Content Analysis Using Keyword Extraction

Bushra Yesmeen Anika
bushra.yesmeen.anika@g.bracu.ac.bd

Rafid Rahman
rafid.rahman@g.bracu.ac.bd

Md Farhadul Islam
md.farhadul.islam@g.bracu.ac.bd

Md Sabbir Hossain
md.sabbir.hossain1@g.bracu.ac.bd

Annajiat Alim Rasel
annajiat@gmail.com

***Abstract*—The ability to glean relevant information from vast amounts of text data has become more crucial in a variety of industries, including social media analysis, business, and marketing. A method for automatically determining the most crucial words and phrases in a piece of text is a content analysis using keyword extraction. The state-of-the-art methods for content analysis employing keyword extraction are thoroughly reviewed in this study, together with their advantages, disadvantages, and practical uses. We also provide a case study in which these methods were used to examine a sizable collection of customer evaluations and determine the key themes and subtopics. Our findings show how content analysis utilizing keyword extraction may effectively summarize and comprehend enormous quantities of text material.**

***Index Terms*—Content analysis, Keyword Extraction, Natural Language Processing, Data Analysis, Text Mining**

I. Introduction

The amount of digital data produced in recent years has grown exponentially, particularly when it comes to text. As a result, methods for extracting useful information from massive volumes of text data have been developed, including content analysis utilizing keyword extraction. A method for automatically determining the most crucial words and phrases in a piece of text is a content analysis using keyword extraction. This work aims to present a thorough examination of the state-of-the-art methods for content analysis utilizing keyword extraction, including their advantages, disadvantages, and practical applications. Additionally, this paper offers a case study that illustrates how content analysis with keyword extraction can effectively summarize and comprehend enormous amounts of textual data.

II. Literature Review

The fields of natural language processing and data mining have extensively investigated content analysis utilizing keyword extraction. The most popular methods for extracting keywords are TF-IDF, TextRank, and LDA. By comparing a word's frequency in a document to its frequency over the entire corpus, the TF-IDF algorithm determines how significant a word is within that document. Using a graph-based ranking system called TextRank, words in a document are ranked according to how important they are to each other. A topic modeling technique called LDA finds the latent themes in a corpus and distributes each document over them.

III. Methodology

This research paper's methodology contains a case study that shows how content analysis utilizing keyword extraction may be used to analyze a sizable dataset of customer reviews. 10,000 customer reviews for a popular product made up the dataset. Preprocessing of the data included stop word elimination, word stemming, and punctuation removal. Then, the reviews' keywords were extracted using TF-IDF, TextRank, and LDA, and the outcomes were compared.

IV. Results and Discussion

The case study's findings demonstrated that while TF-IDF and TextRank identified comparable keywords, TextRank was superior at collecting multi-word phrases and spotting crucial terms in context. A high-level overview of the themes and issues covered in the reviews was supplied using LDA, which found latent topics in the data. The reviews' collected keywords were utilized to create summaries, pinpoint the significant themes and topics, and guide product changes.

V. Application and Limitation

Market analysis, sentiment analysis, and social media analysis are just a few of the many uses for content analysis with keyword extraction. However, this method has certain drawbacks, including the inability to fully capture a text's meaning and the challenge of sarcasm or irony.

VI. Conclusion and Future Work

Content analysis using keyword extraction is a powerful technique for summarizing and understanding large amounts of textual data. Our case study demonstrated the effectiveness of TF-IDF, TextRank, and LDA for identifying important words and phrases in customer reviews. Future work could explore the use of other keyword extraction techniques and compare the results to those presented in this paper. Additionally, further research is needed to evaluate the effectiveness of content analysis using keyword extraction in other domains and applications.

REFERENCES

[1] Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing Management, 24(5), 513-523.

[2] Mihalcea, R., Tarau, P. (2004). Textrank: Bringing order into text. Association for Computational Linguistics.

[3] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022