ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1121022021.pdf
https://doi.org/10.30534/ijatcse/2021/1121022021

An Analysis on Text Mining Techniques for Smart Literature Review

Ahmed Mateen[1]*, Muhammad Yasir[1],Qamar Nawaz[1], Salman Afsar[1], Qasim Yasin[2] and Muhammad Yunusi[3]
[1]Department of Computer Science, University of Agriculture Faisalabad, Pakistan,
[2]Department of Computer Science, NUML Faisalabad Campus, Pakistan.
[3]Head of Informatics Department, Tajik National University, Tajikistan.
*Corresponding Author Email: ahmedbuttar@uaf.edu.pk

ABSTRACT

With the development of web technologies, databases and social networks etc. a large amount of text data is generated each day. Most of the data on the internet is in unstructured form. This unstructured data can provide valuable knowledge. For getting valuable knowledge from text data text mining techniques are used widely. As each day large amounts of research papers were published in journals and conferences. These research papers are very valuable for future research and investigations. These research papers act as a source for future innovations. Researchers write review papers to give updated knowledge about the specific field. But review papers used a limited number of papers and involved manually reading each paper. Due to the large volume of research papers published each day, it is not possible for the researchers to go through each paper to find the updated knowledge about their field of interest. To automate the literature analysis process different techniques of text mining were used. This paper provides a review of text mining techniques used in automatic literature analysis. We collected papers in which previous literature is used with text mining techniques to get valuable knowledge. This review paper presented an overview of text mining techniques, their evaluation criteria, their limitations and challenges for exploring literature to find research trends.

Key words: Text mining, Topic modelling. Keyword Extraction, Literature Analysis

1. INTRODUCTION

Every day many research papers are published in different journals and conferences to solve different problems or to provide updated knowledge. These research papers are very valuable for the current and future research and for generating new knowledge. Due to the large volume of these publications the process of manually reading a large number of papers for the purpose of exploring knowledge about a specific field is very time consuming and difficult. As there is a large amount of research data present on the research repositories like ScienceDirect, IEEE and Web of Science etc. there is a need to automatically analyze this data to get valuable knowledge from it. To cope with this issue text mining plays an important role by providing techniques which can be used to automate the complete process of getting valuable knowledge from the previous research literature.

Text mining approach is used to mine large amounts of text data to get valuable knowledge. There are many text mining techniques for solving different types of problems. Figure 1shows the process of text mining.

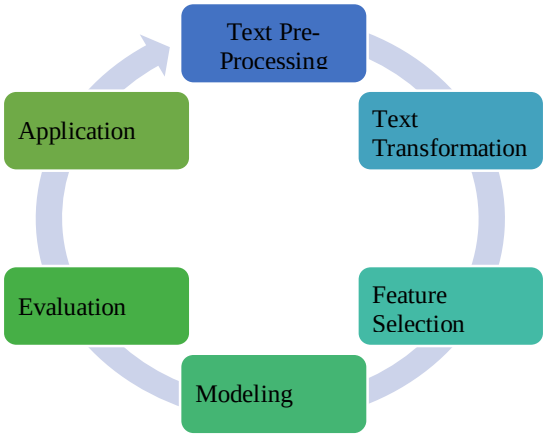

Figure 1: Process of Text Mining

The process of literature analysis is straightforward. The first step in literature analysis study is to collect papers related to their field from research repositories. For this purpose researchers need to know terms or keywords which can be used to search related documents, so there is a need to define a criteria for searching related material. After collection of data the next step is to clean this data. After cleaning the next step is data preprocessing in which different tasks performed like converting text to lowercase as lowercase and uppercase letters have the same meaning but for the computer uppercase and lowercase letters are different so converting all text to lowercase is the recommended step. As text data contains punctuations so removing these punctuations from text data is also important as these punctuations does not contribute to the subject discussed in the document. Other preprocessing tasks include tokenization, stop words removal etc. The text data need to be converted to a form on which different text mining techniques can be applied for analyzing data. After this evaluation of technique is performed based on different evaluation criteria.

Researchers used different text mining techniques to analyze a large volume of research works for solving a wide range of problems. In this study we present a review of text mining techniques usage in literature analysis. We explore the techniques used in different research papers for literature analysis, their evaluation criteria, limitations and challenges.

2. LITERATURE REVIEW

Sharma et.al [1]used Bibliometric analysis and STM to explore evolution of research trends in the Information management field by systematically analyzing research publications. Dataset consists of bibliographic data of 19,916 research articles, chapters of book, and reviews paper related to Information Management from Scopus database for the time span of 1970-2019.

Porturas et.al [2]used LDA, Hierarchical clustering using Heatmap and t-SNE for identifying thematic topics and research trends of emergency medicine. Dataset consists of 20528 abstracts of research papers related to emergency medicine from the OVID database published between the time span of 1980 to 2019.

Authors[3]used Term Frequency analysis and MALLET implementation of LDA to find thematic topics and topic evolution of Twitter research within the time span of 2006-2019. Dataset consists of abstracts of 18,000 research papers related to twitter research from EBSCO, IEEE and WOS repositories.

Kim et.al [4]used Word2vec, Spherical k-means clustering (W2V-LSA) to identify thematic topics and trends of the Blockchain technology field by using abstracts of 231 related research papers. Dataset consists of abstracts of research papers related to Blockchain published between the time span of 2014 to 2018.

Maede and his team[5]usedTF-IDF and keyword computation algorithm on the software AntConc for Exploring Telemedicine research trends in South Africa. Dataset consists of abstracts and titles of 36 research papers of 2019 related to Telemedicine from PubMed.

Zuliani et.al [6]used TF-IDF, LDA with Gibbs sampling, and Hierarchical clustering (Average linkage method with agglomerative clustering algorithm) on XLSTAT tool to explore research trends of mountain livestock farming research. Dataset consists of abstracts of 2,679 research papers related to Mountain livestock farming from Scopus from the year 1980 to year 2018.

Ding et.al [7]used MALLET implementation of LDA with 1800 Gibbs sampling iterations for exploring the trending topics studied in the big data field. Dataset consists of abstracts, title and keyword of 17,599 research papers related to Big Data from Elsevier's Scopus database published between 2009 and 2018.
Article [8]used TF-IDF and LDA for Identifying research trends in Data Science Articles. Dataset consists of 214 research papers related to data science from IEEE and ScienceDirect.

Author[9]used Noun Phrase Mining, and Term Frequency to explore research trends of data mining techniques. Dataset consists of abstracts and metadata of 5,843 data mining articles from ScienceDirect for the time span of 2014-2018.

Researchers[10]used TF-IDF, n-grams and LDA for exploring research trends of building saving energy research. Dataset consists of 1600 building energy saving research papers from three databases WOS, ScienceDirect and JSTOR within the time span of 1973 and 2016.

Scientist[11]used LDA for identifying thematic topics of bioinformatics research. Dataset consists of titles and abstracts of 143,000 research papers related to the Bioinformatics field from the PubMed database published between 1987 and 2018.

3. MATERIALS AND TECHNIQUES

Researchers used different text mining techniques for the purpose of exploring research literature. Topic modeling, keyword selection, term frequency, TF-IDF, Noun Phrase Mining etc. are some of the techniques used for analyzing research literature.

3.1 Topic Modeling

The most widely used text mining technique for literature analysis is topic modeling, topic modeling is an efficient and systematic technique in analyzing a large number of documents in a small amount of time by identifying thematic topics from these documents [12]. TM is used to identify the thematic topics of a large number of text documents and occurrence of these topics in these documents. It is an unsupervised learning technique. Themes which are identified using TM technique are present within these documents. Every topic generated by TM technique is a collection of words that are semantically similar. So, we can use TM technique to find the relationship between large numbers of text documents by identifying the thematic topics of these text documents [1].

LDA is an efficient and most widely used topic modeling technique among different topic modeling techniques [13]. Latent Dirichlet Allocation (LDA) used statistical distributions. LDA works by assuming a certain number of topics which can represent the documents and then assign words to these topics by calculating probabilities of each word in the corpus. LDA used a bag of words model to represent documents. The topics which LDA generates consist of words that have similarity among themselves. The collection of words for each topic then analyzed by a human to assign them an expressive topic or theme[3].

LDA is used by many researchers for different purposes.Scientist [14]used LDA to identify latent themes of teacher self-assessment survey dataset. Article [15]used the LDA model for creating a system for recommending TV programs to the user. A team[16]used the LDA model in identifying spam messages. Bastani et.al [17] applied the LDA model on the corpus of consumer complaints to find consumer issues.

Gurcan and his team[7]used MALLET implementation of LDA topic model with Gibbs Sampling to find the major topic studied in big Data research. This study used abstracts, titles and keywords of 17599 journal articles related to Big Data from Elsevier's Scopus database. However, Karami et.al[3] used frequency

analysis and MALLET implementation of LDA topic model to find major themes in twitter related researches and evolution of research trends using 18000 abstracts of research papers related to twitter from three research repositories IEEE, Web of Science and EBSCO for the period of 2006 to 2019. Used log-likelihood to evaluate LDA model. For finding the number of topics this study applied the topic model for topics of range 2 to 100 and selected 40 as the number of topics. This study used frequency analysis to find frequent terms in the whole corpus which is a good starting point for getting a top level view of terms which are used frequently in documents. This study did not collect all the papers from a single research database but used three research repositories which make this research better in terms of using more than one source for collecting research papers. This study makes 3 clusters of topics manually.

While study [6] used LDA for finding research trends of Mountain livestock farming research. This study used LDA for topic generation and then used hierarchical clustering to assign topics to relevant clusters.

Scientist[8]used TF-IDF and LDA topic model to find research trends in data science research. This study used 214 research papers from IEEE and ScienceDirect. Evaluated the LDA model by visualizing the topics. This study used complete research papers but did not provide criteria for selecting these papers.

Researchers[2] implemented LDA using Mallet to find research trends of emergency medicine. Also perform clustering of generated topics using Heatmap. Used t-SNE to visualize abstracts with topic distributions. The t-SNE is a technique which is used to visualize high dimensional data in lower dimensions [18]. Used coherence score as a criteria for deciding number of topics. Proposed t-SNE and Heatmap are good measures to find relationships between topics.
Study[1]used bibliometric analysis and structured topic modeling to identify the research trends of Information management with respect to authors, institutions and countries. STM is a new TM technique. STM is not widely used yet. Few studies used STM. Study[19]used STM to explore opinions of physically disable persons about autonomous vehicles. Study[20]used STM for sentiment analysis of political speeches.

Study[4]used word2vec with LSA to identify research trends of Blockchain research using abstract of related research papers. This study proposed that use of word2vec with topic models can improve the quality of topics. Word2vec is a neural network-based model, which is used to represent words present in the text documents as a vector on the basis of context [21]. This study uses word2vec with spherical clustering to explore the literature. This study also compares this proposed method with PLSA and proves word2vec-LSA technique gives better results.

3.2 Keyword Selection

Keyword Selection is the process of finding keywords from the corpus that can represent the subject discussed in the document. In this mechanism more general terms are excluded to get only those terms which can represent the subject of the document. Study [5]applied keyword computation algorithm on the abstracts and titles of telemedicine research papers for the purpose of exploring Telemedicine research trends in South Africa. Study[5]also provides a comparison between frequency analysis using TF-IDF and Keyword Selection. This study uses a small amount of research papers (n=36) for applying keyword computation and TF-IDF analysis. This study suggests that using a large number of research papers may improve the result of keyword computation algorithm but not provide any evidence of this. This study also suggests that TF-IDF results can also be improved by using n-grams or context phrases.

3.3 Frequency Analysis

TF-IDF is a term frequency analysis technique. TF is measured by counting the occurrence of a word in the document and dividing it by total count of words in the document.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad (1)$$

IDF is used to find the unique words weight within all documents in the corpus. It is measured by logging the total number of text documents present in the corpus divided by the total number of text documents that have word w.

$$idf(w) = log\ (\frac{N}{dft}) \qquad (2)$$

We can calculate TF-IDF by multiplying term frequency and IDF.

$$Tf - IDF = TF * IDF \qquad (3)$$

Study [3]used TF for frequency analysis. Study [8]used frequency analysis with word frequency rank. Study [9]used Term frequency analysis with noun phrase mining to identify data mining techniques used in data mining research. This study used abstracts and metadata of 5543 research papers related to data mining from IEEE and ScienceDirect published within the span of 2014 to 2018. This study applied noun phrase mining to shorten the corpus as they only need to find the trending techniques in data mining research. They rank the data mining techniques by frequency analysis. This study suggests the use of noun phrases with TF for finding trends of data mining techniques used in research. This noun phrase mining technique with term frequency analysis can be used in future for exploring trending techniques of any research field.

4. RESULTS AND DISCUSSION

This paper provides a detailed review of text mining techniques used for literature analysis. We first collect the papers related to literature analysis in which text mining techniques are used. Then we provide complete details of techniques used for literature analysis. In literature analysis studies we explored different evaluation criteria that were used like coherence score

for topics, perplexity score for topics, and qualitative measure etc. Coherence score is the measure of finding semantic similarity between words of a topic [22]. Many studies rely on the qualitative evaluation. In the Table 1 we provide different evaluation criteria used for evaluating techniques used in different studies.

The most widely used technique for literature analysis is LDA. And most of the studies used MALLET for LDA implementation. LDA uses a bag of words model which does not consider order or context of words. This limitation of the LDA model can be removed by using some technique like word2vector or other word embedding model which consider order or context of words. Hence, this is an opportunity for researchers to use context based techniques with LDA for improving its performance.

Some studies used only frequency analysis to find research trends. In these studies there is no quantitative criteria to check the accuracy of these frequency analysis techniques. Using only words frequency gives very commonly used words which do not represent the topic discussed in the document. We can use POS tagging with words frequency techniques for getting better results. Most of the studies use frequency analysis with topic modeling techniques for literature analysis. Most studies used topic modeling techniques for exploring literature but only one study evaluated their topic model results with other topic modeling techniques. So comparing performances of different topic models is also an opportunity for future research. Most of the studies used the LDA topic model for analyzing research trends. Some studies only generate topics and do not present evolution of these topics, however most of the studies provide evolution of topics for a specific time span. The topic modelling technique can also be used to identify research trends with respect to countries or authors etc.

Deciding the number of topics in topic modeling technique is a very important step. Most of the studies use log-likelihood and coherence scores for identifying appropriate numbers of topics that can be generated from the corpus. However, automating the process of determining the number of topics for topic modeling techniques is a challenge for future research.

Few studies used STM for literature analysis. It is a new topic model, so there is an opportunity for researchers to improve this model performance by using different techniques and compare it with other topic models like LDA, LSA and NNMF. Some studies used t-SNE to evaluate LDA generated topics. The t-SNE is a good method to visualize documents with their topic distributions. One study proposed W2V-LSA and proved its significance in literature analysis by comparing it with PLSA. Word2Vector text representation can be used with other topic models like LDA, NNMF in future research for improving these models' performances.As topic models generate words for the topic, which researchers manually interpret and give them a label. Automatic generation of labels from words is also a challenge for future research.In the Table 1 we present the evaluation criteria used in different studies.

Table 1: Evaluation Criteria

Reference	Evaluation Criteria
[1]	Average semantic coherence and exclusivity of topics.
[4]	Compare results of W2V-LSA with PLSA. Use qualitative measures to evaluate the model by checking the document of a topic. Use quantitative measures like Topic Coherence and Keyword Matching.
[6]	For selecting the number of topics for LDA log-likelihood and perplexity used and 10 topics generated. Jaccard distance for calculating distance among topics. Topic assigned to cluster on the basis of entropy.
[3]	For evaluating the LDA model this study used a log-likelihood measure for 5 sets of one thousand (1000) iterations. Used ldatuning R package for determining the number of topics to be generated by using LCM on the range of 2 to 100 topics, and found 40 topics.
[5]	Compare results of TF-IDF with keyword computation. Manually evaluate the performance of these two techniques by visualizing terms generated by these techniques on the corpus consisting of abstracts and titles of 36 Telemedicine research papers.
[2]	Used coherence score to find the number of topics to be generated using LDA on MALLET.
[7]	For determining number of topics the LDA model was implemented for varied number of topics in the range {10, 12, 14, ... ,50}, and generate 36 topics Evaluate the LDA model by visualizing topics.
[8]	Evaluate results by qualitative measures.
[9]	Evaluate the results of noun phrase mining and frequency analysis by visualizing the generated terms and using Zipf's law for frequency analysis.
[10]	Used qualitative measures for evaluation of results.
[11]	Used coherence measure to find the number of topics to be generated using LDA.

5. CONCLUSION

This paper provides a detailed review of studies related to literature analysis using text mining techniques. We first collect the papers related to literature analysis in which text mining techniques are used. In this review paper we provide a detailed analysis of these studies. We divide this review paper into four sections, Introduction, Literature Review, Text Mining

Techniques, Result and Discussions of these studies and challenges for future research. In the introduction section we introduce the topic and give historical knowledge about it. In the Literature Review section we provide complete details of studies we used for this review paper. In the Text Mining techniques section we explore the techniques used for literature analysis. The most used techniques in literature analysis studies are frequency analysis and topic modeling. LDA is the most widely used topic model for literature analysis. We also explore different criteria for evaluating these text mining techniques. We also present limitations of these studies and challenges for future research in the result and discussion section.

REFERENCES

1. A. Sharma, N. P. Rana, and R. Nunkoo, **Fifty years of information management research: A conceptual structure analysis using structural topic modeling**, *International Journal of Information Management*, vol. 58, 2021.
2. T. Porturas and R. A. Taylor, **Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling**, *American Journal of Emergency Medicine*, Aug. 2020.
3. A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, **Twitter and Research: A Systematic Literature Review through Text Mining**, *IEEE Access*, vol. 8, 2020.
4. S. Kim, H. Park, and J. Lee, **Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis**, *Expert Systems with Applications*, vol. 152, 2020.
5. A. Maeder, M. George, and B. Naveda, **Identifying recent telemedicine research trends using a natural language processing approach**, in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa, 2020, pp. 1-6.
6. A. Zuliani *et al.*, **Topics and trends in Mountain Livestock Farming research: a text mining approach**, *Animal*, vol. 15, no. 1, 2021.
7. F. Gurcan and S. Sevik, **Big Data Research Landscape: A Meta-Analysis and Literature Review from 2009 to 2018**, in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Ankara, Turkey, 2019, pp. 1-5.
8. S. H. Adil, M. Ebrahim, S. S. A. Ali, and K. Raza, **Identifying Trends in Data Science Articles using Text Mining**, in *2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, Karachi, Pakistan, 2019, pp. 1-7.
9. S. Nazir, M. Asif, and S. Ahmad, **The Evolution of Trends and Techniques used for Data Mining**,in *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan, 2019, pp. 1-6.
10. Z. Ding, Z. Li, and C. Fan, **Building energy savings: Analysis of research trends based on text mining**, *Automation in Construction*, vol. 96, pp. 398-410,Oct. 2018.
11. A. Youssef and A. Rich, **Exploring trends and themes in bioinformatics literature using topic modeling and temporal analysis**, in *2018 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, Farmingdale, NY, USA, 2018, pp. 1-6.
12. A. Karami, M. Ghasemi, S. Sen, M. F. Moraes, and V. Shah, **Exploring diseases and syndromes in neurology case reports from 1955 to 2017 with text mining**, *Computers in Biology and Medicine*, vol. 109, pp. 322–332, Jun. 2019.
13. J. C. Campbell, A. Hindle, and E. Stroulia, **Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data**, in *The Art and Science of Analyzing Software Data*, vol. 3, no. 4–5, 2015, pp. 139–159.
14. D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, **Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach**, *IEEE Access*, vol. 8, pp. 35318–35330, Feb. 2020.
15. S. Pyo, E. Kim, and M. Kim, **LDA-Based Unified Topic Modeling for Similar TV User Grouping and TV Program Recommendation**, *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1476–1490, Aug. 2015.
16. N. Al Moubayed, T. Breckon, P. Matthews, and A. S. McGough, **Sms spam filtering using probabilistic topic modelling and stacked denoising autoencoder**, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9887 LNCS, pp. 423–430.
17. K. Bastani, H. Namavari, and J. Shaffer, **Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints**, *Expert Systems with Applications*, vol. 127, pp. 256–271, Jul. 2019.
18. L. Van Der Maaten and G. Hinton, **Visualizing Data using t-SNE**, *Journal of Machine Learning Research*. vol. 9, no. 86, pp. 2579–2625,Nov. 2008.
19. R. Bennett, R. Vijaygopal, and R. Kottasz, **Attitudes towards autonomous vehicles among people with physical disabilities**, *Transportation Research Part A: Policy and Practice*, vol. 127, pp. 1–17, Sep. 2019.
20. D. Liu and L. Lei, **The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election**, *Discourse, Context and Media*, vol. 25, pp. 143–152, Oct. 2018.
21. T. Mikolov, K. Chen, G. Corrado, and J. Dean, **Efficient estimation of word representations in vector space**, in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*,2013.
22. M. Röder, A. Both, and A. Hinneburg, **Exploring the space of topic coherence measures**, in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408.