Acta Scientific COMPUTER SCIENCES

Volume 4 Issue 4 April 2022

Review Article

Basic Protocol to Conduct an NLP Based Study

Santhosh Priya[1]* and R Kalaiarasi[2]

[1]Research Scholar, School of Computer Science, Tamil Nadu Open University, Chennai, India

[2]Assistant Professor, School of Computer Science, Tamil Nadu Open University, Chennai, India

***Corresponding Author:** Santhosh Priya, Research Scholar, School of Computer Science, Tamil Nadu Open University, Chennai, India.

Received: December 25, 2021

Published: March 14, 2022

Abstract

There are tons and tons of raw data available on the Internet, hence, the purpose of this study is to explore the need for text mining with the terminology. This paper details the study of text mining with its application areas and its terminology critiques its working with Natural Language Processing (NLP).

Keywords: Artificial Intelligence; Textmining; Textmining Process; Natural Language Processing; Areas of NLP

Abbreviations

TM: Text Mining; AI: Artificial Intelligence; NLP: Natural Language Processing; HL: Human Language; MIS: Management Information System

Introduction

Text mining is a relatively young concept in the data mining industry for information retrieval. The ability to extract relevant information from natural language text is known as text mining [1]. Text analytics is another name for text mining. Text mining is thought to be a very promising field. Knowledge can be obtained from a variety of sources.

A type of text mining is data mining. For example, data is organized or disorganized. The term "structured" refers to how the data is laid out in rows and columns. Structured data is best represented by a calendar. Data that is structured is simple, verifiable, and simple to comprehend. Unstructured data, as well as semi-structured data, is available in a variety of forms and formats. The retrieving of unstructured information is more challenging and requires intuition.

Unstructured text, which is widely accessible, is thought to account for around 80% of all text [2]. Handling massive amounts of unstructured text from sources such as email, full-text documents, HTML files, and other sources is a time-consuming and costly procedure [3,30].

The value of data in business analysis cannot be overstated. It offers value to the success of the company. Text mining entails not just the extraction of information from any unstructured source, but also the use of Natural Language methods [27]. It incorporates retrieving information as well as a pre-process of verifying the format of text to identify the data as quantitative or qualitative [4]. The analysis phase is completed after the text has been successfully done, with repeated text in the dataset being trimmed down [5]. The last stage of text mining is known as Knowledge or Management Information Systems (MIS).

Text mining involves the following process.

- Collection of data
- Extracting of information
- Preprocessing of format

- Analysis phase key
- Final results outcome in form or MIS

Text mining process

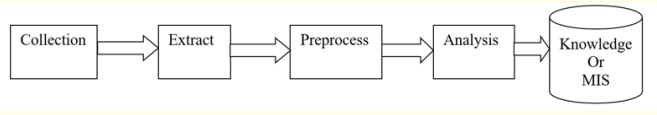

Figure 1

To search for something in a search engine, we must first enter a keyword, such as "Devotion," and the retrieved output will be content connected to the word "Devotion". Now we must decide which document is the most relevant to our search. It has been observed that a keyword search is not always sufficient when findings are speculative as there will be a larger result, therefore the search has to be reduced, but caution must be exercised as this can result in the loss of valuable data [5,28]. It may also make it impractical to analyze every document because each document describing "Devotion" may yield different conclusions and information, necessitating the use of text mining [6,26].

Natural language processing

Natural Language Processing (NLP) is an area of computer science and artificial intelligence concerned with Human Language (HL) [9,29]. it's a known fact that a computer can only read and write 0s and 1s. Text mining includes NLP, which is a part of it [7].

Areas of natural language processing

To the best understanding of the difference between Text mining and NLP. It can be readily observed that text mining is the process to carry the data while NLP is the method to do the processing [8,10]. The following are the different areas where NLP is implemented in form of applications.

- **Sentiment Analysis**: (also known as opinion mining or emotion AI) is a type of artificial intelligence that analyses people's feelings. It's an NLP technique for detecting emotions in the text [11,12,14].
- **Chatbot:** A chatbot is a piece of software that mimics human communication [13].
- **Speech Recognition:** Exercising Recognition Alexa and Google Assistant are two of the most popular virtual assistants on the market today.
- **Machine translation:** Google Translator is the most famous example of machine translation [15,18].
- **Spell Check:** The spell is examined, and the findings are obtained afterward.
- **Keyword search:** After removing any unnecessary terms, the keyword is used to search.
- **Information extraction:** The right data is need to be extracted [25].
- **Advertisement Matching:** The recommendation of ads based on search history is projected [17].

Terminologies used in text mining

- Text cleanup—removes hyperlinks, special characters, and advertisements from websites, as well as figures and formulae from websites and documents [19].
- **Tokenization:** Tokenization is the act of breaking down unstructured data into tokens such as words, phrases, keywords, and other pieces [15].
- **Stemming:** This is a method of reducing words to their simplest form. For example, stemming is frequently used to denote "doing", "done", and "do." Remove all stop words from the manuscript [16], including nouns, verbs, adjectives, pronouns, singular nouns, plural nouns, and other parts of speech.
- N-grams might be used in tokenization. It is necessary to create n-grams to understand the information. For example, "good" may be a positive attitude, while "not" is not, but when "not good" are combined, it becomes a negative feeling [21].

Working on text mining

From the perspective of A REAL-TIME REVIEW, "It's a good phone for the price, but it doesn't charge quickly, even with the 15w plug."

- A two-word combination is referred to as a bi-gram. For example, "good phone" denotes a good phone, but "not charging" denotes a sluggish charging device. As a consequence, this may be used to compare the phone and the battery, with the former being characterized by its capabilities and the latter by its features [20,24].

- A three-word combination is referred to as a tri-gram [22,23]. For example, "excellent phone pricing" denotes a reasonable and worthwhile price, but "not charging quickly" denotes a sluggish charging phone. You'll receive incorrect results if you try to analyze them without using the n-gram. It's pointless to analyze data like "good" "phone" "not" "quick" "charging" "price" individually.

Conclusion

It is a known fact that industries depend on technology for their business development. Hence equivalent importance should be provided to utilize the right way of data, to have an easy way of using the right data for the right purpose with the use of the right technology. Considering the growth, in the trending field of NLP, more research on the effective way in mining the right data for the required purpose should be processed as the mining processes involves various terminology specifically under human supervision. The paper concludes with a study that research should be undertaken in the text mining process which should involve less human supervision.

Bibliography

1. https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

2. https://www.altexsoft.com/blog/structured-unstructured-data/

3. D Khanaferov., *et al*. "Social Network Data Mining Using Natural Language Processing and Density-Based Clustering". 2014 IEEE International Conference on Semantic Computing (2014): 250-251.

4. https://thesai.org/Downloads/Volume7No11/Paper_53Text_Mining_Techniques_Applications_and_Issues.pdf

5. Hernández-Blanco Antonio., *et al*. "A systematic review of deep learning approaches to educational data mining". *Complexity* 2019 (2019).

6. https://www.educba.com/text-mining/

7. ApurwaYadav., *et al*. "A comprehensive review on resolving ambiguities in natural language processing". *AI Open* 2 (2021): 85-92.

8. Kevin Bretonnel Cohen. "Chapter 6 - Biomedical Natural Language Processing and Text Mining". Editor (s): Indra Neil Sarkar, Methods in Biomedical Informatics, Academic Press (2014): 141-177.

9. https://www.sciencedirect.com/science/article/pii/B9780124016781000063

10. https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

11. Saad Saidah and Saberi Bilal. "Sentiment Analysis or Opinion Mining: A Review". *International Journal on Advanced Science, Engineering and Information Technology* 7 (2017): 1660.

12. Gupta Vishal and Gurpreet S Lehal. "A survey of text mining techniques and applications". *Journal of Emerging Technologies in Web Intelligence* 1.1 (2009): 60-76.

13. https://bigdata-madesimple.com/how-do-chatbots-work-an-overview-of-the-architecture-of-a-chatbot/

14. Vijayarani S and Janani Ms. "Text Mining: open Source Tokenization Tools – An Analysis". *Advanced Computational Intelligence: An International Journal (ASCII)* 3 (2016).

15. https://www.kdnuggets.com/2017/09/machine-learning-translation-google-translate-algorithm.html

16. Jivani Anjali. "A Comparative Study of Stemming Algorithms". *International Journal of Computer Applications in Technology* 2 (2011): 1930-1938.

17. Jin-A Choi and Kiho Lim. "Identifying machine learning techniques for classification of target advertising". *ICT Express* 6.3 (2020).

18. http://jkhighereducation.nic.in/jkrjmcs/issue1/15.pdf

19. https://towardsdatascience.com/nlp-building-text-cleanup-and-preprocessing-pipeline-eba4095245a0

20. Shervin Minaee., *et al*. "Deep Learning--based Text Classification: A Comprehensive Review". *ACM Computing Surveys* 54.3 (2022): 1-40.

21. https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058

22. Xiaojin Zhu and R Rosenfeld. "Improving trigram language modeling with the World Wide Web". 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 1 (2001): 533-536.

23. Richard C Angell., *et al*. "Automatic spelling correction using a trigram similarity measure". *Information Processing and Management* 19.4 (1983).

24. Fatma Elghannam. "Text representation, and classification based on a bi-gram alphabet". *Journal of King Saud University - Computer and Information Sciences* 33.2 (2021).

25. https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP=Natural%20language%20processing%20 (NLP)%20is,in%20the%20field%20of%20linguistics

26. https://www.jatit.org/volumes/Vol96No6/4Vol96No6.pdf

27. https://link.springer.com/book/10.1007/978-1-4757-4305-0

28. Nan Li., *et al*. "Using text mining and sentiment analysis for online forums hotspot detection and forecast". *Decision Support Systems* 48.2 (2010).

29. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems May 2019 Paper No.: 185 (2019): 1-12.

30. Fernando Enríquez., *et al*. "A comparative study of classifier combination applied to NLP tasks". *Information Fusion* 14.3 (2012): 1566-2535.