Content Analysis Using Keyword Extraction

***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.

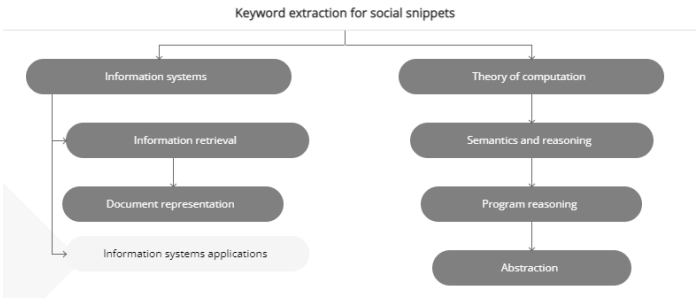


Fig. 1. title

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.

The entropy of cluster j is:

$$E(j) = -\sum_{i=1}^{I} P(i,j)\log_2 P(i,j)$$

The entropy of the entire clusters is the sum of the entropy of each of the clusters weighted by its size:

$$E = \sum_{j=1}^{J} \frac{n_j}{n} E(j)$$

Purity measures how far each cluster contained objects from primarily one class. good clustering solution result from the large purity values. Similar to the entropy, the purity of each cluster is calculated as :

$$P(S_r) = \frac{1}{n_r} \max_i \left(n_r^i\right)$$

Where Sr is a particular cluster of size nr.

The purity of the entire clusters is computed as a weighted sum of the individual cluster purities and is defined as

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r)$$

Fig. 2. title

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. Below the methodology is described step by step:

Table 2. Entropy Results

Datasets	Term Frequency	Word Position	Semantic TF	Semantic WP	Proposed Model
DS1	0.324	0.319	0.312	0.305	0.239
DS2	0.326	0.325	0.313	0.3	0.229
DS3	0.324	0.315	0.313	0.295	0.218
DS4	0.322	0.311	0.311	0.29	0.17

Table 3. Purity Results

Datasets	Term Frequency	Word Position	Semantic TF	Semantic WP	Proposed Model
DS1	0.864	0.868	0.887	0.889	0.909
DS2	0.863	0.866	0.889	0.9	0.917
DS3	0.864	0.869	0.889	0.908	0.922
DS4	0.865	0.889	0.89	0.906	0.94

Fig. 3. title

A. Data Collection

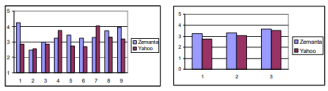


Fig. 1. The average of keyword relevancy grading per presentation. For each of the 9 presentations (X-axis), the users were grading the relevancy of 8 keywords from Zemanta and 10 keywords from Yahoo, with grades 1-5 (5 being the most relevant). The average of grades is calculated for two services separately (Y-axis). The grades for the same keywords were equally distributed among users.

Fig. 2. The average of keywords relevancy grading (Y-axis) per presentation topic (1 – open source, 2 – databases, 3 – gravity) on the X-axis.

The relevant data that will be studied must first be collected as part of the technique. This information may be presented as text documents, posts on social media, news pieces, or in any other way that involves text.

B. Data Preprocessing

Preprocessing is the next stage after data collection. In addition to cleaning and normalizing the text data, this entails deleting any unnecessary or redundant material. To make sure the data is reliable and prepared for analysis, this is done.

C. Keyword Extraction

Extraction of keywords from the text data is the following stage. For keyword extraction, a variety of methods can

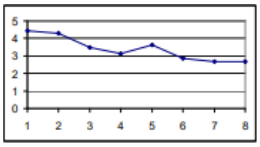


Fig. 3. The average user grading of keywords per particular Zemanta rank. The X-axis presents 8 Zemanta internal ranks. The Y-axis presents the average of user grades for the keywords in each Zemanta rank. In this diagram, the keywords from all 9 presentations were included.

Fig. 4. title

be utilized, including machine learning-based methods like LDA (Latent Dirichlet Allocation) and statistical methods like TF-IDF (term frequency-inverse document frequency) and TextRank. These methods assist in locating the key terms or phrases within the text data that may be utilized to sum up or group the content.

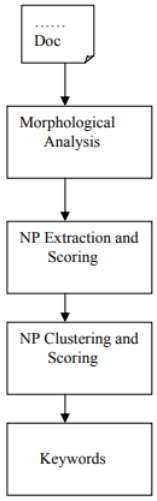


Fig .1. Keyword Extraction Algorithm

D. Keyword Analysis

Analysis comes next after the keywords have been extracted. This entails grouping related terms together and figuring out which ones are used the most. This aids in determining the primary themes or subjects covered in the text data.

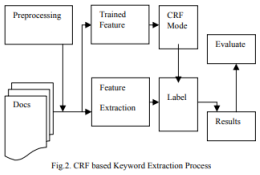
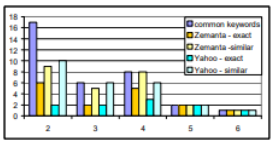


Fig.2. CRF based Keyword Extraction Process

E. Sentiment Analysis

The text data can also be subjected to sentiment analysis in addition to keyword extraction and analysis.



This entails figuring out whether the language is expressing a favorable, negative, or neutral attitude or emotion. This aids in figuring out the content's general tone and might be helpful in applications like customer feedback analysis.

F. Visualization

The final step in the methodology is to visualize the results of the analysis. This can be done using various techniques such as word clouds, bar charts, or network graphs. Visualization helps in presenting the results in an easy-to-understand format and can be useful in communicating the insights to stakeholders.

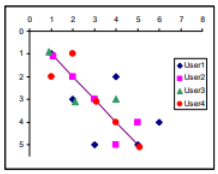


Fig. 5. The relation between the user and Zemanta ranking. The X-axis presents Zemanta ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The ranking itself is marked with a dot of a different type for each user. Ideally, the user and internal rankings would be identical, with all the dots on a diagonal line. Here, the dots are dispersed, but still near the diagonal line. The majority of dots are placed in the first five columns (Zemanta rank 1-5): this shows that users and Zemanta largely agree on what are the 5 most relevant keywords.

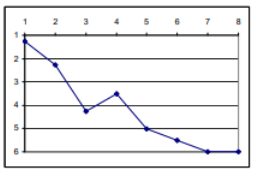


Fig. 6. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for all keywords in a particular Zemanta rank. For instance, the highest-ranked keywords by Zemanta got 1, 1, 1 and 2 as user ranks, which gives an average of 1.25 out of 5. The diagram shows that the user ranking lowers together with Zemanta ranking; the keywords with the lowest Zemanta rankings are not among the most relevant to the users. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

In conclusion, data collection, data preprocessing, keyword extraction, keyword analysis, sentiment analysis, and visualization are all components of the technique for content analysis utilizing keyword extraction. For the process to yield valuable insights from the text data, each stage is crucial.

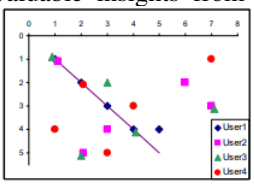


Fig. 7. The relation between user and internal ranking. The X-axis presents Zemanta internal ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The actual ranking is marked with a dot of a different type for each user.

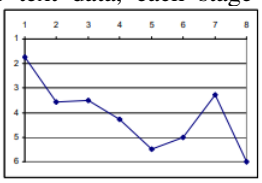


Fig. 8. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for the keywords in a particular Zemanta rank. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

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.

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

Where:

TP= keywords extracted keywords by the algorithm and already found in document's predefined keywords.

FP= keywords extracted keywords by the algorithm and doesn't found in document's predefined keywords.

FN= document's predefined keywords that are not extracted by the algorithm.

Table 1. Precision and Recall results

	Term Frequency	Position Weight	Proposed Model
Precision	77.2	80.3	83.5
Recall	68.5	73.0	76.3

Fig. 5. title

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
[4] https://citeseerx.ist.psu.edu/document?repid=rep1type=pdfdoi=75a587e18fa8386c5bb6 https://citeseerx.ist.psu.edu/document?
[5] repid=rep1type=pdfdoi=75a587e18fa8386c5bb67a69d36251da3ee71fa1page=54
[6] https://dl.acm.org/doi/abs/10.1145/1772690.1772845
[7] https://ieeexplore.ieee.org/abstract/document/9754064

[8] https://citeseerx.ist.psu.edu/document?repid=rep1type=pdfdoi=5620a132b106f8280ab01dc5331c09e7fd41b3bb
[9] https://ieeexplore.ieee.org/abstract/document/1234007