Application of Text Mining on the Editorial of a Newspaper of Bangladesh

Tarequl Islam Manir
Department of Statistics
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh

Md. Moyazzem Hossain
Department of Statistics
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh

ABSTRACT
The development in the fields of web, digital libraries, technical documentation, and medical data has made it easier to access a larger amount of a textual document. Owing to the increasing demands to obtain knowledge from a large number of textual documents accessible on the web, text mining is gaining significant importance. The aim of this paper is to find the topmost ten frequent words of the whole writing of the Editorial writing appeared in the most popular Bangladeshi Daily English newspaper entitled "The Daily Star" over the period 01 January 2018 to 30 June 2018. Finally, we representation of text data visually with the help of word cloud. The results indicate that the most frequent word highlighted is 'government' in the writing of all months considered in this study. Also, the words "Bangladesh", "Myanmar", "people", "must" and "will" emerge in the analysis.

General Terms
Data Mining, Natural Language Processing.

Keywords
Text Mining; Editorial; Frequent Words; Information Extraction; Word Cloud.

1. INTRODUCTION
Nowadays, the data size is accumulating at an exponential rate. More or less the organizations, institutions, and business industries are storing their data electronically. A huge amount of text is also flowing through the internet from different digital libraries, repositories, and other textual information such as social media network, blogs and e-mails (Sagayam [1]). Therefore, to determine the appropriate patterns and trends or extract valuable knowledge from this large volume of data is too much difficult (Padhy, et al., [2]). However, Text mining is a useful technique to extract exciting and significant patterns to explore knowledge from the textual data sources (Fan, et al. [3]). Thus, these techniques are incessantly applied in academia, web applications, internet, industry and other fields (Liao, et al. [4]). Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud recognition, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis (He [5]).

In this modern culture, the text is one of the most common means of transportation for the formal exchange of information. It is essential to extract useful information from texts however it is a very difficult task indeed. Text mining is an interdisciplinary arena which includes information retrieval, data mining, machine learning, statistics, and computational linguistics (Jusoh and Alfawareh [6]). Research in text mining has been carried out since the mid-80s when the US academic, Professor Don Swanson, realized that, by combining information slice from seemingly unrelated medical articles, it was possible to deduce new hypotheses (Nightingal [7]). However, Text Mining or knowledge discovery from the text (KDT) was first introduced by Fledman and Dagan [8].

In the early years, only information specialists used text mining systems. However, recent's research on text mining has been carried out by the researchers of various fields (Jusoh and Alfawareh [9]). With the help of text mining application, unstructured textual information is used to discover the structure and implicit hidden meanings of the text (Rao [10]; Karanikas, et al. [11]). Through text mining, one can uncover hidden patterns, relationships, and trends of a text. Hale [12] disputed that the benefits of using text mining are to reach a decision more quickly, at least 10x speedup over previous methods, and find hidden information. Chakrabarti [13] addressed that text mining facilitates organizations to discover interesting patterns, models, directions, trends, rules, contained in a text in the same way of data mining which explores tabular or "structured" data.

Roul et al. [14] presented a top down and bottom up approach for web-based text mining process. Sumathy and Chidambaram [15] gave an overview of applications, tools and issues arise to mine the text. Patel and Sharma [16] discuss the different method of text categorization and cluster analysis for text documents. Jhanji and Garg [17] provide an overview of text mining in the contexts of its techniques, application domains, and the most challenging issue. They also stretch the emphasis on the fundamentals methods of text mining including natural language possessing (NLP) and information extraction (IE). Kaushik and Naithani [18] reviewed the text mining techniques, tools, and it's various applications. Nie and Sun [19] performed a two-dimensional text mining approach, including bibliometric and network analysis, in order to detect trends of major academic branches. This paper attempt to examine the Editorial writing of the most popular Bangladeshi daily English newspaper named "The Daily Star". So, firstly we find out the topmost ten frequent words of the whole writing. Finally, the word cloud is used to represent the text data visually.

2. MATERIALS AND METHODS

2.1 Date Preparation
This paper examines the writing of Editorial section of a Bangladeshi Daily English newspaper which is The Daily Star. Firstly, the documents are collected from the official website of this newspaper over the period 01 January 2018 to 30 June 2018. The link of the official website of this newspaper is http://www.thedailystar.net/editorial.

2.2 Methods

Scientists in the text mining community have been trying to apply numerous techniques or methods such as rule-based, knowledge-based, statistical and machine-learning-based approaches. However, among the methods, the most popular, as well as essential methods for text mining, are natural language processing (NLP) and information extraction (IE) techniques. The subsequent diagram illustrates the process of text mining.

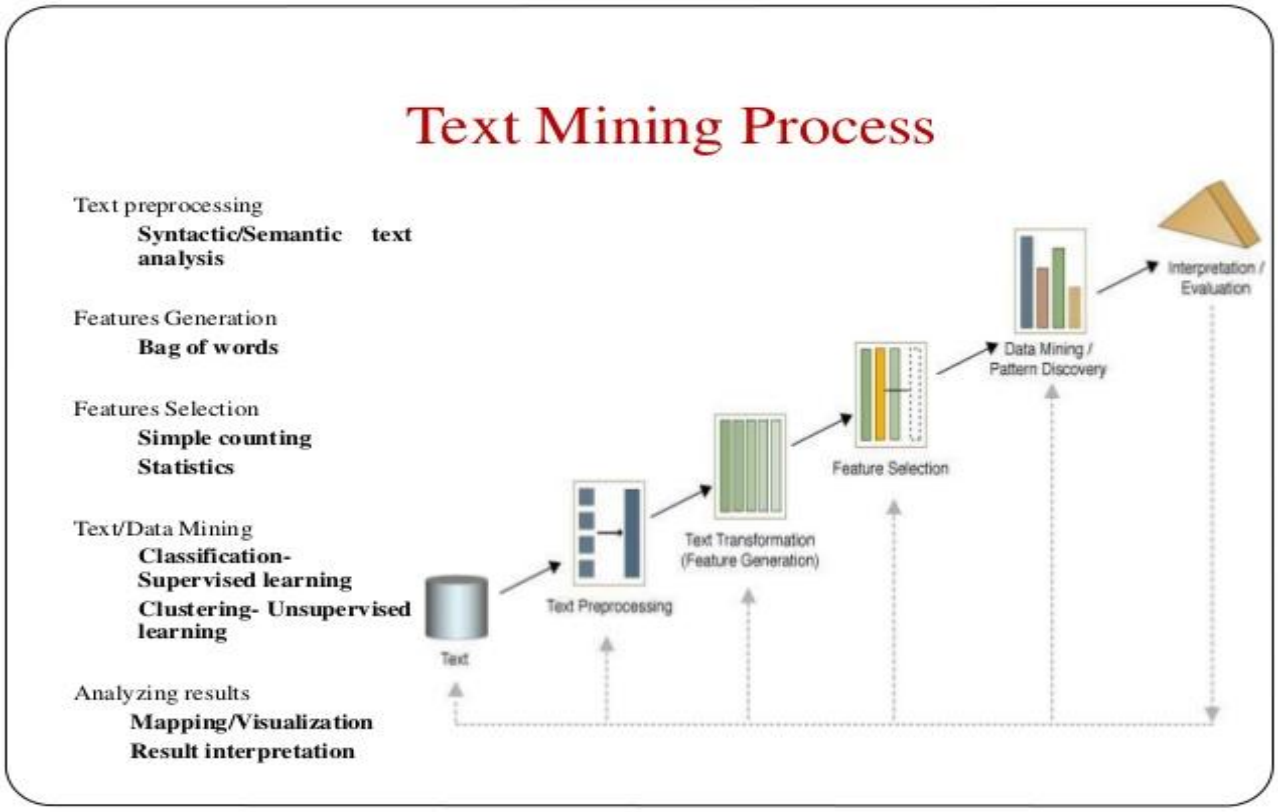

Fig 1: Text Mining Process

2.2.1 Natural Language Processing (NLP)

Automatic processing and analyzing the unstructured textual information is the main concern of Natural language processing (NLP). It executes different types of analysis such as Named Entity Recognition (NER) is used for abbreviation as well as their synonyms extraction and find the relationships among them (Laxman, and Sujatha [20]). However, NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG usages several levels of underlying linguistic representation of the text in order to make sure that the generated text is grammatically correct and fluent. Moreover, NLU calculates the meaning illustration, essentially restricting the discussion to the domain of computational linguistics. In addition, NLU consists of at least of one the components from tokenization, morphological or lexical analysis, syntactic analysis, and semantic analysis. A sentence is segmented into a list of tokens by tokenization. The token represents a word or a special symbol such as an exclamation mark (Jusoh and Alfawareh [9]).

2.2.2 Information Extraction (IE)

Information Extraction (IE) encompasses directly with the text mining process with the help of extracting valuable evidence from the texts. IE can also be defined as the formation of a structured representation of selected information drawn from texts. In IE, natural language texts are recorded to be predefine, structured representation, or templates, which, when it is occupied, represent an extract of significant evidence from the original text (Rao [10]; Karanikas, et al. [11]). Moreover, IE systems are applied to extract specific attributes and entities from a document and establish their relationship (Dang and Ahmad [21]). The extracted corpus is stored into database for supplementary processing. In order to achieve more appropriate results from information extraction process, in-depth and complete information about the relevant field is required (Steinberger [22]). The goal of text mining is to find specific data or information in natural language texts. Therefore, the IE task is defined by its input and extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, which includes tables or itemized and enumerated lists. Then data mining procedures can be applied to the data for discovering new knowledge.

2.2.3 Information Retrieval (IR)

Information Retrieval (IR) is an extracting procedure of finding the appropriate and associated patterns according to a supplied set of words or phrases. In IR systems, different algorithms are used to track the user's behavior and search relevant data accordingly (Steinberger [22]). Most popular search engines i.e., Google and Yahoo are using more frequently the information retrieval system in order to extract the relevant documents according to a phrase on Web. These search engines provide more relevant and appropriate information to users that satisfying their needs (Zhong, et al. [23]).

2.2.4 Text Summarization

Text summarization is a procedure of gathering and generating a concise representation of original text documents (Mukhedkar, et al. [24]). In summarization, pre-processing and processing operations are executed on the raw text. Tokenization, stop word removal, and stemming methods are applied for pre-processing. However, Lexicon lists are produced at the processing stage of text summarization. In the past, automatic text summarization was performed on the basis of existence a certain word or phrase in the document. Later on, additional methods of text mining were introduced with the standard text mining process to improve the relevance and accuracy of results (Chen and Zhang [25]). Moreover, the weighted heuristics method extract features of a text by using specific rules in order to summarize the text documents. Sentence length, fixed phrase, paragraph, thematic word, and upper case word identification features can be implemented and analyzed for text summarization. Text summarization techniques can be applied to multiple documents at the same time. Quality and type of classifiers depend on the nature and theme of the text documents (Al-Hashemi [26]).

2.2.5 Word Cloud

Word clouds or tag clouds are defined as the visual representation of words for a certain written content structured as per its frequency (Jayashankar and Sridaran [27]). Among the most commonly used method of presenting text data in a graphical manner; Word cloud is helpful for analyzing various forms of text data such as essays and short answers or written opinions to a survey or questionnaire (De Paolo, and Wilkinson [28]). Furthermore, word cloud used as a preliminary stage for in-depth analysis of certain text material (Sinclair and Cardew-Hall [29]; Viegas, et al. [30]). Nonetheless, the method has certain drawbacks as well. One of the major drawbacks that is it does not consider the linguistic knowledge about the words and their respective link to the given subject while providing a purely statistical summary to the segregated words. As a result, in most systems, the word clouds are often employed in a statistical manner for summarizing text, providing very little or no means for correlating the data. It is perceived that this could be one of the most influencing paradigms of visualization for most of the analysis conditions. Thus, in this paper, we have employed the use of word clouds as the central method of text analysis.

3. RESULTS AND DISCUSSION

This paper considers the Editorial section which contains two writing topics in each day of a daily Bangladeshi English newspaper named "The Daily Star". This paper also collects the documents for the month of January 2018 to June 2018. It is observed that the word limit of the editorial of a month is above fifteen thousand and just above the eighteen thousand. The following figures represent the most ten frequent words in the writing.

Among the words, "Government" is used 329 times, and the frequency of the most frequent ten words are more than 170. Among the popular ten words "Law" has the lowest frequency which is about half of the topmost frequent word "Government". In the editorial writing published in the month of January 2018 in "The Daily Star", Bangladesh, we observed that most of the writing issues are related to the Government of Bangladesh. In this month, the authors also discussed the enormous influx of Rohingyas to Bangladesh came from Myanmar fleeing horrific atrocities in their homeland. In response, Bangladesh hosted the Rohingyas and gave shelter these oppressed people. At that time, the discussed people of Bangladesh and came from Myanmar. The authors of the editorial addressed different issues related to several laws. Already there are more than three lakh Rohingyas in Bangladesh as a result of previous exoduses (**Fig 2**).

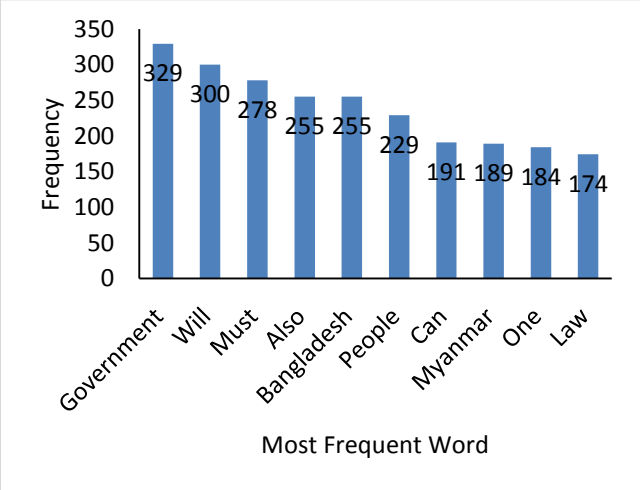

Fig 2: Bar diagram of most frequent word of Editorial writing in January 2018.

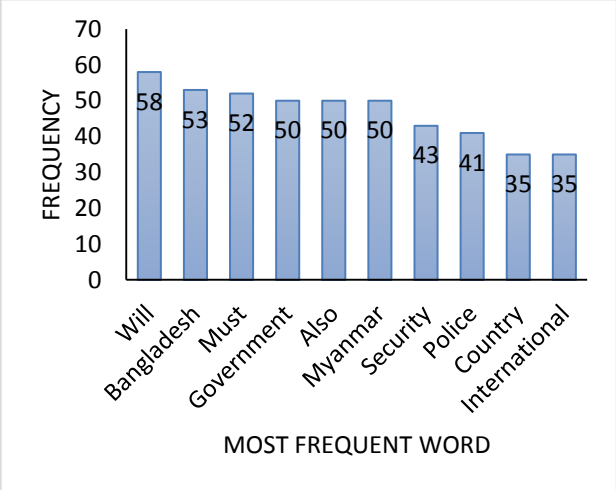

Fig 3: Bar diagram of most frequent word of Editorial writing in February 2018.

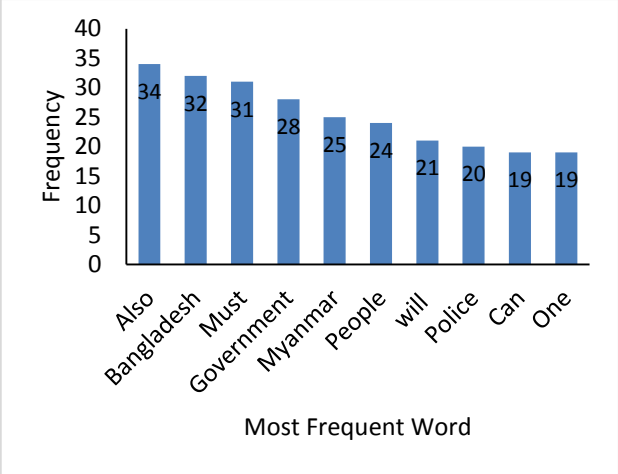

Fig 4: Bar diagram of most frequent word of Editorial writing in March 2018.

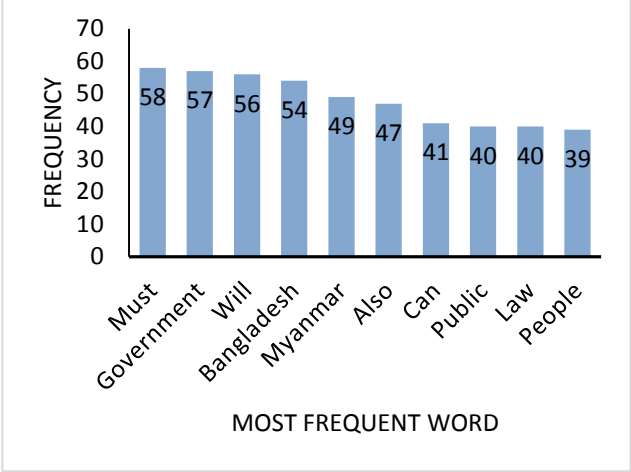

Fig 5: Bar diagram of most frequent word of Editorial writing in April 2018.

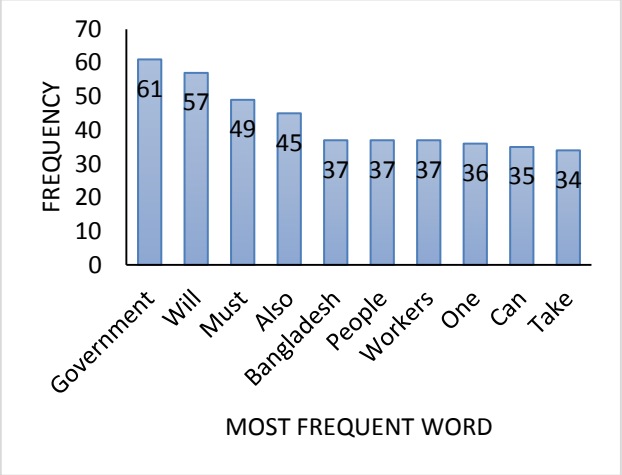

Fig 6: Bar diagram of most frequent word of Editorial writing in May 2018.

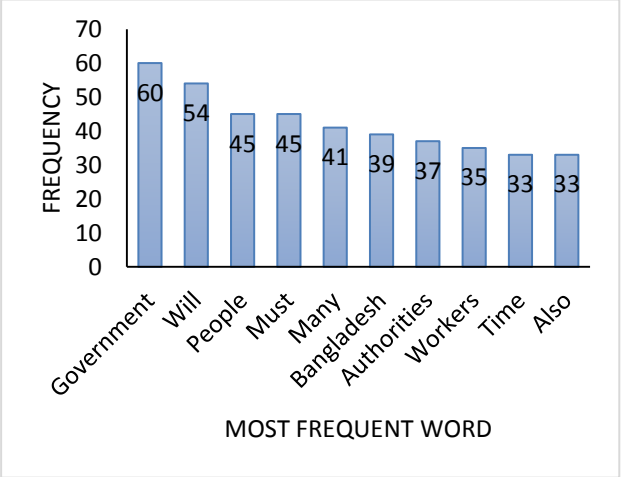

Fig 7: Bar diagram of most frequent word of Editorial writing in June 2018.

However, in the month of February 2018, we observed that the writers continued their writing related to Rohingya issues those are also associated with the Government of Bangladesh as well as Myanmar. Authors also raised the issues about the role of police and security of peoples of both countries (**Fig 3**). While the generosity of Bangladesh has been lauded by the international community, this does not help to solve the problem of continuing to give shelter to almost one million people in a country that is already burdened with overpopulation, poverty and acute scarcity of land. In addition, from the most frequent words, it can be said that the authors write the topics about Rohingya issues and the roles of the people, police and the Government of Bangladesh and Myanmar. Furthermore, from the most frequent words used in writing of the remaining months, it may be concluded that the authors continued their writing about issues related to Rohingya.

The word cloud of the most frequent words used in the editorial section of the renowned daily newspapers of Bangladesh entitled "The Daily Star" for the month of January to June 2018 are presented below.

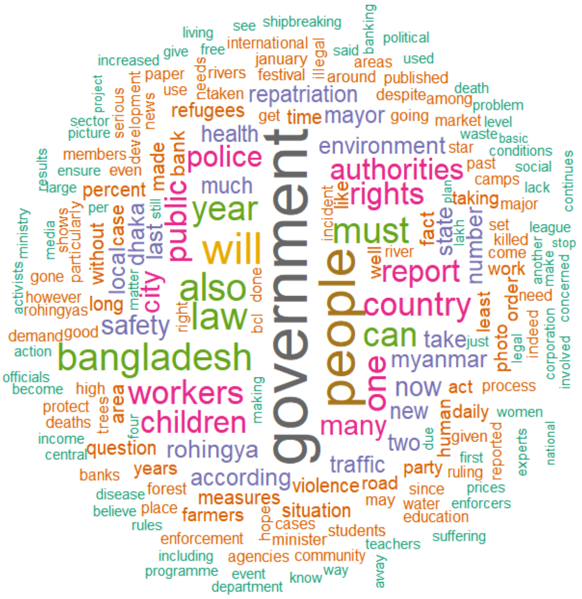

Fig 8: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of January 2018.

Fig 9: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of February 2018.

Fig 10: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of March 2018.

Fig 11: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of April 2018.

Fig 12: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of May 2018.

Fig 13: Word cloud of the most frequent words of Editorial writing of "The Daily Star" in the month of June 2018.

From the **Fig 8**, we see that the most 10 frequent words were government, will, must, also, Bangladesh, people, can, Myanmar, one, law and will, may, one, public. However, from the **Fig 3**, it can be seen that the most frequent words are will, Bangladesh, must, government, also, Myanmar, security, police, country, international. Firstly, it can be seen that the most protuberant word highlighted is 'government' in the editorial writing of the most popular daily English newspaper entitled "The Daily Star" published in the months January to June 2018. Moreover, the words like "Bangladesh", "Myanmar", "people", "must" and "will" emerge in the analysis.

4. CONCLUSION

At present, text mining is one of the utmost essential, curious and interesting fields. With the passage of time its importance is only going to increase because the rate of data production is very high and through text mining, one can uncover hidden patterns, associations, and tendencies in a text that is the benefits of making the decision more quickly. From the results, it can be seen that the most prominent word highlighted is 'government' in the writing of all months considered in this study. The terms "Bangladesh", "Myanmar", "people", "must" and "will" emerge in the analysis.

5. ACKNOWLEDGMENTS

Authors would like to give thanks to the Editor as well as the anonymous referees who check carefully the manuscript and provide the suggestions which improve the quality and readability of the paper.

6. REFERENCES

[1] Sagayam, R. 2012. A survey of text mining: Retrieval, extraction and indexing techniques. International Journal of Computational Engineering Research. 2(5), 1443-1446.

[2] Padhy, N., Mishra, D. and Panigrahi, R. 2012. The survey of data mining applications and feature scope. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT). 2(3), 43-58. Doi: 10.5121/ijcseit.2012.2303.

[3] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2006. Tapping the power of text mining. Communications of the ACM. 49(9), 76–82. Doi: 10.1145/1151030.1151032.

[4] Liao, S. H., Chu, P. H. and Hsiao, P. Y. 2012. Data mining techniques and applications–a decade review from 2000 to 2011. Expert Systems with Applications. 39(12), 11303–11311. Doi: 10.1016/j.eswa.2012.02.063.

[5] He, W. 2013. Examining students online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior. 29(1), 90–102. Doi: 10.1016/j.chb.2012.07.020.

[6] Jusoh, S. and Alfawareh, H. M. 2012. Techniques, Applications and Challenging Issue in Text Mining. IJCSI International Journal of Computer Science Issues. 9(6), 431-436.

[7] Nightingal, J., 2006. Digging for data that can change our world. The Guardian, 10 January, 2006. Accessed on 12 February 2018. Available at https://www.theguardian.com/education/2006/jan/10/elearning.technology14.

[8] Feldman, R. and Dagan, I. 1995. Knowledge Discovery in Textual Databases

[9] (KDT). KDD'95 Proceedings of the First International Conference on Knowledge Discovery and Data Mining. 112–117.

[10] Jusoh, S. and Alfawareh, H. M. 2007. Natural language interface for online sales. Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007). Malaysia: IEEE, 224–228.

[11] Rao, R. 2003. From unstructured data to actionable intelligence. IT Professional. 5(6), 29-35. Doi: 10.1109/MITP.2003.1254966.

[12] Karanikas, H., Tjortjis, C. and Theodoulidis, B. 2000. An approach to text mining using information extraction. Proceedings of Workshop of Knowledge Management: Theory and Applications. September 13-16, 2000, Lyon, France.

[13] Hale, R. 2005. Text mining: Getting more value from literature resources. Drug Discovery Today. 10(6), 377–379.

[14] Chakrabarti, S., 2000. Mining the Web: Analysis of Hypertext and Semi Structured Data. San Francisco, CA: Morgan Kaufman.

[15] Roul R.K., Varshneya S., Kalra A., Sahay S.K. 2015. A Novel Modified Apriori Approach for Web Document Clustering. In: Jain L., Behera H., Mandal J., Mohapatra D. (eds) Computational Intelligence in Data Mining - Volume 3. Smart Innovation, Systems and Technologies, vol 33. Springer, New Delhi. Doi: 10.1007/978-81-322-2202-6_14.

[16] Sumathy, K. and Chidambaram, M. 2013. Text mining: Concepts, applications, tools and issues-an overview. International Journal of Computer Applications. 80(4), 29-32. Doi: 10.5120/13851-1685.

[17] Patel, M. R. and Sharma, M. G. 2014. A survey on text mining techniques. International Journal of Engineering and Computer Science. 3(5), 5621-5625.

[18] Jhanji, D. and Garg, P. 2014. Text Mining. International Journal of Scientific Research and Education. 2(8), 1642-1648.

[19] Kaushik, A. and Naithani, S. 2016. A Comprehensive Study of Text Mining Approach. International Journal of Computer Science and Network Security. 16(2), 69-76.

[20] Nie, B. and Sun, S. 2017. Using Text Mining Techniques to Identify Research

[21] Trends: A Case Study of Design Research. Applied Sciences. 7, 401; Doi: 10.3390/app7040401.

[22] Laxman, B. and Sujatha, D. 2013. Improved method for pattern discovery in text mining. International Journal of Research in Engineering and Technology. 2(1), 2321–2328.

[23] Dang, D. S. and Ahmad, P. H. 2015. A review of text mining techniques associated with various application areas. International Journal of Science and Research (IJSR). 4(2), 2461–2466.

[24] Steinberger, R. 2012. A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation. 46(2), 155–176. Doi: 10.1007/s10579-011-9165-9.

[25] Zhong, N., Li, Y. and Wu, S. T. 2012. Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering. 24(1), 30–44. Doi: 10.1109/TKDE.2010.211.

[26] Mukhedkar, B. A., Sakhare, D. and Kumar, R. 2016. Pragmatic analysis based document summarization. International Journal of Computer Science and Information Security. 14(4), 145-149.

[27] Chen, C. P. and Zhang, C. Y. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 275, 314–347. Doi: 10.1016/j.ins.2014.01.015.

[28] Al-Hashemi, R. 2010. Text summarization extraction system (tses) using extracted keywords. International Arab Journal of e-Technology. 1(4), 164– 168.

[29] Jayashankar, S. and Sridaran, R. 2016. Superlative model using word cloud for short answers evaluation in e-Learning. Education and Information Technologies. 22(5), 2383-2402. Doi: 10.1007/s10639-016-9547-0.

[30] DePaolo, C. A. and Wilkinson, K., 2014. Get your head into the clouds: using word clouds for analyzing qualitative assessment data. Tech Trends. 58(3), 38–44. Doi: 10.1007/s11528-014-0750-9.

[31] Sinclair, J., and Cardew-Hall, M. 2008. The folksonomy tag cloud: when is it useful?. Journal of Information Science. 34(1), 15–29. Doi: 10.1177/0165551506078083.

[32] Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J. and McKeon, M. 2007. Many eyes: a site for visualization at internet scale. IEEE Trans. Vis. Computer Graphics. 13(6), 1121–1128.