10-2019

TopicSummary: A tool for analyzing class discussion forums using topic based summarizations

Swapna GOTTIPATI
Singapore Management University, SWAPNAG@smu.edu.sg

Venky SHANKARARAMAN
Singapore Management University, venks@smu.edu.sg

Renjini RAMESH
Carnegie Mellon University

Citation

GOTTIPATI, Swapna; SHANKARARAMAN, Venky; and RAMESH, Renjini. TopicSummary: A tool for analyzing class discussion forums using topic based summarizations. (2019). *2019 IEEE Frontiers in Education Conference 49th FIE: Covington, KY, October 16-19: Proceedings*. 1-10.
Available at: https://ink.library.smu.edu.sg/sis_research/4518

TopicSummary: A Tool for Analyzing Class Discussion Forums using Topic Based Summarizations

Swapna GOTTIPATI[1], Venky SHANKARARAMAN[1], Renjini RAMESH[2]
School of Information Systems[1],Singapore Management University, Carnegie Melon University[2]
Swapnag,venks@smu.edu.sg, renjinir@andrew.cmu.edu

Abstract— **This Innovative Practice full paper, describes the application of text mining techniques for extracting insights from a course based online discussion forum through generation of topic based summaries. Discussions, either in classroom or online provide opportunity for collaborative learning through exchange of ideas that leads to enhanced learning through active participation. Online discussions offer a number of benefits namely providing additional time to reflect and synthesize information before writing, providing a natural platform for students to voice their ideas without any one student dominating the conversation, and providing a record of the student's thoughts. An online discussion forum provides a repository comprising the discussion threads related to the topics discussed. One approach to extracting useful knowledge from the repository is through generation of concise summaries of the discussion for each topic. This summary information can help both the instructor and the student in being able to focus on the key learning points discussed in the forum threads. The focus of our research is directed towards analysis of online discussion forums (ODFs) and generating topic based summaries that can be viewed by both instructor and the students. We have developed a tool, Topic Based Summarization (TBS), that takes an excel sheet with discussion forum thread posts as input and generates visual reports of summaries that are clustered into different topics. We evaluate the tool using discussion thread posts for an undergraduate course titled "Business Process Modelling and Solutioning". We also performed qualitative analysis of the tool to investigate the strengths and weakness of various summarization algorithms.**

Keywords— Online discsusion forum, content analaysis, text mining, topic based summary, clustering

I. INTRODUCTION

Learning is a process which occurs in a social context and involves interaction between students and instructors. Effective learning process occurs when both instructors and students interact and actively participate in the learning activities. Discussions, either in classroom or online, is a type of active learning, where there is a sustained exchange between the students and the instructor and between the students, with a purpose to enhancing learning through active participation.

Class discussions as well as online discussion forums, should be carefully designed and executed by the instructors, to ensure high quality and high quantities of "student talk", that lead to better learning experience. A primary by-product of these discussions is the generation of a knowledge repository that comprises the discussion threads related to the topics discussed. This knowledge repository when analysed, can generate useful information and insights for both instructors and students. For example, the instructor will be able to gain insights on which topics had more student participation, and an individual student can obtain a summary of the discussion. Current research studies of analysing discussions mainly focus on student academic achievement correlations, discourse or conversation analysis, sentiment analysis, students' traits identification and cognitive behaviour analysis.

In this paper, our focus is in the analysis of online discussion forums (ODFs) and to generate topic based summaries that can be viewed by both instructor and the students. Compared to live classroom discussion, the use of ODFs addresses two key challenges related to student participation is discussions; communication and time. ODFs remove some of the communication impediments associated with face-to-face discussions. For students, the online environment is less intimidating, less prone to be dominated by a single participant and less bounded by convention [1]. Hence it provides an equitable forum to address issues through argumentative and collaborative discourse [2]. It also provides students the flexibility of time and place to reflect on the previous postings to the discussion thread [3] and thus actively engage them to share their experience in a more meaningful and thoughtful manner.

Figure 1 shows the sample of a discussion forum from the learning management system used in our school. The column "Forums", represents the title given to the discussion forum. An instructor can setup several discussion forums during the delivery of a course. Column "Questions" shows the thread posts. This is the initial post by the instructor to facilitate the discussion. We also refer to this data as a "question thread", a question requesting responses from the students. Each Question thread may have a title constituting a broad area of the concepts discussed in the question thread which is identified by the column "Question Title". For each question thread, students will provide the answers which are saved under the column "Body".

One of the key challenges of such online discussion forums is the voluminous information that is generated. We surveyed 16 students in higher education on challenges of

Name	Forums	Question Title	Questions	Body
Student 1	Subway process case study	Performance Target	Q1: What are the performance targets for New Subway sandwich sales process in addition to the initial targets?	1. Reducing the total time taken for each customer from ordering to payment (efficiency) 2. Increase total manpower (efficiency)
Student 2	Subway process case study	Improvements & Rationale	Q2: What are some recommendations to improve the process and state the rationales for it?	1. Inrease the number of manpower in the shop so as to speed up the process of making sandiwches because there will be more staffs to cater to more customers.2. In accordance to increase the manpower, increasing the machinery is also a must inorder to be efficient.3. Hire more part timers just to come during the peak hours (can save some cost for the company raher than letting them to work full day) to solve the peak hour problem.

Figure 1: Sample posts from discussion forum in our school's Learning Management Systems depicting various key components along with the spelling and grammar errors

extracting useful insights from the discussion forums. 87.5 % of students agreed that the summarised views of the topics discussed in a question thread can help them in learning from the discussions with other students. Topic-based summaries from threads curated by the instructor is the key focus of this paper. The topic based concise summaries provides three key benefits for learning; prepare students for assessments, instructors can analyse the topics of strength and weakness among student, and finally encourage peer learning among students.

Manually analysing such knowledge repositories and generating high quality information such as topics and summaries is a pain staking process. To the best of our knowledge there is no study on the topical based summarization of student discussion forums in the education domain. The novelty of the work is two-fold. Firstly, the task that we explore for summarization is a novel scenario where an instructor drives the discussions with a question posted in learning management systems such as MOOC discussion for forum. Based on the main topic identified in the "question thread", our research goal is to extract the sub topics from the students' discussions. Secondly, through the provision of adjustable technical parameters, the auto generated sub-topics and the concise summaries can further be improved by intervention of the instructor. The topic based summaries are then shared with the students for efficient learning.

In the traditional approach to document summarization, a sentence is usually treated as an individual unit of text and summaries are constructed by extracting most relevant sentences from a document. However, the text in discussion a thread is generated by multiple users where each post comprises a distribution of sub-topics. Therefore, traditional document summarization techniques are not suitable for our task. To solve this challenge, the problem is treated as a multi-document summarization task [4].

Clustering is a popular data mining technique used for text categorization and topic discovery from textual documents [6]. We employ clustering techniques to extract the topics from the textual posts. Since clustering is unsupervised and manual labelling of each cluster is tedious, we use the top words of the cluster to tag the summaries. At the same time, the solution also provides the instructor with the facility for adjusting the parameters to improve the quality of the clusters [5].

The main contribution of our work is the innovative application of text clustering, natural language processing, and summarization and visualization techniques in the education domain. The tool empowers the instructors with insights that can be gleamed from students participations in the discussion forums and help continually improve the student learning experience through the provision of three capabilities; (1) Web based environment for uploading and analysing discussions; (2) User friendly interface that supports the selection of clusters, and summarization techniques to view high quality topical based concise summaries; (3) Quantitative figures on the contributions of individual students towards each topic for a given discussion thread.

The rest of this paper is structured as follows: Section II describes the research problem statement and defines some of technical terms. In Section III, we review related work in two areas namely, use of discussion forums in education and the application of analytics to gain insights from discussion forums.

Section IV describes the overall solution design and the details of each stage of the solution process. In Section V, we present the details of the dataset used for the research. In section VI, we present the results of evaluation of the tool and its limitations. We present interesting future directions of this research and conclude in Section VII.

II. Problem Definition

In this section, firstly, we introduce the key components and terms related to discussion forums [7]. Secondly, we formally define the topic based summarization task and the challenges associated with it.

A. Components of Student Discussion Forum

Based on the Learning Management System used in our university, the discussion forum comprises of three components namely, forum, thread and post.

1. Forum: Online discussion forum (ODF) is a web-based application that brings people together with shared interest and mind-set. It has a tree like structure. Top nodes are sub-forums and sub-nodes are the threads in the sub forums. In a class based ODF the instructor and the students automatically enrolled. It provides features for the instructor to create threads and collect the responses from the students, usually in a HTML or excel format. In Figure 1, "Subway process case study" is the discussion forum.

2. Thread: In a discussion forum, the messages posted by different students participating in the forum are visually grouped with their replies. This grouping is referred to as a thread. They are the placeholders under which the students can post their discussions related to a key topic. The instructor can structure the discussions to align it with the content covered during the classroom lectures, by initiating the discussion through a question. We refer to this as a "question thread". In Figure 1, values under the column "Questions" are referred to as question threads. Additionally, the instructor can further motivate the students by awarding discussion participation marks.

3. Posts: Posts are the messages posted by the instructor or students, which can be in text or image or video formats. Usually, the instructor posts the first question, followed by the students' posts which are responses to the question. Sometimes, the instructor can also give feedback to a specific student post. A question posted by the instructor is labelled as "Questions". The answer posted by the students are labelled "Body". Figure 1 shows the student posts under the column, "Body".

The posts under each thread discuss various topics for the given thread. In our paper, we focus on extracting topics from the posts and generating concise summaries.

B. Topic Based Summarization and Challenges

1. Topic: A topic or theme is a key idea or subject or matter dealt with in an article or text or document or discussion. Each topic or theme is identified through a representative set of words. For example, Table 1 shows the themes and topics for a case study discussion used in the course, "Business Process Modelling" along with the corresponding representative words that identify the topic or theme.

Table 1: Example theme and representative words for the topic.

Themes	Topic Representative words
Process Cycle Time	Time, waiting, reduce, resource, cycle
Payment	System, customer, online, kiosk, payment

2. Summary: A summary can be defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and usually significantly shorter than combined length of the texts [4].
3. Topic-based summary: It can be defined as the concise summaries generated from posts that are clustered or grouped under a single topic. Such summaries provide the details of the concepts and some examples as well.

In our research, we define the Topic Based Summarization task as follows:

"The ability to automatically cluster the discussions into unique sets of summaries that correspond to a specific topic or theme".

This task poses two types of challenges; input data challenges and text mining challenges. The input data challenges include spelling errors, grammar syntax errors in the posts as shown in Figure 1. The text mining challenges include appropriately labelling the clusters and ensuring acceptable quality levels for the generated summary.

The first data challenge, spelling errors, is handled using NLP techniques [8]. NLP tools provide APIs to autocorrect the spell errors by replacing the error with the closest possible correctly spelled word. The second data challenge, grammar syntax errors, is handled by using use tokenised words and stopword removal technique when generating clusters of posts, thus not majorly effecting the quality of the clusters.

The challenge of labelling the clusters is handled by using few top topic words as labels for a cluster. Every cluster is named using the top 5-6 frequent words that appear in the posts that belong to that cluster.

III. Related Work

We review related work in two areas namely use of discussion forums in education and the application of analytics to gain insights from discussion forums.

A. Discussion Forums in Education:

The key advantage provided by online discussion forums (ODF) is the asynchronous interactions. In other words, the ability to communicate with peers and instructors independent of time and space [9]. There are two reasons for wider adoption of ODFs by instructors in tertiary education. Firstly, the advancements and easy access to discussion forum technology. For example, most universities use Learning Management Systems (LMS) which natively support discussion forums. Secondly, the characteristics of millennial and Gen Z students, who have a greater dependence on technology, and their desire to embrace online social learning environments. They expect on-demand services that are available at any time and with low barriers to access. Hence, making ODFs a good choice for this group of students.

Students can use the discussion forum to discuss key concepts, enabling them to share ideas as well as learn within the group [10]. This helps the student in becoming a part of a vibrant learning community, rather than being an independent learner who completes and submits assignments without any peer interaction [11]. When effectively used, discussion forums can help in encouraging student leadership, giving them more voice in the class [12]. They build classroom dynamics by promoting discussion on different course topics. They allow students to reflect deeply on course concepts. Students have more time to research, reflect, and compose their thoughts prior to participating in discussions [13]. Moreover, meeting course objectives and aligning course content are other purposes of discussion boards [14]. It is important to manage participants' interaction time and ensure that forum interactions are relevant and enriching [15].

B. *Analytics on the Discussion Forums*

Analysing quantity and quality of online postings and comparing students' performance provides insights to the instructors on the effectiveness of ODF in learning process. Ravi Seethamraju conducted quantitative analysis on the aspects such as timing of responses, number of posts for various questions, etc. This research also focused on manually performing content analysis on a number of aspects. For example, evidence that the student read and understood others' ideas and contributions; evidence of good analysis of the case study data; demonstrable understanding of the questions, and identification of issues in case study. When compared to the previous cohort that did not use the discussion forum, this study observes a significant improvement in student learning through the effective use of discussion forum [16].

At times, instructor tend to believe things are going well when they are not, or conversely think the class is not understanding things and is not progressing when in fact they are. Therefore, instructors might need to know how the class is doing to make timely interventions and motivate the students. Schubert et al., proposed text analytics based approaches for assessing the sentiment of a large population of learners, through the learner generated discussion forum posts and without the benefit of face to face interaction [17]. We also use similar text mining approaches for our project. However, our research is not focused on sentiment analysis but on knowledge extraction by performing content analysis on the discussion forum posts.

Content analysis is a key area of research that enables to perform analysis on textual data. The input to content analysis can include all sorts of recorded communication such as transcripts of interviews, discourses, protocols of observations, video tapes, documents, discussion forums, etc. [18].

In the context of analysing the content of student discussion forums, it can be further sub-divided into a number of sub-tasks such as interaction analysis, learning pattern analysis, and behaviour pattern analysis. Understanding students' online interaction is important because interaction influences the quality of online learning [19]. Interactions among students in online classes can further motivate them to learn through engagement with other peers [20]. Hence, discovering students' evolving interaction patterns and identifying different types of interaction patterns among students in the same class can provide useful insights in discovering issues related to the learning process [13]. Our paper focuses on the content analysis on the discussion posts submitted by the students and our goal is to discover knowledge that can further enhance the learning process. We adopt text mining approaches to perform content analysis on the qualitative data and develop a solution for extracting knowledge insights in the form of a summary, from the discussions.

IV. Solution design

In this section we present the solution design of our proposed system.

A. *System Overview*

Figure 2 shows the three stages in the Topic-Based Summarization (TBS) tool namely Data Processing, Topic Extraction and Summary Generation. The data from the discussion forum obtained from the learning management system is the input to the tool. The outputs are the topic-based summaries for each thread.

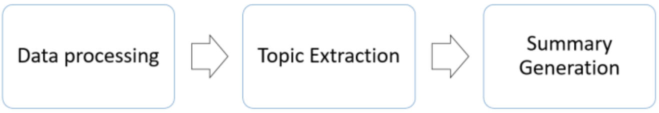


Figure 2: Overview of solution approach based on text mining techniques.

B. *Data Processing*

During the data processing stage, the discussion posts are converted into lower case, and the trailing and ending spaces are removed. Each post is tokenized into sentences, so that in the later stage it helps with categorizing the sentences into groups.

Stopwords such as prepositions, determiners, to-be verbs etc., can create "noise" which can affect the performance of the text mining algorithm. Hence stop words are removed using the NLTK stop words English list [21] and additional stop words, which are commonly used in academic discussion forums, are added by creating a custom list.

Lemmatisation is the process of grouping together the inflected forms of a word so they can be analysed as a single item, which are identified by using word's lemma, or dictionary form [8]. Lemmatization of words is carried out using NLTK's WordNet based Lemmatizer API [21]. This process helps with effective clustering, and to generate for each cluster, the representative words, which are accurate, and non-repetitive. Hyperlinks, special characters and numbers are removed from all sentences as they generally do not provide added value to the word corpus and tend to distort tokenization and clustering results. Hence, some information may be lost due to this, but at the cost of attaining better quality of clustering.

C. Topic Extraction

Research work in Topic Detection and Tracking (TDK) aims to identify stories in several continuous news streams that pertain to new or previously unidentified events [22]. We apply some of the techniques used in (TDK), where the main task is to cluster a group of news items, blogs or tweets and then discover the labels of these clusters based on the content of text within the particular cluster. These cluster labels are actually the topics extracted from a group of news items, blogs, or tweets [22, 23].

In this solution design we use k-mean clustering algorithm and we have adopted the tool CLUTO [5] for implementing this algorithm. The algorithm treats each document as a vector in a high-dimensional space, and it computes the clustering distances between the documents to find the groups. In our solution, we tokenize the posts into sentences first and then create vectors for each sentence. Each sentence is the input document for the clustering algorithm.

Several algorithm choices are provided by CLUTO for clustering: I2 criterion, I1 criterion, E1 criterion, G1 criterion, H1 criterion, H2 criterion [5]. For instructors who are not technically inclined, the tool will select a default algorithm that will be used. Instructors who are not technically inclined, can analyse the results from each of the algorithms and then use the most suitable one. The number of clusters required can be set by the user. Recall that as no automated labelling is generated by the clustering tool, we use the top descriptive words for each cluster as representatives of a label to the cluster.

D. Topic-Based Summary Generation

Content reduction is a process of sentence elimination through sentence extraction. Most sentence extraction algorithms work in a constructive way: given a document and a sentence scoring mechanism, the algorithm ranks sentences by score, and then chooses sentences from the ranked list until a compression rate is reached [4].

For our solution design, we adopt multi-document summarization. Each post is first tokenized into sentences. Each sentence is considered as a document. It the process of producing a single summary of a set of related source documents. We propose two approaches for the summarization; TextRank Summarizer [24] and LSA Summarizer [25].

TextRank summarizer is an unsupervised algorithm. It does not need any training or external knowledge. Algorithm is a graph -based model which takes into consideration local vertex-specific information as well as full graph global statistics repeatedly for determining significance of vertex. In the context of summarization, sentences are considered as vertices and similarity between sentences is used to obtain a weighted graph. The ranking algorithm is run on this graph and top sentences with higher scores are selected to generate the summary of the documents [24].

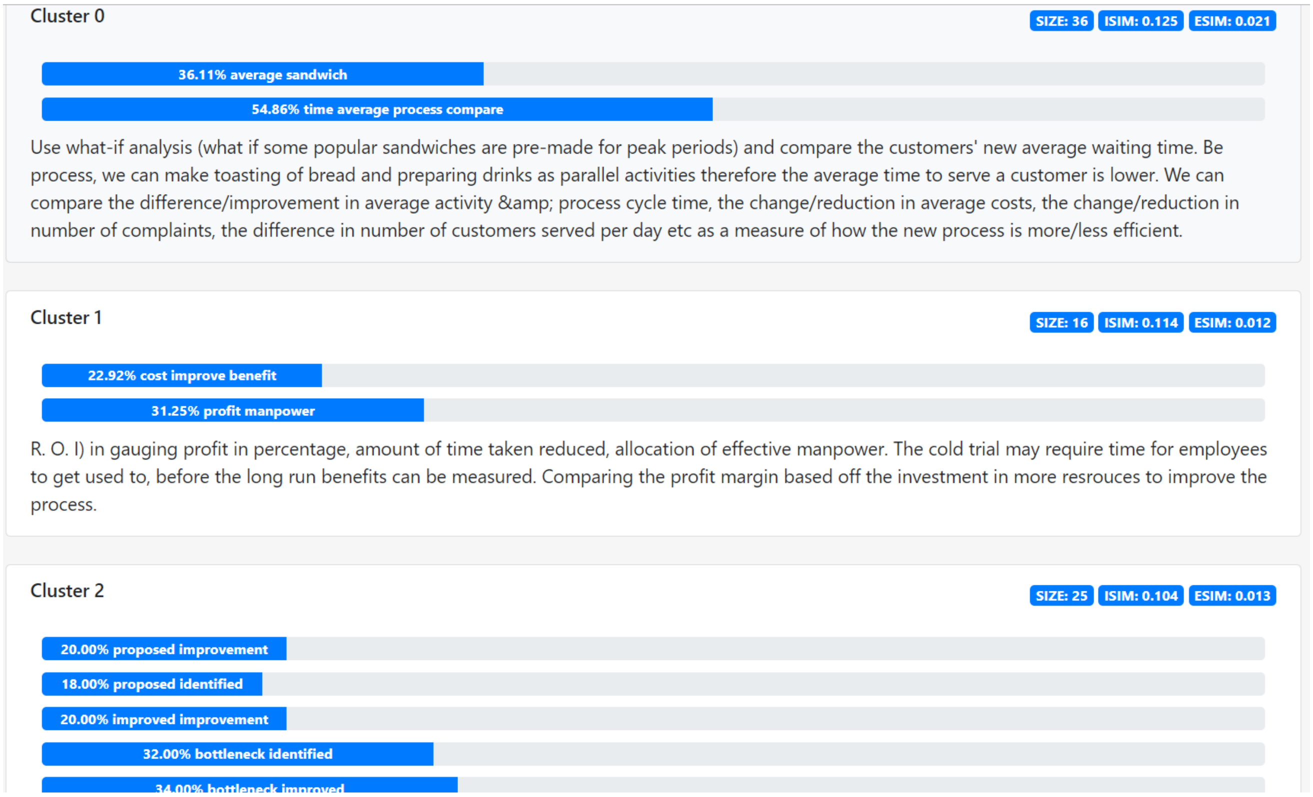
LSA summariser is based on algebraic-statistical method, Latent Semantic Analysis (LSA). Similar to TextRank, LSA algorithm is an unsupervised approach that extracts hidden semantic structures of words and sentences. LSA uses context of the input document and extracts information such as which words are co-occurring in and which common words are seen in different sentences. High number of common words among sentences indicates that the sentences are semantically related. Meaning of a sentence is decided using the word it contains, and meaning of words are decided using the sentences that contains the word. Sentences are scored based on its relevance to the concepts of the documents. Top ranked sentences from each concept are selected for the summary [25].

V. DATASET

For developing and evaluating the Topic-Based Summarization (TBS) tool, we collected the data from the discussion forum from an undergraduate course, “Business Process Modelling and Solutioning”, a second year undergraduate course within the BSc (Information Systems) degree program. One of the main learning outcomes of this course is to ensure students can perform an analysis of a given business process, identify the bottleneck and propose an improved process through use effective use of technology. The data is derived from the LMS which has an in-built online discussion forum, where the instructor and students can engage in discussions pertaining to the course topics. The students were given a case study of the sales process currently implemented in a sandwich shop (e.g. Subway). Students were required to read this case study before participating in the online discussion forum. The discussion forum was setup in the LMS and relevant questions were added by the instructors. The students subsequently submitted their posts individually for each question. The LMS technical team extracted the discussion posts as a excel spreadsheet. Table 2 shows the statistics of the posts for each question. We used this excel spreadsheet as an input to TBS. The details of the user interface and tool evaluations will be described in the next section.

Table 2: The thread questions we use in our evaluation and the corresponding posts

Thread Question	**# of posts**
Q1: What types of analysis can be done on the current process? Describe with examples	42
Q2: What are some recommendations to improve the process and state the rationale for it?	89
Q3: What are the performance targets for the new subway sandwich sales process?	129



Cluster 0

SIZE: 36 ISIM: 0.125 ESIM: 0.021

36.11% average sandwich

54.86% time average process compare

Use what-if analysis (what if some popular sandwiches are pre-made for peak periods) and compare the customers' new average waiting time. Be process, we can make toasting of bread and preparing drinks as parallel activities therefore the average time to serve a customer is lower. We can compare the difference/improvement in average activity &amp; process cycle time, the change/reduction in average costs, the change/reduction in number of complaints, the difference in number of customers served per day etc as a measure of how the new process is more/less efficient.

Cluster 1

SIZE: 16 ISIM: 0.114 ESIM: 0.012

22.92% cost improve benefit

31.25% profit manpower

R. O. I) in gauging profit in percentage, amount of time taken reduced, allocation of effective manpower. The cold trial may require time for employees to get used to, before the long run benefits can be measured. Comparing the profit margin based off the investment in more resrouces to improve the process.

Cluster 2

SIZE: 25 ISIM: 0.104 ESIM: 0.013

20.00% proposed improvement

18.00% proposed identified

20.00% improved improvement

32.00% bottleneck identified

34.00% bottleneck improved

Figure 3: Visual dashboard depicting the topics and summaries for each question thread. Each cluster represents a cohesive topic in the student's posts. Top left blue bars shows the representative words for each cluster. Top right blue bars represent the cluster statistics. The text below is the concise summary for each cluster or topic.

To understand and analyse the outputs of the TBS tool, we shall first give a quick background of the case study. Business process is a value chain in an organization. The goal of business process management team is to analyse the business process and provide recommendations to optimize the as-is business process. The improvements are measured by performance targets and the recommendations to optimise the process are usually about process changes or technology introductions to the process. Depending on the activities in the business process, the team provides specific recommendations which are practical to implement. The discussion questions are linked to the analysis and improvement of "Subway Sales Process". The goal of the tool is to extract the topic-based summaries for each question in the thread. In the next section, we describe the TBS tool and the findings.

VI. Tool Description & Evaluations

A. *Visual Dashboard and User Interface*

Figure 3 shows the visual dashboard for viewing the topics and topic-based summaries for a given question thread. Cluster 0, Cluster 1 etc., represent the grouping of the posts by topics. Recall that each post may have multiple topics and hence we used tokenised sentences for clustering. The numbers on the top right of each cluster indicate the statistics of the cluster:

1. Size: Number of mentions of a topic in the cluster
2. ISIM: Displays the average similarity between the objects of each cluster (i.e., the internal similarity). High ISIM refers to high quality cluster.
3. ESIM: Displays the average similarity of the objects of each cluster and the rest of the objects (i.e., external similarities). Low ESIM depicts high quality of the clusters.

The users of the TBS tool can analyse these statistics and appropriately adjust the number of clusters and the quality of the topics.

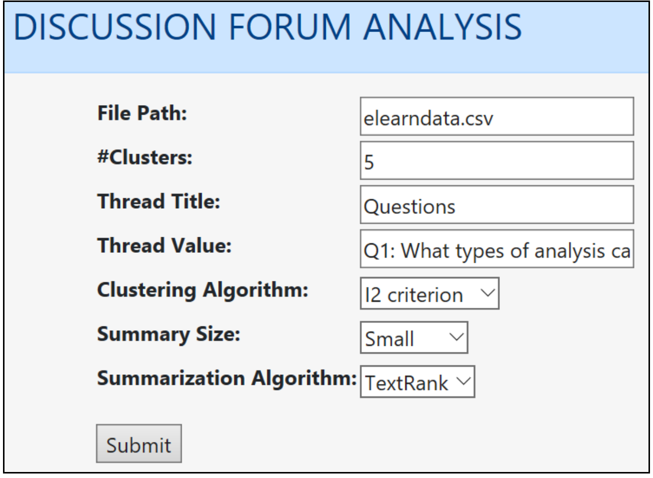
The top left corner depicts the label for the cluster (e.g. Cluster 1). In addition each cluster includes the following below the label

1. Words: The most descriptive set of words for the cluster, that is, the high frequency words from all the posts in this cluster. These are also referred to as the representative words for the cluster (e.g. cost, improve, benefit for Cluster 1).
2. Percentage: Right next to each word or group of words, the tool displays a number which is the percentage of the intra-cluster similarity that this particular word can explain [5]. For example, for the Cluster 1, the feature "cost" explains 22.92% of the

average similarity between the objects of the Cluster 1.

The paragraph at the bottom of each cluster depicts the summary derived from all the sentences within this cluster. The short summary is derived using TextRank technique.

Figure 4 shows the user interface for the instructor to upload the discussion forum data as an excel spreadsheet and specify the options for clustering and summarization. The details of each input field along with an example is explained in Table 3. If the instructor is unsure of which clustering technique or the summarization technique to use, the default technique will be selected by the TBS tool for processing the discussion forum data and generating the topic-based summaries for each question thread.



DISCUSSION FORUM ANALYSIS

File Path: elearndata.csv

#Clusters: 5

Thread Title: Questions

Thread Value: Q1: What types of analysis ca

Clustering Algorithm: I2 criterion

Summary Size: Small

Summarization Algorithm: TextRank

Figure 4: User interface to input the discussion forum data.

Table 3: Description of input fields in the user interface

Field	Description
File Path	Location and name of the spreadsheet file containing the discussions.
# Clusters	Number of clusters required
Thread Title	The column header that we wish to classify on. This is same as the question thread. In this example, the column title is “Questions”. All the questions are under this column.
Thread Value	The column “Questions” will have many questions for discussion within the given discussion forum (e.g. Q1, Q2). In the example shown in Figure 3 the question thread is for the first question, “Q1: What types of analysis can be done on the current process? Describe with examples”
Clustering Algorithm	Choices provided are I2 criterion, I1 criterion, E1 criterion, G1 criterion, H1 criterion, H2 criterion. User can experiment with different algorithms to evaluate best outcome for the given dataset. Default is “I2 criterion” [REF].
Summary Size	Choices provided are Small, Medium and Large. Summary size, i.e. number of sentences extracted in summary are adjusted accordingly. Default size is “Small”.
Summarization Algorithm	Choices provided are TextRank and LSA. Default algorithm is “TextRank”.

B. Evaluations

1) Topics Evaluatons

In this section, we show the clustering results for each question. Recall that the cluster labels are the top representative words which are also frequent words in the sentences within the cluster. Table 4 shows the top representative words for each cluster of the thread questions.

Cluster top representative words: resource time reduce cycle waiting, Size: 74, ISim: 0.061, ESim: 0.014

Automating manual task will allow an overall reduction in average process cycle time, leading to higher capacity of production. During such periods of **congestion** and heavy footfall, the **manager might be required to take on a more "directive"** or managerial position to ensure processes are followed, rather than be the hands-and-feet in the activities. Better allocation of resources: may be can **appoint one person in charge** of completing meals and place bread in toaster to save overall time duration.

(a) TextRank Summarization

Invest in better and more efficient toasters (shorter toasting time and to enable more sandwiches to be toasted at one time) -- this can help reduce the preparation time for each sandwich and thus improve efficiency. reduce time to process each order by **expanding employee job scope** (letting them handle more than role)reduce customer wait time through pre-orders/ online ordering effective use of resources through **reallocating staff** from the affected branches to the other branches to **cope with peak periods**. **Adding resources**(e.g hire another cashier and another machine) would help to reduce the process cycle time as when there is cashier, the process cycle time will be reduce by half.

(b) LSA Summarization

Figure 5: Summarization comparison for Q2, “What are some recommendations to improve the process and state the rationales for it?” The comparison is on the same cluster, “Resources for reducing cycle time”. The comparison is on small size summary. Highlighted words shows the organizing of ideas in this cluster by the summarization algorithms.

Table 4: Topics generated by tool for each question thread

Question	Cluster	Topic: Top representative words
Q1	1	utilisation understand resource utilization
	2	process analysis current cycle time
	3	construction customer entire process wait
	4	bottleneck peak complain time waiting
	5	day produced sandwich prepare sold
	6	solution manpower shortage due branch
	7	manager analysis activity path analysis determine
Q2	1	help role sandwich bottleneck cashier
	2	resource time reduce cycle waiting
	3	sandwich customer bread prepare drink
	4	hour branch subway peak manpower
	5	system customer online payment kiosk
Q3	1	waiting time peak hour queue
	2	target include initial addition performance
	3	manual task reduce manpower efficiency
	4	cost reduce process sale sandwich sale
	5	average time total day customer

From Table 4, we observe that each question has different number of clusters. Recall that the instructor can use the ISIM, ESIM parameters and manual qualitative analysis to adjust the number of clusters that are required by defining it through the user interface in order to generate high quality clusters. In our evaluations of TBS tool, we took this approach to select the optimum number of clusters for each question, which explains the reason for the difference in the number of clusters. We also observe the top representative words for each cluster are coherent and align well with the question.

2) Summarization Evaluations

Figure 5 shows the comparison of summaries generated by both the algorithms, TextRank and LSA. The summaries are generated for the Q2: “What are some recommendations to improve the process and state the rationale for it?” The summary comparison is on the second topic, “resources for reducing cycle time”

From TextRank summary we observe the key recommendations such as peak time, hiring, and multi-tasking of the managers are extracted from the discussion posts. On the other hand, the LSA summary extracts additional an recommendation, buying more toasters. This shows LSA summarization is slightly better than TextRank for this specific discussion forum post. However, the TBS tool provides the instructor with the choice of using both the algorithms and the instructor can choose the summary of one algorithm or combine the summaries before sharing with the students.

C. Limitations and Future Work

We identify a number of limitations of the TBS tool that will be addressed in the future work of this research. Firstly, a key limitation of the TBS tool is its performance with regard to spell check, this requires further investigation and selection of better techniques for doing the spell check. Secondly, the current approach of manual analysis of ESIM and ISIM to determine the optimum number of clusters can be a tedious and time consuming process. Going forward we will also investigate in improving the process of choosing optimum cluster number. For example, giving some recommendations to the instructor based on preliminary analysis of the dataset. Thirdly, we intend to expand the tool to include a feature to generate pdf file of the summary which the instructor can share with the students. Finally, we will be conducting a survey involving the instructors and students on the effectiveness of the current TBS tool in enhancing the learning process and identifying new features that can be useful for them.

VII. Conclusions

In this paper, we presented a text mining based approach to analyse the discussion forums and generate topic based summaries. The TBS tool uses clustering techniques to generate the topics from the posts submitted by the students for the given question thread and summarization technique is used generate topic-based concise summaries for each topic. We evaluated the tool on the discussion forum created for an undergraduate Information Systems course and the qualitative evaluations show the effectiveness of the tool in extracting both topics and topic-based summaries.

References

[1] Redmon, R., & Burger, M. (2004). Web CT discussion forums: Asynchronous group reflection of the student teaching experience'. Curriculum and Teaching Dialogue, 6(2), 157–166.

[2] Karacapilidis, N., & Papadias, D. (2001). Computer supported argumentation and collaborative decision making: The HERMES system. Information Systems, 26(4), 259-277.

[3] Anderson, T. & Kanuka, H. (1997). On-line forums: New platforms for professional development and group collaboration. Journal of Computer-Mediated Communication, 3(3).

[4] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. Comput. Linguist. 28, 4 (December 2002),

[5] Karypis, G. 2002. {CLUTO} a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota.

[6] Leouski, A. and Croft, W. 1996. An evaluation of techniques for clustering search results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst.

[7] Biriyai, Alabo and Emmah, Victor, Online Discussion Forum: A Tool for Effective Student-Teacher Interaction (November 16, 2014). Available at SSRN: https://ssrn.com/abstract=2525047 or http://dx.doi.org/10.2139/ssrn.2525047

[8] Ellen M. Voorhees. 1999. Natural Language Processing and Information Retrieval. In Information Extraction: Towards Scalable, Adaptable Systems, Maria Teresa Pazienza (Ed.). Springer-Verlag, London, UK, UK, 32-48

[9] Blackmon, S. (2012). Outcomes of chat and discussion board use in online learning: A research synthesis. Journal of Educators Online, 9(2), 2.

[10] Balaji, M. & Chakrabarti, D. (2010). Student interactions in online discussion forum: Empirical research from media richness theory perspective. Journal of Interactive Online Learning, 9(1), 1-22.

[11] Harris, N. & Sandor, M. (2007). Developing online discussion forums as student centred peer e-learning environments. In: ICT: Providing Choices for Learners and Learning. Proceedings Ascilite Singapore 2007, pp. 383-387.

[12] Dringus, L. & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. Computers & Education, 45(1), 141-160.

[13] Song, L. & McNary, S.W. (2011). Understanding Students' Online Interaction: Analysis of Discussion Board Postings. Journal of Interactive Online Learning, 10(1), 1-14. Retrieved March 14, 2019 from https://www.learntechlib.org/p/109405/.

[14] Streveler RA, Smith KA, Pilotte M (2012). Aligning course content, assessment, and delivery: creating a context for outcome-based education. In: Outcome-Based Education and Engineering Curriculum: Evaluation, Assessment and Accreditation, ed. K Mohd Yusof,S Mohammad, N Ahmad Azli, M Noor Hassan, A Kosnin, and SK Syed Yusof, Hershey, PA: IGI Global.

[15] Biggs, J. (2012). What the student does: Teaching for enhanced learning. Higher Education Research & Development, 31(1), 39-55.

[16] Seethamraju, R. (2014). Effectiveness of using online discussion forum for case study analysis. Education Research International, 2014, 1-10.

[17] Schubert, M., Durruty, D., & Joyner, D. A. (2018). Measuring Learner Tone and Sentiment at Scale via Text Analysis of Forum Posts. In Proceedings of the 8th Edition of the International Workshop on Personalization Approaches in Learning Environments (PALE). London, United Kingdom.

[18] Krippendorff, K. (2004). Content analysis. An introduction to its methodology (2nd ed.). Thousand Oaks: Sage. (Orig. 1980).

[19] Trentin, G. (2000). The quality-interactivity relationship in distance education. Educational Technology, 40(1), 17-27.

[20] Gabriel, M. A. (2004). Learning together: Exploring group interactions online. Journal of Distance Education, 19(1), 54-72.

[21] Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117

[22] Y. Choi, Y. Jung, and S. Myaeng," Identifying Controversial Issues and Their Sub-topics in News Articles" Book Title: "Intelligence and Security Informatics" Book Series Title:" Lecture Notes in Computer Science" , 2010,Chen,H. et al. (Eds.) Springer Berlin / Heidelberg pp. 140-153

[23] X. Dai, Q. Chen, X. Wang, and J. Xu , "Online topic detection and tracking of financial news based on hierarchical clustering," Machine Learning and Cybernetics (ICMLC), 2010 International Conference on , vol.6, no., pp.3341-3346, 11-14 July 20

[24] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts", In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP),

[25] Makbule G. Ozsoy, Ferda N. Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. Journal of Information Science, 37(4):405–417.