deposited in amorphous texts cannot simply be used for next processing by computers, which classically maintain text as simple sequences of character strings. Therefore, specific pre-processing methods and algorithms are essentially in order to mine useful patterns. Text mining states generally to the process of extracting required information and knowledge from unstructured text. In this paper, we converse text mining as a young and interdisciplinary field in the intersection of the related areas information retrieval, machine learning, statistics, computational linguistics and especially data mining. We designate the main study tasks pre-processing, classification, clustering, information extraction, visualization and analyse and display a amount of successful applications of text mining.

Key words: Text Mining, Information Extraction, Information Retrieval, Natural Language Processing, Clustering, Text Summarization.

1. Introduction

Data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Text mining is the process of mining significant information or knowledge or patterns from the available unstructured text documents. Text mining can be categorized as text categorization, text clustering, association rule extraction, and text visualization. They are discussed in the following sub-sections (7) .It projected that up to 80% of professional data comprises of unstructured facts such as text. Text mining permits businesses to extract more valuable information from the unstructured text generated every day in email messages, social media posts, customer service tickets, chatbots and other sources. Without an automated process, it can be extremely time-consuming or even impossible to analyze all this information. Automatic processing of text documents can also produce more accurate and consistent information. Text mining can help businesses quickly discover and respond to problems in manufacturing or customer service, anticipate competitive threats and provide more personalized customer service.

Today the internet has massive amount of text in the form of digital libraries, repositories, and other textual information such as blogs, reports, reviews, news, social media network and e-mails. It is difficult task to find out appropriate patterns and trends to extract important knowledge from this large volume of data.[3], The financial segment generates a vast amount of data like customer data, logs from their financial products, transaction data that can be used in order to support decision making, together with external data, like social media data and data from websites[5].The following table 1 shows the Data Generated rate from various sectors during the year 2010 to 2023.

Table1: The Data Generated rate from the year 2010 to 2023

SNO	Year	Data Generated Each Year	Equate Above Forgoing Year
1	2010	2 zettabytes	-
2	2011	5 zettabytes	**3 zettabytes**
3	2012	6.5 zettabytes	**1.5 zettabytes**

4	2013	9 zettabytes	**2.5 zettabytes**
5	2014	12.5 zettabytes	**3.5 zettabytes**
6	2015	15.5 zettabytes	**3 zettabytes**
7	2016	18 zettabytes	**2.5 zettabytes**
8	2017	26 zettabytes	**8 zettabytes**
9	2018	33 zettabytes	**7 zettabytes**
10	2019	41 zettabytes	**8 zettabytes**
11	2020	64.2 zettabytes	**23.2 zettabytes**
12	2021	79 zettabytes	**14.8 zettabytes**
13	2022	97 zettabytes	**18 zettabytes**
14	2023	120 zettabytes	**23 zettabytes**
15	2024	147 zettabytes	**27 zettabytes**
16	2025	181 zettabytes	**34 zettabytes**

Table 2: **The Type of Data Generated rate**

SNO	Data Type	Data Generated %
1	Video	**53.72%)**
2	social media, TikTok	**67.45%**
3	Facebook	**51%**
4	Snapchat	**70.02%**

The above table 2 displays different types of data and rate. The Snap chat is the multimedia messaging Application; It is that pictures and messages are usually only available for a short time before they become inaccessible to their recipients. Highest Data's or Zeta bytes Data's are generated by Snap chat In 2020, the amount of data on the internet hit 64 zeta bytes. A zeta byte is about a trillion gigabytes. One way to estimate the size of the Internet is to look at the amount of information created, captured, copied, and consumed on the web. Streaming, downloading, and watching videos (YouTube, Netflix, etc.) and downloading or streaming music (Pandora, iTunes, Spotify, etc.) dramatically increases data usage.

In zeta bytes, that equates to 120 zeta bytes per year, 10 zeta bytes per month, 2.31 zeta bytes per week, or 0.33 zeta bytes every day. The amount of data generated annually has grown year-over-year since 2010.In fact, it is estimated that 90% of the world's data was generated in the last two years alone. In the space of 13 years, this figure has increased by an estimated 60x from just 2 zeta bytes in 2010.The 120 zeta bytes generated in 2023 is expected to increase by over 150% in 2025, hitting 181 zeta bytes. Video is responsible for over half (53.72%) of all global data traffic and social media is brimming with video content. TikTok is entirely based on videos and continues to grow its user base year-over-year. While Facebook has evolved to the point where 51% of content shared on the platform is video-based. Although public data related to Snap chat is limited, it is estimated that each snap sent requires 1MB, many of which are videos.

In this paper we describe text mining as a truly interdisciplinary method drawing on information retrieval, machine learning, statistics, computational linguistics and especially data mining. We first give a short sketch of these methods and then define text mining in relation to them. Later sections survey state of the art approaches for the main analysis tasks pre-processing, classification,

clustering, information extraction and visualization. The last section exemplifies text mining in the context of a number of successful applications.
Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß
A Brief Survey of Text Mining
Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß
A Brief Survey of Text Mining

2. Related Works

The analysis of the data with data mining algorithms can be supported by databases and thus the use of database technology in the data mining process might be useful [1]. The unstructured text documents from various source contains huge amount of information which are not to be used for any processing to extract useful information. Text mining is the process of extracting significant information or knowledge or patterns from the available unstructured text documents. Text mining tasks includes text categorization, text clustering, document summarization and sentiment analysis.[2]In the present time, text mining helps the business organization to analyse the vast text document and extract valuable data from this text. This study has adopted the secondary qualitative data collection methods and from this secondary data, this study has conducted the thematic analysis to find the authentic result [4]. Finacle Connect (2018) indicates the top 10 technologies for financial industries, including the rise of API economy, cloud business enablement; block chain for banking, and usage of artificial intelligence. In order to adapt to the economic environment, decision-makers across the public and private sectors require accurate forecasts of the economic outlook [6]. The second research question was answered by providing the analysis of techniques for text mining in the financial sector. Analysis of big amounts of data represents the transition to analytic-driven business, conducted by big companies, small enterprises or research teams, in order to identify significant information and transform it into new knowledge. Text analytics or text mining of big data, conducted by various techniques (keyword extraction, named entity recognition, gender prediction, sentiment analysis, topic extraction, and social network analysis) has moved from research centres to real-world institutions, such as financial and banking institutions. The third research question was answered by the analysis of data sources used for text mining techniques. Results revealed that most of the research focuses on external data sources, such as news and online media posts for the purpose of stock market predictions, and fraud detections. The number of research studies using internal data sources is low. Therefore, the utilization of internal data sources will be a rich source of future research with both theoretical and practical contributions. Various research using internal text sources, such as emails, corporate wikis, financial statements, and project reports could be useful for various purposes, such as human resource management, internal audit, and customer relationship management.

In addition, various multimedia files could also be the high-value additional component of text mining analysis (Pouli et al., 2015 [81]; Stai et al., 2018 [82]; Ma et al., 2011 [83]).The main limitation of our work is the usage of bibliometric approaches to the literature analysis,which has certain limitations. By selecting the database for studies search (Web of Knowledge), specific studies remain invisible to this analysis (Batistiˇc et al., 2017 [12]). Research results also generate several paths for future research directions. First, more up-to-date outlook to the usage of text mining in finance could be attained with the use of so-called “grey” literature sources, such as case studies, corporate reports, and text-mining software projects (Adams et al., 2017 [84]). Second, usage of text mining in finance should be reviewed according to different decisions that are made based on its results (e.g., tactical, operational and strategic decisions). Taxonomy of various decisions based on text mining in finance could be developed in order to support decision making in a more effective manner, following the work of Gray et al. (2014) [18].Third, characteristics of organizations that have implemented text mining in their business processes should be investigated, with the goal of identifying best-practice approaches, but also obstacles that stand on the way to the successful

implementation of text mining in finance. Finally, more in-depth for text mining in finance should be conducted, focusing more on the internal documents as the domain of the analysis.

F ER EN CE S
[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] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.
[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.
[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.
[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012

R EF ER EN CE S
[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] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.
[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.
[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.
[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012Sagayam.R, A survey of text mining: Retrieval, extraction and in-dexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

A huge amount of text is flowing over the internet in the form of digital libraries, repositories, and other textual information such as blogs, social media network and e-mails [1]. An amount of text is flowing over the internet in the form of digital libraries, repositories, and other textual information such as blogs, social media network and e-mails [8].challenging task to determine appropriate patterns and trends to mining. Extract valuable knowledge from this large volume of data [2It is challenging task to determine appropriate patterns and trends to extract valuable knowledge from this large volume of data [9].

Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [3]. Text mining deals with natural language text which is stored in semi-structured and unstructured format [4]. Text mining techniques are continuously applied in industry, academia, web applications, internet and other field Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields [12]. [14] presented innovative and efficient pattern discovery techniques. They used the pattern evolving and discovering techniques to enhance the effectiveness of discovering relevant and appropriate information. They performed BM25 and vector support machine based filtering on router corpus volume 1 and text retrieval conference data to estimate the effectiveness of the suggested technique. [15] Performed various experiments of classification using multi-word features on the text. They proposed a hand-crafted

method to extract multi-word features from the data set. To classify and extract multi-word text they divide text into linear and nonlinear polynomial form in support of vector machine that improve the effectiveness of the extracted data techniques. They used the pattern evolving and discovering techniques to enhance the effectiveness of discovering relevant and appropriate information. They performed BM25 and vector support machine based filtering on router corpus volume 1 and text retrieval conference data to estimate the effectiveness of the suggested technique. [15] Performed various experiments of classification using multi-word features on the text. They proposed a hand-crafted method to extract multi-word features from the data set. To classify and extract multi-word text they divide text into linear and nonlinear polynomial form in support of vector machine that improves the effectiveness of the extracted data.

Different text mining techniques are available that are applied for analysing the text patterns and their mining pro-cess [16].Information Extraction systems are used to extract specific attributes and entities from the document and establish their relation- ship [18].IE systems are used to extract specific attributes and entities from the document and establish their relation-ship. There is a close relationship in text mining and information retrieval for textual data. In IR systems, different algorithms are used to track the user's behaviour and search relevant data accordingly [19].

Conclusion

By reviewing 50 papers, this paper aims to provide answers to the three research questions, and for that purpose, a qualitative analysis of literature has been conducted using a systematic literature review, citation, and co-citation investigation. The first research question was answered using the biometric analysis. The most important studies with the highest number of citations in the field have been identified, and a brief overview of the themes is given. In addition, papers that are the source of the field have been presented prior to the critical connection with recent studies identified. Based on this, the paper contributed to the existing literature through an overview of the most significant studies published in the Web of Science databases. Research trends have been identified as well. After reviewing the papers, it is possible to conclude that the research focus is on stocks price prediction, financial fraud detection and market forecast utilizing online text mining. The research results reveal that the current research trends of text mining are related to the need to analyse large amounts of data on websites and pages on social media, and to identify and test various text-mining techniques.

References

1. Ackermann U, Angelini B, Brugnara F, Federico M ,Giuliani D, Gretter R, Lazzari G and H. Niemann G ,"SpeeData: Multilingual Spoken Data Entry", International Conference, IEEE, Trento, Italy., 2211 - 2214.
2. Andreas Hotho, Andreas Nürnberger, and Gerhard Paa B',"A Brief Survey of Text Mining", July 2005 [17]
3. Agrawal R and Batra M, "A detailed study on text mining techniques,"International Journal of Soft Computing and Engineering (IJSCE) ISSN,pp. 2231–2307, 2013.
4. Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.
5. Brin S., and Page L.(1998), " The anatomy of a largescale hyper textual Web search engine", Computer Networks and ISDN Systems, 30(1-7): 107-117
6. Dang D. S and Ahmad P. H., "A review of text mining techniquesassociated with various application areas," International Journal ofScience and Research (IJSR), vol. 4, no. 2, pp. 2461–2466, 2015
7. Dean J. and Henzinger M.R. (1999), "Finding related pages in the world wide web", Computer Networks, 31(11-16):1467-1479.

8. Dion H. Goh and Rebecca P. Ang (2007),“ An introduction to association rule mining: An application in counseling and help seeking behavior of adolescents”, Journal of Behavior Research Methods39 (2), Singapore, 259-266
9. Emilio Sanchis, Davide Buscaldi, Sergio Grau, Lluis Hurtado and David Griol (2006),“ SPOKEN QA BASED ON A PASSAGE RETRIEVAL ENGINE”, Proceedings of IEEE international confernce, Spain, 62-65
10. Feuerriegel, S.; Gordon, J. News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. European Journal of Operational Research, 272(1),2019,pp. 162–175.
11. Fan.W, Wallace.L ,Rich.S,Zhang.Z, “Tapping the power of text mining,” Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.
12. Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy and Asghar Tajoddin (2008), “Optimizing Text Summarization Based on Fuzzy Logic”, Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer society, 347-352.
13. Fang Chen, Kesong Han and Guilin Chen (2008), “An approach to sentence selection based text summarization”, Proceedings of IEEE TENCON02, 489- 493.
14. Guihua Wen, Gan Chen, and Lijun Jiang (2006), “Performing Text Categorization on Manifold”, 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan , IEEE , 3872-3877. [20] JIAN-SUO XU (2007), “TCBPLK: A new method of text categorization ”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,, IEEE , 3889-3892.
15. Gupta and G. S. Lehal, “A survey of text mining techniques andapplications,” Journal of emerging technologies in web intelligence,vol. 1, no. 1, pp. 60–76, 2009
16. Hwee-Leng Ong, Ah-Hwee Tan, Jamie Ng, Hong Pan and Qiu-Xiang Li.(2001),“FOCI : Flexible Organizer for Competitive Intelligence”, In Proceedings, Tenth International Conference on Information and Knowledge Management (CIKM'01), pp. 523-525, Atlanta, USA, 5- 10.
17. Haralampos Karanikas and Babis Theodoulidis Manchester, (2001), “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management, UK
18. Jignashu Parikh and M. Narasimha Murty (200 2),“Adapting Question Answering Techniques to the Web”, Proceedings of the Language Engineering Conference, India, IEEE computer society.
19. Joby P. J. and Korra, “Accessing accurate documents by min-ing auxiliary document information,” in Advances in Computing andCommunication Engineering (ICACCE), 2015 Second InternationalConference on. IEEE, 2015, pp. 634–638.
20. Joe Carthy and Michael Sherwood-Smith (2002), “Lexical chanins for topic tracking”, International Conference, IEEE SMC WP1M1, Ireland
21. Kleinberg J.M., (1999), “Authoritative sources in hyperlinked environment”, Journal of ACM, Vol.46, No.5, 604-632.
22. 22.. Kanya N andGeetha S. (2007), “Information Extraction: A Text Mining Approach”, IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE, Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118.
23. Liao S.-H., Chu P.-H., Hsiao, P.Y., “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
24. Liu Lizhen, and Chen Junjie, China (2002), “ Research of Web Mining”, Proceedings of the 4th World Congress on Intelligent Control and Automation, IEEE, 2333-2337.
25. Liritano S. and Ruffolo M., (2001), “ Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining”, IEEE, 454-458 , Italy.
26. Navathe, Shamkant B., and Elmasri Ramez, (2000), “Data Warehousing And Data Mining”, in “Fundamentals of Database Systems”, Pearson Education pvt Inc, Singapore, 841-872.

27. Li Gao, Elizabeth Chang, and Song Han (2005),“ Powerful Tool to Expand Business Intelligence: Text Mining”, PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 8, 110-115.
28. 28.Mirjana M Peji´c Bach , Živko Krsti´c , Sanja Seljan, Lejla Turulja ,“Text Mining for Big Data Analysis in Financial Sector: A Literature Review”,
29. 29. Ming Zhao, Jianli Wang and Guanjun Fan (2008), “Research on Application of Improved Text Cluster Algorithm in intelligent QA system ”, Proceedings of the Second International Conference on Genetic and Evolutionary Computing, China,, IEEE Computer Society , 463-466.
30. 30. Pak Chung Wong, Paul Whitney and Jim Thomas,“ Visualizing Association Rules for Text Mining”, ”, International Conference, Pacific Northwest National Laboratory, USA, 1-5
31. 31. Padhy N., Mishra P, R.Panigrahi, (2012), "The Survey of Data Mining applications And Feature Scope” International Journal of Computer Science,Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
32. 32. Padhy. N, Mishra.D, Panigrahi.R , “The survey of data mining applications and feature scope,” arXiv preprint arXiv:1211.5723, 2012.
33. 33. RONALD NEIL KOSTOFF (2003),“TEXT MINING FOR GLOBAL TECHNOLOGY WATCH”, article, OFFICE OF NAVAL RESEARCH, Quincy St. Arlington, 1-27.
34. 34. Rajender Singh Chhillar (2008),“ Extraction Transformation Loading –A Road to Data warehouse”, 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries,India, 384-388.
35. 35. Siddth Kumar Chhajer, Rudra Bhanu Satpathy,“Adaptation of Text Mining as Text Data Mining, Similar to Text Analytics for the Process of Driving High-Quality Information from Text” November 2022
36. 36. Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference , IEEE computer society, 1-8.
37. 37. Seth Grimes (2005),“The developing text mining market”, white paper, Text Mining Summit05 Alta Plana Corporation,,Boston, 1-12.
38. 38. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan,“Using Text Mining Techniques for Extracting Information from Research Articles”, Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740,,Springer International Publishing AG 2018.
39. 39. Shantanu Godbole, and Shourya Roy, India (2008), “Text to Intelligence: Building and Deploying a Text Mining Solution in the Services Industry for Customer Satisfaction Analysis”, IEEE, 441-448.
40. 40. Sungjick Lee and Han-joon Kim (2008), “News Keyword Extraction for Topic Tracking”, Fourth International 74 JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009 © 2009 ACADEMY PUBLISHER Conference on Networked Computer.
41. 41. Steinberger.R, “A survey of methods to ease the development ofhighly multilingual text mining applications,” Language Resources andEvaluation, vol. 46, no. 2, pp. 155–176, 2012
42. 42. Sergio Bolasco , Alessio Canzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2002), “Understanding Text Mining:a Pragmatic Approach”, Roam, Italy.
43. 43. ting and Advanced Information Management, IEEE, Koria,554-559.
44. 44. Wang Bo and Li Yunqing (2008),“ Research on the Design of the Ontology-based Automatic Question Answering System”, International Conference on Computer Science and Software Engineering, IEEE, Nanchang, China, 871-874
45. 45. Wang Xiaowei, JiangLongbin, MaJialin and Jiangyan (2008), “Use of NER Information for Improved Topic Tracking”, Eighth International Conference on Intelligent Systems Design and Applications, IEEE computer society, Shenyang, 165-170.

46. 46. Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg.
47. 47. Weiss.S.M, Indurkhya.N ,Zhang.T, Damerau.F , “Text mining:predictive methods for analyzing unstructured information.”,Springer Science and Business Media, 2010.
48. 48. Wen Z., T. Yoshida, and X. Tang, “A study with multi-word featurewith text classification,” in Proceedings of the 51st Annual Meeting ofthe ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45
49. 49. XiQuan Yang, DiNa Guo, XueYa Cao and JianYuan Zhou (2008), “Research on Ontology-based Text Clustering”, Third International Workshop on Semantic Media Adaptation and Personalization, China,, IEEE Computer Society , 141-146.
50. 50. Zhou Ning, Wu Jiaxin, Wang Bing and Zhang Shaolong (2008), “A Visualization Model for Information Resources Management”, 12th International Conference Information Visualisation, China, IEEE, 57- 62.