Identifying Content-Related and Non-content-related Queries in Online Discussion Forums Using Voyant Tools

Neha* and Eunyoung Kim

***Abstract*—The overarching goal of this study was to assess the suitability of Voyant Tools to identify the frequency of content-related and non-content-related query subjects (thread title) and prioritize them based on their occurrence and importance in the online discussion forum. The dataset consisted of 296 query subjects collected from the discussion forums of practical and theoretical massive open online courses (MOOCs). The cirrus, correlation, and scatter plot features of Voyant Tools (a web-based application) were used to analyze the dataset. The Cirrus feature assisted with word frequency, and the Correlation feature helped with their co-occurrence in the online discussion forum. The Scatter plot feature was the most appropriate tool among the three tools implemented in the current study for generating the clusters of content-related and non-content-related query subjects. Overall, Voyant Tools was an effective resource capable of analyzing quantitative and qualitative data and providing visual output in various forms.**

***Index Terms*—Massive open online courses (MOOC) discussion forum, query analysis, Voyant tools**

I. Introduction

In recent years, interest in online education has increased due to the needs of learners for flexible learning hours, an adaptive study environment, and access to distance education [1]. In the past two years, online education has become an essential part of our education system due to the COVID-19 global pandemic [2–4]. Due to COVID-19, recent trends in online education, including obtaining knowledge and developing practical and communication skills, have led to an increase in the number of learners engaged in distance learning [5].

Massive open online courses (MOOCs) are one of the leading platforms for online education. Some of the examples of MOOCs are Coursera, edX, Udacity, and Udemy. In recent times, MOOCs has gained attention because of flexibility in participation, motivation, language, and open access to lifelong-learning opportunities because they provide substantial content to a large number of learners in a cost and time-efficient way [6, 7]. In terms of time efficiency, MOOCs provide pre-recorded video lectures convenient for both learners and instructors [8] but result in less quality interaction with other learners and instructors than in traditional classroom settings. This lack of interaction and collaboration has been identified one of the primary reasons for learner dropout from MOOCs [9].

Manuscript received January 4, 2022; revised March 4, 2022; accepted March 20, 2022.

The authors are with the School of Knowledge Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan.

*Correspondence: neha11@jaist.ac.jp (N.)

Online discussion forums are one of the primary platforms for interaction among learners and instructors in MOOCs. In MOOCs, the discussion forum is also considered a primary platform for knowledge construction through social interaction, sharing information, egocentric elaboration, allocentric elaboration, application and transfer, coordination, and reflection [10]. Consequently, MOOC discussion forums provide a wide opportunity for researchers for data mining due to the broad range of online courses and various type of learners [11]. The existing literature on discussion forums focuses specifically on learning behavior patterns [10, 12, 13]. However, most of these studies focused only on features used to find content-related queries in the discussion forum [14–16].

A well-structured discussion forum is considered an essential requirement for smooth interaction and collaboration [17, 18]. Currently, a systematic understanding of how discussion forum posts contribute to interaction and collaboration is still lacking. This study aimed to analyze online discussion forums and identify content-related and non-content-related queries using query subjects. Previous approaches were limited to analyzing content-related queries and learning behavior. This study aimed to determine the type of queries using Voyant Tools to better understand online forum discussions. Initially, extracting data from the discussion forums is a resource-intensive task. Several approaches have been used to extract discussion forum data, such as the Board forum [19], iRobot [20], and Vigi4Med [18]. These forum-based data extraction approaches were powerful in obtaining all the information from the web page of the discussion forum. However, only query subjects were used to analyze the discussions in this study, so Voyant Tools was a more effective and more straightforward approach to extracting and analyzing the dataset. Voyant Tools, which includes the Cirrus, Correlation, and Scatter plot features, was also previously found to be successful in discovering words and their relationships from the massive dataset [21, 22].

This paper is divided into four parts:

1) A brief overview of the recent history of work in the field of online discussion forum analysis;
2) The methodology followed for this study;
3) The experimental results obtained using Voyant Tools;
4) Finally, the conclusions of the present study.

Through this research, we chose to use Voyant Tools in the data mining of the discussion forum and further investigate the feasibility of using that tool to identify key discussion topics of content-related and non-content-related queries. The Voyant Tools comprises a wide range of features for evaluating term frequencies and distributions and can support various data formats, including plain text, HTML, XML, MS

doi: 10.18178/ijiet.2023.13.4.1845

Word, or PDF. In this study, one of the primary purposes of using Voyant Tools was to stop the word-removal feature, automatically removing the stop words such as in, for, of, the, and as from documents. This feature distinguishes Voyant Tools from other software such as Concordle or Leximancer used for similar purposes.

II. Related Work

A. Research and Efforts to Address Overload in Online Discussion Forum

Generally, MOOCs are characterized by many learners; subsequently, their discussion forums become difficult for instructors to manage, and the quality of the interaction between learners and instructors gradually declines. Baek *et al*. found that the size of the discussion forum is linearly dependent on the content contribution of each participant [23]. However, the contribution of online discussion forums due primarily to the participation of some active learners. Gillani *et al*. noted that higher-performing learners engage more actively in the discussion forum than low-performing learners. However, high-performing learners were not interacting with other high-performing learners [24]. The literature on discussion forums also highlighted the patterns of learner interaction through various parameters such as cognitive engagement, critical thinking, coherent dialogue, and interactive dialogue in learner posts. However, Thomas observed that interactive, collaborative learning did not occur in the online discussion forums analyzed [25]. Previous research also established the relationship between confusion in the learning process due to the massive learner's posts and the rate of dropout from MOOCs [26]. The intermingling of queries with different categories was also observed in the discussion forum posts [27]. Xing *et al*. tracked discussion forum posts to examine the relationship between the achievement emotions of learners and the weekly dropout rate [28].

Manual analysis of text or discussion forum posts is time-consuming and resource demanding. Therefore, several tools and techniques were developed to analyze discussion forum posts automatically [7, 29, 30]. These techniques include supervised learning, unsupervised learning, and text mining. Text mining is one of the techniques most widely used in the analysis of discussion and feedback forums [31, 32]. Text mining transfers unstructured text into a structured form using word trends, patterns, and categorization with keywords to identify valuable information. Unsupervised text mining [33] collected convincing keywords from four different models, including latent dirichlet allocation (LDA), Key phrase extraction (topic rank), Text rank, and Frequency-based word cloud to form word clusters that identify the topic of discussions. Peng *et al*. focus on the discussion posts, including the number of posts, words in the post, discussion topics, learner's emotions, learner's behavior, word index, time index, and topic time distribution using the text mining technique [34]. Text mining can be accomplished using Voyant Tools to extract useful information from massive text datasets of any format [35]. In an unstructured feedback forum, Voyant Tools worked well in the textual analysis function [21].

The literature on facilitating online discussion forums suggests that the forum can be divided into social-emotional and group discussion (content and task-oriented) using a discussion rubric to make it more productive [36]. This study considered group discussion to be content-related posts and social discussion to be non-content-related posts.

B. Research and Efforts to Classify Posts in MOOC Forums

Several methods have been used to classify MOOC discussion forum posts. A large volume of published articles describes the role of data mining tools in assessing asynchronous discussion forums. Using data mining techniques and diverse visualization, successful modeling opportunities can be found for discussion forum posts [31]. Rovai developed a model using the concept of design and facilitation to construct practical knowledge through social-emotional and task-related discussions [36]. Brinton *et al*. focused on improving the quality of online discussion forums through the learners' activities that lead to course-relevant discussions [37]. Yusof *et al*. [38] developed a model for community question answering that determines the quality of questions using good and bad questions to facilitate relevant queries in the discussion forum. Most forum users do not consider other learners' similar questions or repetitions of the same questions before starting the new discussion; eventually, it increases the load in online discussion forums.

Ocharo *et al*. developed a model to make reviewer comments more meaningful by categorizing them into content-related and non-content-related for revising the entire document, but the model was limited to research articles [39]. Other efforts have been made to identify content-related queries from MOOC discussion forums using queries' starting posts and linguistic features [27]. However, that study failed to identify non-content-related queries found in large numbers in any discussion forum. Later, Feng *et al*. developed a language and content independent model to analyze discussion threads using twenty-three limited interactive features, including structure, popularity, and social work [40]. Xing *et al*. also extracted seven features, including language summary features, linguistic features, grammar, punctuation, function words, and social and LDA topics to identify the learner's expression in the online discussion forum [28]. Subsequently, a topic tracking model was created using thread posting, replying, quoting, and common posting labels instead of linguistic features [34]. The use of common posting labels aids in creating clusters of content-related queries and non-content-related queries.

III. Research Design

This study aimed to create clusters of content-related and non-content-related queries as query subjects from discussion forum posts. Non-content-related query subjects are those that have social, technical, and management queries.

These queries are related to submitting assignments, tests, or quiz-related queries that forum administrators can answer. Content-related query subjects are those that contain subject-specific words, and which vary from course to course. In this study, technical queries were considered content-related queries.

The primary task in carrying out this study is data pre-processing. In previous studies, several classification models were used to pre-process discussion forum data. Linguistic features were used in the classification models to categorize the queries [41, 42]. Previous research was also limited to conduct a study on the online discussion forum of two programming computer courses or theoretical courses instead of comparing the practical and theoretical course discussion forums in MOOCs [43].

While Voyant Tools is widely used for both qualitative and quantitative measurement, to our knowledge, no previous work has employed this tool to focus on non-content-related query subjects of Google group discussion forums. Voyant Tools works with the text analysis and data exploration functions to visualize datasets that traditional means may not achieve [44]. Text mining has also been widely used in analyzing various aspects of discussion forum posts. However, in this study, we focused instead on the usability of the Cirrus, Correlation, and Scatter plot features of Voyant Tools to analyze queries. Voyant Tools also enables us to remove stop words automatically. Therefore, Voyant Tools is a particularly suitable option for digital humanities study [45].

The content of a discussion forum can be analyzed at various levels (query subjects, queries itself, replies, number of views on the query). For the following three reasons, the query subject (query thread) was the most valuable unit of analysis for creating a model:

1) MOOC Google group discussions are represented to learners as a threaded conversation in the form of query subjects providing an idea of the query type. Learners decide what to read based on this query subject.
2) Query subjects may change direction when other learners join the conversion. To verify this, an analysis was conducted to check each query conversation.
3) Query subjects aid in identifying relevant features for categorization and creating a model for clustering queries. In particular, the Scatter plot feature in Voyant Tools can aid in clustering queries with similar attributes. A previous study found that instructors responded more frequently to clusters of overlapping forum queries [46].

This study aimed to determine whether the query subject could accurately obtain the idea of a query and further aid in creating clusters of similar queries using Voyant Tools.

The research supporting this study had three primary goals. It sought to determine if:

1) Query subjects of content-related threads have linguistic features that distinguish from non-content-related posts.
2) Linguistic features can be used to create a model that reliably identifies query subjects of content-related posts in a MOOC Google group discussion forum.
3) Voyant Tools can determine the frequencies of content-related and non-content-related query subjects.

IV. Method

A. Data Sources

This study was conducted on text data from two MOOCs. The analysis was undertaken explicitly on a practical course and a theoretical course offered by study webs of active learning for young aspiring minds (SWAYAM). This open-source platform is the provider of MOOCs initiated by India's government for online education (https://swayam.gov.in/global).

B. Targeted Courses

The practical course, C and C++ (named for two programming languages) is ongoing, while the theoretical course, computer networks, has finished. The practical course comprises 20 audio-video spoken tutorials for learning the programming language. In contrast, the theoretical course is a 12-week course with three videos each week and an online quiz every week. There was a significant disparity in the number of learners enrolled in the practical course (26,721) and theoretical course (11,973). Learners in both the courses were invited to post questions and comments in the discussion forum of Google groups. The duration of both courses was 12 weeks, each was taught by a single instructor.

C. Dataset

The dataset consists of a selection of discussion forum posts from the two courses. Query subject (thread title) from the discussion forum was included in the dataset as shown in Table I.

TABLE I: Sample Dataset

Query number	Query subject	Type of query
1	C in windows	CR
2	Lecture slides	CR
3	for subject assignment	NCR
4	matrix multiplication	CR
5	Certificate Detail reg	NCR
6	thankyou	NCR
7	System operator	CR
8	About certification	NCR
9	related test	NCR
10	confusion	CR
11	assignment	NCR
12	Regarding assignment	NCR
13	Error in C++	CR
14	c	CR
15	online test	NCR

D. Data Preparation

1) Data sample

The entire set of 296 query subjects was analyzed. These query subjects were labeled manually as content-related or non-content-related as shown in Table I. During the evaluation phase, the results from Voyant Tools were compared with the result obtained by manual coding. CR (Content-related) queries contain domain-specific keywords for the course and direct academic queries. They require appropriate discussions between learners and instructors. NCR (non-academic or management) queries are related to assignments, quizzes, and tests related queries.

Examples of query subject pre-processing:

Content-related example:

- Before pre-processing: compilation of the program in Windows OS.
- After pre-processing: compilation, program, Windows, OS.

Non-content-related example:

- Before pre-processing: how to submit an assignment.
- After pre-processing: submit, assignment.

In the examples, stop words like "he, of, in, I, how" were removed manually. Stop words occur frequently and have no meaning in the context of queries. Content-related query words such as compilation, program, Windows, and OS are retained. Likewise, words such as submit, and assignment were considered for the categorization of non-content-related queries.

2) Tool used

The entire 296 query subjects were analyzed using the online Voyant Tools (a web-based application). Duplicate or repetitive query subjects were also considered to determine their frequency of occurrence. Three different tools out of twenty-four were used for experiments. These tools were the Cirrus tool, Correlation tool, and Scatter plot tool.

3) Cirrus tool

Cirrus is a word cloud generator that creates a visual image by ranking the words of a corpus or document according to the frequency of occurrence. It automatically filters stop words, saving time and effort during data pre-processing.

4) Correlation tool

This tool enables us to find the co-occurrence of two or more lexical items. The co-occurrence can be positive or negative. The tool provides correlating pairs of words in the text. This tool aids in finding the meaning of the query directly without focusing on stop words.

5) Scatter plot tool

This tool is designed for data visualization using dimensionality reduction methods. It includes analysis functions with the dimensional representation of the data. The Analysis function provides four techniques. These are Principal component analysis, Correspondence analysis, t-SNE (t-Distributed Stochastic Neighbour Embedding) analysis, and document similarity.

In this study, t-SNE analysis was carried out on the qualitative textual data. The t-SNE technique aids in finding the most complex information [47]. Data were analyzed using the parameters perplexity (P) and iteration (I). The level of perplexity ranges from 5 to 100, and the number of iterations that can be performed is between 100 and 5000.

E. Feature Extraction and Modeling

Discussion forum queries include course content-specific queries, repetitive queries, queries categorized as frequently asked questions (FAQ), and management queries. Several words in the query subject were used to classify various categories of queries. Feature extraction aided in data mining of discussion forum queries. In the current dataset, there were 1,182 total words with 392 unique word forms. However, the most frequent words were provided by Voyant Tools for feature extraction (Table II). The most frequent words in the discussion were associated with non-content-related queries, i.e., assignment, exam, certificate, regarding, and test. Finding content-related features from the online discussion forum corpus was difficult because the keywords varied from course to course. This study lacked various categories of classification because it was limited to two discussion forums.

TABLE II: SUMMARY OF THE DATASET

Words	Quantity
Total words	1,182
Unique word forums	392
Vocabulary density	0.33
Average words per sentence	62.2
Most frequent words	assignment (45); exam (39); certificate (30); test (22)

V. RESULTS AND DISCUSSION

A. Features Identification to Classify Content-Related and Non-content-related Query Subjects (Thread Title)

The cirrus tool and the correlations tool facilitated in identifying the features to classify query subjects. The cirrus tool aids in identifying the most frequent discussions. The tool helped in investigating the high-frequency queries in both the practical and the theoretical courses. The high frequency of non-content-related queries for instance assignment, exam, certificate, and test can be seen in Fig. 1.

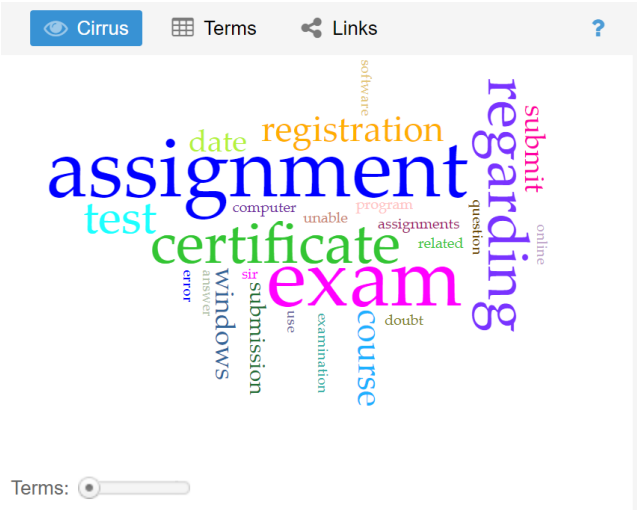

Fig. 1. Word cloud using cirrus tool.

Table III shows several collocations such as "test date", "assessment date", "exam date", and "test week" from the dataset of query subjects. Each collocation shows a perfect positive relationship (value is greater than zero). The strongest correlations between the words of the query subject (Term 1 and Term 2) were determined using a correlation tool. The correlation of words describes the query type without reading the complete query detail. For example, "test date" describes the learner was asking regarding the date of the test; "test week" describes asking the week of the test. These features were later used to classify content-related queries and non-content-related queries.

TABLE III: CORRELATIONS OF WORDS IN QUERY SUBJECTS

Term 1	Term 2	Correlation	Significance
Date	Test	0.97	0.0000022
Function	Submission	0.96	0.0000039
Date	Week	0.95	0.0000251
Correct	Submitted	0.93	0.0000636
File	Header	0.93	0.0000636
Start	Time	0.93	0.0000636
Completion	Related	0.93	0.0000950
Matrix	Related	0.93	0.0000950
Operating	Related	0.93	0.0000950
Enquiry	Examination	0.92	0.0001175
Examination	Wrong	0.92	0.0001175
Test	Week	0.92	0.0001277
Assessment	Date	0.92	0.0001341
Dates	Exam	0.92	0.0001525
Exams	Sir	0.91	0.0002267

B. Feasibility of Three Tools from the Toolkit of Voyant Tools in Clustering Content-Related and Non-content-related Query Subjects

The priorities of queries can be set using the cirrus tool and the instructor can focus according to the significance of the queries in the online discussion forums. In the case of two discussion forums of practical and theoretical courses, the most frequent query asked by the learner was assignment related so it can be considered to the prioritized query. It should be answered first by the instructors. However, the cirrus tool cannot generate clustering of content-related and non-content-related query subjects. Query subjects (thread title) are short headings to introduce the query type. However, several query subjects were found which were too long and time-consuming for the reader and writer to understand the query type. Using the Voyant Tools, these types of long query subjects were dealt with efficiently by removing unnecessary stop words.

The Scatter Plots were created using the t-SNE tool. The technique behind the analysis is tf-idf (term frequency-inverse document frequency). The tf-idf aids in determining the importance of a word in the document and clustering of terms. The range of perplexity (P) and the number of iterations (I) are the two important factors for an analyst. A high level of data resolution and more accurate data interpretation can be achieved by implementing perplexity and iterations [48].

In this study, seven experiments were carried out in three phases. In the first phase of the experiment, clusters were generated by setting the value of perplexity as to the highest range and the number of iterations as to the lowest range. However, the clusters were not sufficiently accurate to distinguish content-related and non-content-related queries (Fig. 2).

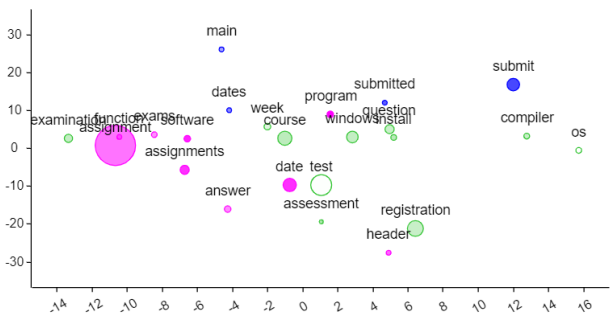

Fig. 2. t-SNE generated clusters at P and I =100 (same).

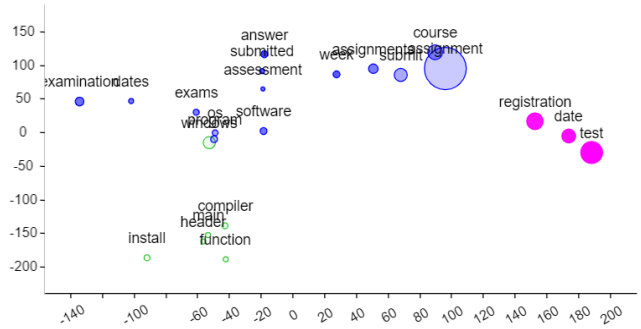

Fig. 3. t-SNE generated clusters at P=5, I=2000.

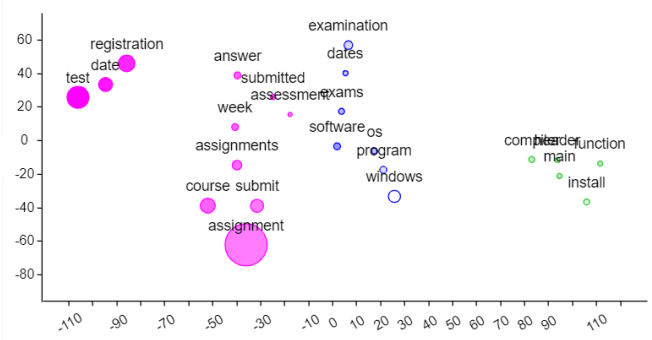

Fig. 4. t-SNE generated clusters at P=5, I=3500.

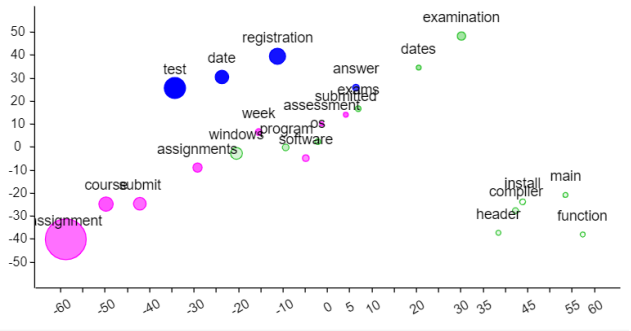

Fig. 5. t-SNE generated clusters at P=5, I=5000.

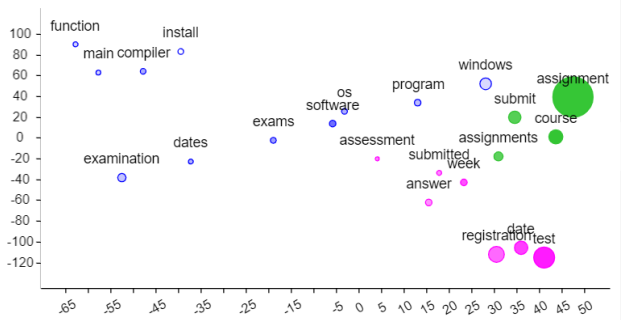

Fig. 6. t-SNE generated clusters at P=100, I= 1000.

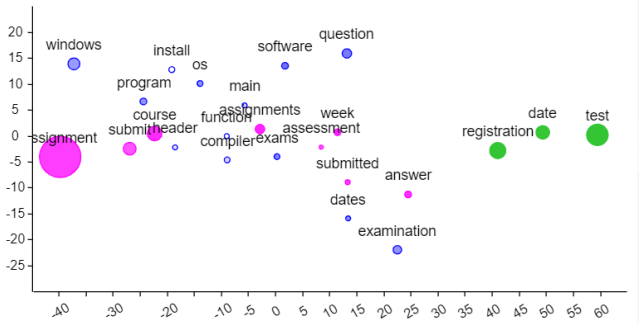

Fig. 7. t-SNE generated clusters at P=100, I=1500.

In the second phase, the level of perplexity remained

constant, with varying iterations to test how the model changes at different iterations (Figs. 3, 4, 5). In the third phase, the level of perplexity was set to the highest value (Figs. 6, 7). The results were able to demonstrate the clusters of content-related queries and non-content-related queries independently. The finding suggested that, in general, there were more non-content-related queries than content-related queries. The best result in a clustering of queries of content-related and non-content-related queries was found with P = 100 and I = 1000, as shown in Fig. 6. We observed that increased perplexity yields better visualization in our dataset. The Voyant Tools is effective in analyzing discussion forum query subjects in terms of quality and quantity. The findings of this study indicate that the Voyant Tools offer several benefits to researchers and that their use as a tool for discussion forums should be further explored.

VI. Conclusion

The main aim of this study was to assess the suitability of the Voyant Tools as a means of identifying the frequency of content-related and non-content-related query subjects. The Cirrus tool and the Correlation tool performed well in prioritizing query subjects based on their occurrence and importance in the online discussion forum. The study used the scatter plot tool to identify the features for automatically clustering various queries as either content-related or non-content-related. The Voyant Tools was found useful in expanding our understanding of how various toolset can be used to analyze quantitative and qualitative data. A subsequent finding was that the tools were more accurate in identifying non-content-related queries than content-related queries. The finding suggests that repeated non-content related queries can potentially be set as FAQs to reduce the burden on instructors. Removal of non-content-related queries can increase the discussion of relevant content-related queries.

During the experiment with the scatter plot tool, it was observed that the frequency of content-related queries and non-content-related queries also varies between practical and theoretical course discussion forums. For example, learners asked more content-related queries in the practical course discussion forum than in the theoretical course forum.

The small size of the dataset allowed us to identify limited features which can be used to categorize various queries. Despite the relatively limited data sample of 296 query subjects, this work offers valuable initial insights into the differences of discussions in practical and theoretical courses. This research has also raised questions regarding the need for further investigation of different types of online courses. For example, a significant difference in content-related queries was observed in theoretical and practical course discussions; identifying the factors of interaction based on the queries may be a fruitful area for future research. Large randomized controlled trials of the t-SNE tool could also potentially provide more definitive evidence.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Neha carried out the literature review, data collection, data analysis, and manuscript drafting. Prof. Eunyoung Kim was involved in revising it critically and giving final approval of the version to be submitted. All authors read and approved the final manuscript.

Funding

This work was supported by the 2021 KDDI Foundation International Students Scholarship.

References

[1] M. D. B. Castro and G. M. Tumibay, "A literature review: Efficacy of online learning courses for higher education institution using meta-analysis," *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 1367–1385, 2021, DOI: 10.1007/s10639-019-10027-z.

[2] H. Baber, "Modelling the acceptance of e-learning during the pandemic of COVID-19-A study of South Korea," *Int. J. Manag. Educ.*, vol. 19, no. 2, 100503, 2021, DOI: 10.1016/j.ijme.2021.100503.

[3] M. D. H. Rahiem, "Remaining motivated despite the limitations: University students' learning propensity during the COVID-19 pandemic," *Child. Youth Serv. Rev.*, vol. 120, no. July 2020, 105802, 2021, DOI: 10.1016/j.childyouth.2020.105802.

[4] T. Muthuprasad, S. Aiswarya, K. S. Aditya, and G. K. Jha, "Students' perception and preference for online education in India during COVID -19 pandemic," *Soc. Sci. Humanit. Open*, vol. 3, no. 1, 100101, 2021, DOI: 10.1016/j.ssaho.2020.100101.

[5] Z. Ma *et al.*, "The impact of COVID-19 pandemic outbreak on education and mental health of Chinese children aged 7–15 years: An online survey," *BMC Pediatrics*, 2021.

[6] J. Reich and J. A. Ruipérez-Valiente, "The MOOC pivot," *Science* , vol. 363, no. 6423, pp. 130–131, 2019, DOI: 10.1126/science.aav7958.

[7] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised Modeling for Understanding MOOC Discussion Forums: A Learning Analytics Approach," in *Proc. the Fifth International Conference on Learning Analytics And Knowledge*, 2015, pp. 146–150, DOI: 10.1145/2723576.2723589.

[8] J. Chauhan and A. Goel, *An Analysis of Video Lecture in MOOC*, Jan. 2015.

[9] D. Gamage, I. Perera, and S. Fernando, "MOOCs lack interactivity and collaborativeness: Evaluating MOOC platforms," *Int. J. Eng. Pedagog.*, vol. 10, no. 2, pp. 94–111, 2020, DOI: 10.3991/ijep.v10i2.11886.

[10] V. Vasodavan, D. DeWitt, N. Alias, and M. M. Noh, "E-moderation skills in discussion forums: Patterns of online interactions for knowledge construction," *Pertanika J. Soc. Sci. Humanit.*, vol. 28, no. 4, pp. 3025–3045, 2020, DOI: 10.47836/PJSSH.28.4.29.

[11] J. Reich, "Rebooting MOOC research," *Science*, vol. 347, no. 6217, pp. 34–35, 2015, DOI: 10.1126/science.1261627.

[12] A. Rantanen, J. Salminen, F. Ginter, and B. J. Jansen, "Classifying online corporate reputation with machine learning: A study in the banking domain," *Internet Res.*, vol. 30, no. 1, pp. 45–66, 2019, DOI: 10.1108/INTR-07-2018-0318.

[13] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, W. Alnumay, and A. P. Smith, "A lexicon-based approach to detecting suicide-related messages on Twitter," *Biomed. Signal Process. Control*, vol. 65, no. December 2020, 102355, 2021, DOI: 10.1016/j.bspc.2020.102355.

[14] Y. Cui and A. F. Wise, "Identifying content-related threads in MOOC discussion forums," *L@S 2015 - 2nd ACM Conf. Learn. Scale*, pp. 299–303, 2015.

[15] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Comput. Educ.*, vol. 68, pp. 458–472, 2013, DOI: 10.1016/j.compedu.2013.06.009.

[16] R. M. Marra, J. L. Moore, and A. K. Klimczak, "Content analysis of online discussion forums: A comparative analysis of protocols," *Educ. Technol. Res. Dev.*, vol. 52, no. 2, pp. 23–40, 2004, DOI: 10.1007/BF02504837.

[17] Neha and E. Kim, "Investigating responsible factors for interaction between learners and instructors in the discussion forum of MOOC," pp. 204–207, 2021.

[18] B. Audeh, M. Beigbeder, A. Zimmermann, P. Jaillon, and C. Dric Bousquet, "Vigi4Med Scraper: A framework for web forum structured

data extraction and semantic representation," 2017, DOI: 10.1371/journal.pone.0169658.

[19] G. Yan, L. Kui, Z. Kai, and Z. Gang, "Board Forum Crawling: A web crawling method for Web forum," in *Proc. - 2006 IEEE/WIC/ACM Int. Conf. Web Intell. (WI 2006 Main Conf. Proceedings), WI'06*, pp. 745–748, 2006.

[20] R. Cai, J. M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An intelligent crawler for web forums," in *Proc. 17th Int. Conf. World Wide Web 2008, WWW'08*, no. January, pp. 447–456, 2008.

[21] I. D. Maramba *et al.*, "Web-based textual analysis of free-text patient experience comments from a survey in primary Care," *JMIR Med. Informatics*, vol. 3, no. 2, pp. 1–13, 2015, DOI: 10.2196/medinform.3783.

[22] S. Sinclair and G. Rockwell, "Text analysis and visualization," *A New Companion to Digit. Humanit.*, pp. 274–290, 2015, DOI: 10.1002/9781118680605.ch19.

[23] J. Baek and J. Shore, "Forum size and content contribution per person: A field experiment," *SSRN Electron. J.*, no. May, 2019, DOI: 10.2139/ssrn.3363768.

[24] N. Gillani and R. Eynon, "Communication patterns in massively open online courses," *Internet High. Educ.*, vol. 23, pp. 18–26, 2014, DOI: 10.1016/j.iheduc.2014.05.004.

[25] M. Thomas, "Learning within incoherent structures: The space of online discussion forums," *J. Comput. Assist. Learn.*, vol. 18, pp. 351–366, Sep. 2002, DOI: 10.1046/j.0266-4909.2002.03800.x.

[26] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rosé, "Exploring the effect of confusion in discussion forums of massive open online courses," *L@S 2015 - 2nd ACM Conf. Learn. Scale*, pp. 121–130, 2015.

[27] A. F. Wise, Y. Cui, W. Q. Jin, and J. Vytasek, "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling," *Internet High. Educ.*, vol. 32, pp. 11–28, 2017, DOI: 10.1016/j.iheduc.2016.08.001.

[28] W. Xing, H. Tang, and B. Pei, "Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs," *Internet High. Educ.*, vol. 43, no. May, p. 100690, 2019.

[29] R. Anbalagan, A. Kumar, and K. Bijlani, "Footprint model for discussion forums in MOOC," *Procedia Comput. Sci.*, vol. 58, pp. 530–537, 2015.

[30] A. Ntourmas, S. Daskalaki, Y. Dimitriadis, and N. Avouris, "Classifying MOOC forum posts using corpora semantic similarities: A study on transferability across different courses," *Neural Comput. Appl.*, 0123456789, 2021, DOI: 10.1007/s00521-021-05750-z.

[31] L. P. Dringus and T. Ellis, "Using data mining as a strategy for assessing asynchronous discussion forums," *Comput. Educ.*, vol. 45, no. 1, pp. 141–160, 2005, DOI: 10.1016/j.compedu.2004.05.003.

[32] R. Deng and P. Benckendorff, "What are the key themes associated with the positive learning experience in MOOCs? An empirical investigation of learners' ratings and reviews," *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, 2021, DOI: 10.1186/s41239-021-00244-3.

[33] A. Adikari, G. Gamage, D. Silva, N. Mills, S. M. J. Wong, and D. Alahakoon, "A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web," *Futur. Gener. Comput. Syst.*, vol. 116, pp. 302–315, 2021, DOI: 10.1016/j.future.2020.10.028.

[34] X. Peng, C. Han, F. Ouyang, and Z. Liu, "Topic tracking model for analyzing student-generated posts in SPOC discussion forums," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, 2020, DOI: 10.1186/s41239-020-00211-4.

[35] R. Hodhod and H. Fleenor, "A text mining based literature analysis for learning theories and computer science education," *Adv. Intell. Syst. Comput.*, vol. 639, no. September, pp. 560–568, 2018, DOI: 10.1007/978-3-319-64861-3_52.

[36] A. P. Rovai, "Facilitating online discussions effectively," *Internet High. Educ.*, vol. 10, no. 1, pp. 77–88, 2007, DOI: 10.1016/j.iheduc.2006.10.001.

[37] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," *IEEE Trans. Learn. Technol.*, vol. 7, no. 4, pp. 346–359, 2014, DOI: 10.1109/TLT.2014.2337900.

[38] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Soft computing in data science," *Int. Conf. Soft Comput. Data Sci. SCDS 2015*, pp. 129–140, 2015.

[39] H. N. Ocharo and S. Hasegawa, "Using machine learning to classify reviewer comments in research article drafts to enable students to focus on global revision," *Educ. Inf. Technol.*, vol. 23, no. 5, pp. 2093–2110, 2018, DOI: 10.1007/s10639-018-9705-7.

[40] L. Feng, H. Lu, S. Liu, G. Liu, and S. Luo, "Automatic feature learning for MOOC forum thread classification," *ACM Int. Conf. Proceeding Ser.*, pp. 65–70, 2018.

[41] L. Feng, L. Wang, S. Liu, and G. Liu, "Classification of discussion threads in MOOC forums based on deep learning," *DEStech Trans. Comput. Sci. Eng.*, pp. 493–498, 2018.

[42] A. Osman Id, N. Salim, and F. Saeed, "Quality dimensions features for identifying high-quality user replies in text forum threads using classification methods," 2019, DOI: 10.1371/journal.pone.0215516.

[43] D. Waller, K. Douglas, and G. Nanda, "A case study of discussion forums in two programming MOOCs on different platforms," *ASEE Annual Conference & Exposition*, 2020, DOI: 10.18260/1-2--31942.

[44] A. Miller, "Text mining digital humanities projects: Assessing content analysis capabilities of voyant tools," *J. Web Librariansh.*, vol. 12, no. 3, pp. 169–197, 2018, DOI: 10.1080/19322909.2018.1479673.

[45] L. J. Sampsel, "Voyant tools," *Music Ref. Serv. Q.*, vol. 21, no. 3, pp. 153–157, 2018, DOI: 10.1080/10588167.2018.1496754.

[46] Neha and E. Kim, "Designing discussion forum in SWAYAM for effective interactions among learners and supervisors," *HCI International 2020—Late Breaking Posters*, 2020, pp. 297–302.

[47] G. Hetenyi, A. Lengyel, and M. Szilasi, "Quantitative analysis of qualitative data: Using voyant tools to investigate the sales-marketing interface," *J. Ind. Eng. Manag.*, vol. 12, no. 3, pp. 393–404, 2019, DOI: 10.3926/jiem.2929.

[48] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Communications*, vol. 10, no. 1. 2019, DOI: 10.1038/s41467-019-13055-y.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).