Automation of discussion board evaluation through keyword extraction techniques: a comparative study

Gomathi Thiyagarajan 1,3, Prasanna S 2 and Uma B 3

1Research Scholar, School of Computing Science, VISTAS, Chennai, India
2Professor & Head, Department of Computer Applications, VISTAS, Chennai, India
3Department of Computer Applications, CMR Institute of Technology, Bengaluru, India
gomathi.t@cmrit.ac.in

Abstract. With the current trend in lieu of the ongoing pandemic, online learning and technology-based pedagogy receive worldwide attention. There are several pedagogical practices one can use to evaluate students mindful and the method will vary based on the learning objective. In this paper, our goal is to dive deeper into identifying the best keyword extraction technique which can be closely mapped with human evaluation. We report the results of three popular keyword extraction techniques specifically RAKE, TF-IDF, and Semantic Fingerprinting on the dataset generated from the discussion board post of a Learning Management System. The results illustrate that the TF-IDF algorithm shows the highest correlation with a human evaluation with 0.76 correlations, RAKE with 0.36, and Semantic Fingerprinting with 0.58. It was also identified Semantic Fingerprinting had a lowest mean square error of 18.26.

Keywords: Automatic Keyword Extraction, Discussion Board evaluation, RAKE, TF-IDF, Semantic Fingerprinting

1. Introduction

Educational Institutions are investigating new pedagogical approaches to figure out and handle difficulties such as average class size, facilitate dynamic, active cooperation among different companion of students [1]. Assessment is an essential component for viable learning. As Vonderwell et al. [2] indicated, web-based learning contexts include distinct characteristics when contrasted with face-to-face settings because of the asynchronous collaboration among the teacher and learner. A well-organized strategy [3] and assessment components would be required in place as meaningful collaboration in both online and blended settings are incorporated by effective learning communities.

The discussion board has become the most popular model of interaction to create a virtual community of inquiry, allowing peers to comment, question, and respond. It is available as both stand-alone and built-in tools in Learning Management Systems (LMSs) typically allows multi-threaded discussion and records every post. Moore, R. L et al. [4] endowed that few segments of discussion colloquy backed the advancement of higher-order thinking; they further suggested that the type of doled out assignment and the forum elicit the type of higher-order thinking in a networked conversation. For decades, educationalists have done their research on automating the evaluation process.

Martin F et al. [5] stated the recommendations from faculties for assessment includes traditional, authentic, rubrics, interviews, surveys, systematic learning analytics, and peer reviews. Amongst the various evaluation mechanisms, this study primarily focuses on keyword-based evaluation. In this paper,
our goal is to explore the effective keyword extraction technique that can be used to automate the assessment of the posts in Learning Management Software. Keywords from individual posts were extracted using various techniques and compared with Lecturer given keywords to identify the percentage of matches between the extracted and given keywords. The posts were also manually evaluated and this correlation between the two was presented in this work.

The rest of this paper contains; Keyword Extraction techniques in Section 2 covers the popular technique used to excerpt keywords and discuss the popular algorithms that are being employed to pull out the keywords to automate evaluation. Section 3 and 4 briefly explains the data collection and evaluation method used in the current research and their results. Finally, Section 5 furnishes the conclusion of this study.

2. Keyword Extraction Techniques

Keywords are the small scale units that can sum up the content of a document. Keywords are regularly used for many reasons like recovering records during a web search or summing up reports for indexing and they are likewise used to pin down the most applicable information from the context. Automatic keyword extraction methods are generally used in Text Mining, Natural Language Processing, and data recovery to represent large sets of words to a small bunch of words.

Beliga, S et al. [6] stated different techniques that can be used for extracting keywords are based on statistics, linguistic, graph-based, and machine learning approaches. The table below shows the techniques and various approaches used for keyword extraction. For the current research, we used the RAKE, TF-IDF, and Semantic Fingerprinting technique to extract the keywords from the text, and the flow diagram of the techniques is shown in figure 2.

2.1. RAKE

Rapid Automatic Keyword Extraction (RAKE) [7] [8] is an unsupervised, language, and domain-independent technique for extracting keywords. The RAKE technique depicted below Figure automates the keyword build process quickly and expressively through the following steps.

- A document comprises multiple phrases and sentences from which keywords have to be extracted.
- The input parameter for RAKE consists of stop words, phrase delimiters, and word delimiters.
- RAKE through these input parameters partitions the document into candidate keys.
- The document text is split into an array of words called a candidate key.
- Every candidate key is given a score (degree of frequency) by analyzing the co-occurrences graph. The degree of frequency = Degree of word / Degree of frequency
- The keywords with top T-Scores are selected as keywords for the document.

2.2. TF-IDF

The Term Frequency-Inverse Document Frequency [9] [10] is a methodology used to discover the significance of sentences involving words. It registers values for each word in a record through an opposite extent of the recurrence of the word in a particular report to the level of archives the word appears in. The equation used to process tf-idf is given in (1).

    \[
    \text{tf-idf}_{(t,d)} = \text{tf}_{(t,d)} \times \text{idf}_{(t)}
    \]

The formula that is used to compute the IDF is computed as in (2)

    \[
    \text{idf}_{(t)} = \log \left[ \frac{n}{\text{df}_{(t)} + 1} \right]
    \]
### Table 1. Techniques and approaches to Keyword Extraction

| Approaches                      | Techniques                                      |
|--------------------------------|-------------------------------------------------|
| Simple Statistic Approach      | Word Frequency, Word Co-Occurrences, TF-IDF, RAKE|
| Linguistic Approach            | Discourse Markers, Semantic information, Linguistic information |
| Graph-Based Approach           | Word Co-Occurrences, Labelled Edges, Neighbourhood Size |
| Machine Learning Approaches    | Support Vector Machine, Deep Learning, Conditional Random Fields |

The word with a high TF-IDF score has a solid relationship with the report in which it shows up. The term frequency refers to the heaviness of a term that occurs in a report. Document Frequency is the quantity of archives in which the word is available.

TF - denotes the occurrences of a specific term t in document d  
IDF - denotes the no. of documents in the collection containing the term t.

### 2.3. Semantic Fingerprinting

As per ongoing discoveries in neuroscience, the brain utilizes Sparse Distributed Representations (SDR) to handle data and information. Sparse implies neurons are dynamic simultaneously, contrasts to the common dense and distributed represents the significance of the pattern in it. The concept driving Cortical.io's Retina [11] comprises changing over the symbolic representation of language into an SDR structure with the goal that it turns out to be mathematically computable. Language is first decomposed into words. The terms are then changed over into SDRs by the Retina Engine. Each word is described by a heap of 16,000 semantic features, which catch the lexical semantics of a word in its normal setting. Two words become comparable in their SDR structure if they are reasonably related. Semantic similarity can be computed using a simple measure like the Euclidean Distance. Figure 1 represents the semantic folding model used in the Semantic Fingerprinting approach.

![Semantic Folding](https://www.cortical.io/)  

**Figure 1.** Semantic Folding: a new model for natural language understanding [digital image]. Retrieved from https://www.cortical.io/
The semantic fingerprinting [12] approach decides the significance of each term in the content by producing a semantic fingerprint for each term with the fingerprint of text determining the overlap looking at the frequency of the term in content with the frequency of the term in the whole retina semantic space.

3. Data evaluation technique
The dataset used in the current research consists of 52 responses to a post by the PG (Post Graduate) students on a topic discussed in the course Big Data Analytics. Students were asked to post their understanding of the topic covered in the online class simultaneously through Google Classroom. All forum posts have been first extracted from Google Classroom through the script written in PHP. The name of the student, timestamp, and the posted responses were extracted and placed in a spreadsheet to apply keyword extraction algorithms for instant evaluation. Figure 3 illustrates the data collection method and evaluation strategy carried out by the researchers during this work.
• Post and Receive mainly consists of posting a question in a Learning Management System and receiving responses from learners in a stipulated time.
• Extract here performs extraction of student response from the discussion board (LMS used in current research is Google Classroom). The Word Cloud is shown in figure 4 shows the graphical representation of word frequency that appears in the forum posts by students.
• Extraction also represents extracting keywords using keyword extraction techniques or algorithms. Table 2 displays the detection of keywords using various keyword extraction techniques for a sample post.
• Compare consists of logic to compare the extracted keyword with keywords suggested by experts and its synonym using wordnet.
• Display match % displays the matches between keywords percentage.

Table 2. Sample keyword extracted

| RAKE                | TF_IDF     | Semantic Fingerprinting |
|---------------------|------------|-------------------------|
| max value           | Reduce value| value                    |
| key-value pair      | values     | max                     |
| type of output      | particular values | values                |
| Value               | temperature| phase                   |
| particular year     | year       | output                  |
| reduce phase        | reducer    | pairs                   |
| highest recorded temperature | types | temperature |
| type of input       | Output     | types                   |
| intwritable type    | Key        | key                     |
| Text                | pairs      | variables               |

4. Results and discussion
This study examined the keyword extraction algorithm to perform an instant evaluation of discussion board posts during an online class. In general, it is difficult to set up the best technique for keyword extraction since certain algorithms work better in specific circumstances. In this section, we evaluated three keyword extractions methods namely RAKE, TF-IDF, and Semantic fingerprinting methods with the human evaluation of the discussion board post. The results of correlation between Human, RAKE, TF-IDF, and Semantic fingerprinting is represented in figure 5. TF-IDF outperforms the other approaches by returning relevant keywords to a particular query and shows the nearest correlation with human evaluation. RAKE performs the weakest, due to optimizing long keyword for accurate phrases. The third method semantic fingerprinting performs better than RAKE and it represents text as a semantic fingerprint by using a Boolean vector with the lowest error rate. Figure 6 shows the Pearson’s r correlation between Human evaluation and RAKE/TF-IDF/Semantic Fingerprinting evaluation.
### Table 3. Comparison between various techniques

| Source                      | Correlation Coefficient | Mean Square Error | Root Mean Square Error | Mean Absolute Error | Coefficient of Determination |
|-----------------------------|--------------------------|-------------------|-----------------------|--------------------|-----------------------------|
| RAKE                        | 0.358                    | 43.082            | 6.564                 | 5.888              | 0.082                       |
| TF-IDF                      | 0.763                    | 32.354            | 5.688                 | 4.768              | 0.536                       |
| Semantic Finger Printing    | 0.577                    | 18.257            | 4.273                 | 3.419              | 0.261                       |

![Figure 5. Correlation Matrix](image)

![Figure 6. Correlation plot](image)

### 5. Conclusion

This study primarily focused on identifying the keyword extraction technique which can be used to instantly evaluate the discussion board post of students based on their understanding of the topic covered in a session through online mode. This can be an effective pedagogical evaluation for the faculties to quickly get to know the understanding of the just-concluded session and to effectively plan for the next lecture. In the present work for retrieving the posts from the discussion forum we have written the code in PhP and applied keyword extraction techniques on the retrieved results. To compare the extracted keywords, we have written the code using Perl script and identified the percentage match between the human evaluation and keyword extraction technique.

The results laid out in the current study show that the TF-IDF keyword extraction method is the closest technique to human evaluation and can be used instantly to evaluate the post of the discussion forum in a Learning Management System. The present study also presented the various measures of errors and Coefficient of determination of various keyword extraction techniques (see Table 3) between human evaluation and keyword extraction. Furthermore, manual investigation of understanding the learning outcome during online classes for each session is not feasible due to the size and duration of the class. A well-established pedagogical model would help in such investigation and make the teaching and learning effectively in online mode.

### 6. Declaration

The authors declare that consent from students was collected to use the student data in the current research.
References

[1] Seethamraju R 2014 Effectiveness of Using Online Discussion Forum for Case Study Analysis Education Research International 1–10

[2] Vonderwell S 2007. Asynchronous Discussions and Assessment in Online Learning 39(3) 309–328. Retrieved from http://files.eric.ed.gov/fulltext/EJ768879.pdf

[3] Gikandi, J. W., Morrow, D., and Davis, N. E 2011 Online formative assessment in higher education: A review of the literature Computers and Education 57(4) 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

[4] Moore, R. L., Oliver, K. M., and Wang, C. 2019 Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis Interactive Learning Environments 27(5–6) 655–669

[5] Martin, F., Ritzhaupt, A., Kumar, S., and Budhrani, K 2019 Award-winning faculty online teaching practices: Course design, assessment and evaluation, and facilitation Internet and Higher Education 42 34–43

[6] Mahrishi M., Morwal S. (2020) “Index Point Detection and Semantic Indexing of Videos—A Comparative Review”. In: Pant M., Kumar Sharma T., Arya R., Sahana B., Zolfagharinia H. (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol 1154. Springer, Singapore.

[7] Rose, S., Engel, D., Cramer, N., and Cowley, W 2010 Automatic Keyword Extraction from Individual Documents Text Mining: Applications and Theory March 1–20. https://doi.org/10.1002/9780470689646.ch1

[8] Gagliardi, I. and Artese, M. T 2020 Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods Multimodal Technologies and Interaction 4(2) 1–20

[9] Aizawa, A 2003 An information-theoretic perspective of tf-idf measures. Information Processing and Management 39(1) 45–65

[10] Zhang, W., Yoshida, T., & Tang X 2011 A comparative study of TF*IDF, LSI, and multi-words for text classification Expert Systems with Applications 38(3) 2758

[11] Sparse Distributed Representations (SDRs). (n.d.). Retrieved November 14, 2020 from https://www.cortical.io/science/sparse-distributed-representations/

[12] Govind, Alec C., Spaniol M 2018 Semantic Fingerprinting: A Novel Method for Entity-Level Content Classification. In: Mikkonen T., Klamma R., Hernández J. (eds) Web Engineering. ICWE 2018 Lecture Notes in Computer Science 10845 Springer Cham.