Mining Informal & Short Student Self-Reflections for Detecting Challenging Topics – A Learning Outcomes Insight Dashboard

Ong De Lin, Swapna Gottipati, Lo Siaw Ling, Venky Shankararaman
School of Information Systems
Singapore Management University
deling.ong.2017,swapnag,venks,sllo@smu.edu.sg

***Abstract*—: Having students write short self-reflections at the end of each weekly session enables them to reflect on what they have learnt in the session and topics they find challenging. Analysing these self-reflections provides instructors with insights on how to address the missing conceptions and misconceptions of the students and appropriately plan and deliver the next session. Currently, manual methods adopted to analyse these student reflections are time consuming and tedious. This paper proposes a solution model that uses content mining and NLP techniques to automate the analysis of short self-reflections. We evaluate the solution model by studying its implementation in an undergraduate Information Systems course through a comparison of three different content mining techniques namely LDA–bigrams, GSDMM-bigrams, and Word2Vec based Clustering models. The evaluation involves both qualitative and quantitative methods. The results show that the proposed techniques are useful in discovering insights from the self-reflections, though the performance varied across the three methods. We provide insights into comparisons of the perspectives, which are useful to instructors.**

***Keywords*—** *Informal self-reflections, text mining, content analysis, GSDMM, LDA, Word2Vec, K-Means*

I. INTRODUCTION

Self-reflections have become an essential component in higher education that provides the opportunity to help students develop confidence in their learning as they establish learning goals and eventually take ownership of their learning [1], [2]. Analysing these self-reflections can further help instructors identify students who are struggling with their learning and support them in narrowing the learning gap.

Reflective thinking helps find solutions to problems in situations that are highly undetermined [3]. It aids to investigate what occurred, critically analyse experiences to make informed decisions, and test their validity. Given that self-reflection places emphasis on personal experiences, learning has become more personalised. Therefore, it is an act of active learning compared with the passive absorption of lesson content. Reflective writing is used as an educational tool in many disciplines. The most traditional self-reflection formats include journals, directed writings, discussions, and portfolios whereas, the new formats include videos, weblogs and audios using reflection annotation tools. Instructors may engage the students in self-reflection activities in classrooms, but it is not often habitual. With all the challenges faced in teaching the content and assessing learning outcomes, it is very common for the instructor to skip the reflection activity, thus losing an opportunity to allow students to take responsibility for their learning. This can be achieved through the use of mid-term or end of class session short reflections using directed questions [4]. One approach to build reflective practice into a class session is by encouraging students to record their thoughts about what they have learnt. In this paper, we share our experiences with collecting and analysing such self-reflections at the end of each session, that are informal and brief in nature.

To gain useful insights from student reflections, qualitative and quantitative content analysis methods are adopted by the instructors. The type of content analysis used is dependent on the format of reflection writing (e.g., audio or text) and the purpose of analysis [2], [5]. Textual content is usually analysed with text analysis methods or language aspects methods which help to detect insights about the students' learning journey. Videos and audios require more advanced techniques [6]. The most common purpose of content analysis involves the discovery of insights and organizing them into certain categories that indicate the students' status. The frequently used categories include a description of an experience, awareness of feelings, awareness of one's perspective, having a critical stance, considering other's perspectives, and the description of learning outcomes.

Course Learning Outcomes (CLOs) are measurable, observable, and specific statements that clearly indicate what a student should know and be able to do because of learning [7]. The normal practice is to decompose these course-level learning outcomes into a unit or session-level outcomes. For example, "on completing this session, the student must be able to draw the business process workflow diagram". One particular use of the self-reflections is that by analysing and identifying the challenging topics in the session, the instructor can gain insights on the challenges faced by students in acquiring the session-level learning outcomes. With this insight, the instructor can take remedial action or redesign the session-level activities to enhance the acquisition of learning outcomes.

While analysis of self-reflections provides important insights about student's learning experience, it is a time-consuming manual process. In most cases, it involves the use of manual qualitative content analysis on a student's reflection

write-up. With advancements in the field of text mining, there have been some attempts to use more automated methods. Machine learning approaches have been employed by several researchers to analyse student reflections such as topic models and classifiers [8], [9]. Many other related works have applied machine learning for the binary classification of reflective statements [10]. Most of the previous works focused on long or medium size articles that are mostly structured and grammatical in nature. In this paper, our approach differs from earlier work in terms of the specific insights we seek to obtain namely topics of the course that the students are reflecting upon and mapping of learning outcomes to those topics. Additionally, in our study, we focus on informal, short self-reflections at the end of every weekly session of a course that are mostly ill-structured and may not follow grammatical rules.

To the best of our knowledge, there is no previously published work analysing the informal and short end-of-class-session self-reflections using content mining approaches. In this paper, we propose a probabilistic topic modelling method that use Dirichlet Mixture Model (DMM) [11]. This approach has been successfully implemented on short texts such as tweets. The DMM approach, using unsupervised learning, extracts the latent topics from the informal self-reflections which are short in length. We apply Gibbs sampling approach for DMM and GSDMM [12]. We then compare with baseline LDA model [27].

Word2Vec [13] is a two-layer neural network machine learning model that produces word embedding. The embedding is essentially the neural network's internal representation for the word after taking in a corpus and training such that the representation of a word depends upon its neighbouring words. As a result, each word is represented as a vector and semantically similar words have similar vectors. Each self-reflection can make use of the word vectors to form a sentence representation, which can be clustered to capture latent semantic structure or topics in the corpus. Two clustering methods are covered in this study, K-Means, and Agglomerative clustering. Both are well-known approaches and have been used to analyse text. K-Means is used as the clustering method on six benchmarking text datasets and achieved consistently better results [14] while Agglomerative clustering has been shown to achieve a much better quality compared to other clustering models (such as DBSCAN) used for topic detection from news items [15].

This paper is structured as follows. Section II presents a review of related work in the areas of self-reflection and the application of machine learning to automate analysis of self-reflections. Section III describes the research statement and the methodology used for collecting the reflection data. In Section IV, the text analytics-based solution model is presented. In section V, we present the qualitative and quantitative evaluation of the solution and discuss the results and highlight the limitations of our work. In Section VI, we conclude with a summary and present future work.

II. Related Work

A. Characteristics of Self-Reflection

Self-reflection is a key activity and a type of thinking that leads to better learning [16]. Dewey defined that, "reflective thinking is an active, persistent, and careful consideration of a belief or supposed form of knowledge, of the grounds that support that knowledge, and the further conclusions to which that knowledge leads". Based on his theory, instructors and researchers proposed two modes, various formats, and various purposes for student self-reflections. Modes of reflection are either individual or team based. In individual mode, the students reflect on one's own experiences. In team-based mode, throughout a course, students in teams reflect together on their teamwork experiences [17]. The formats of self-reflections include journals, portfolios, essays, discussions, structured survey questions [4], videos, audios, weblogs, etc. [18].

Reflections are adopted in the class for various purposes. According to Noels et al. [2], goal-setting and proactive use of strategies are logical outcomes of the reflective process. In many cases, reflections enable shifting some of the learning responsibilities from the instructor to the learner. According to Marefat [5], reflections provide insights into students' learning process and to get closer to learner needs. Park [1] indicated that learning journals have the potential to assist the introspective examination of the students' learning behaviour process. According to Amulya [19], reflections provides a view of the course characteristics from the students' angle and serve the educators to develop efficient pedagogical practices.

Reflections can be informal or formal. Formal reflections include reflective essays which are like normal academic essays, where the student is required to explain what he or she has learnt and why it is important. A formal reflection is usually backed by reference evidence. An informal reflection can include end of course learning journals, detailed portfolios, or short reflections at end of a weekly class session. In informal reflections, students are expected to simply respond to series of structured questions and there is no requirement to formally reference the work. In our research, we focus on informal short reflections at end of a weekly session, we use the term informal-short-reflections [4], [20]. The main purpose is to get closer to students' needs by gaining insights into the topics and concepts that the students find challenging during each class session.

B. Machine Learning in Self-reflections

Analysis of self-reflections is conducted in both qualitative and quantitative modes. Reflective thinking frameworks [21], [22] provide a structured process to guide the act of reflection. When reflections are structured in this way, instructors can analyse these reflections more efficiently. Machine learning and Natural language processing (NLP) are becoming popular techniques that are used in automating the analysis of self-reflections. Ullmann et al. [23] developed a rule-based system for reflection analysis in students' blog postings using NLP techniques. Gibson et al. [24] used part-of-speech (POS) tagging to analyse student writing.

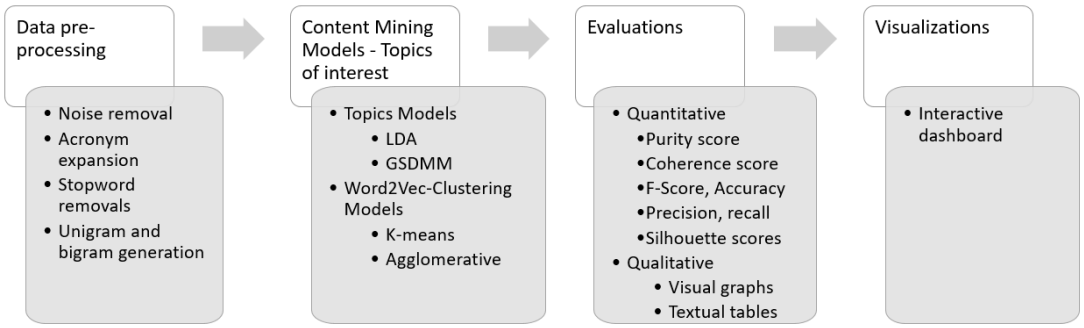

Fig. 1. Solution Model for analysing informal & short reflection

Latent Semantic Analysis (LSA) model was used by Gary [6] to understand reflections in an English language course. Kovanović et al. [25] combined various techniques such as, POS-tagging, name-entity tagging, and syntactic parsing to build a classification system for examining students' levels of critical thinking. Kovanović developed a classifier with an elaborated feature engineering process to categorise reflections into various pre-defined classes. Chen [8] applied LDA models to analyse the topics in the students' reflections. In our paper, we combine NLP techniques with topic mining models to analyse reflections and gain insights on challenging course topics or concepts. We apply various topic mining methods (LDA, GSDMM, Word2Vec-K-Means, Word2Vec-Agglomerative clustering) in our solution design to extract course topics or concepts to understand students' needs in terms of learning outcomes.

III. Research Statement and Methodology

A. Research statement

Using informal-short-reflections captured at end of each weekly session, this paper aims to answer two research questions.

RQ1: Can machine learning algorithms identify the course topics that students find challenging from their informal-short-reflections?

RQ2: How can the insights from self-reflections guide the instructor regarding the acquisition of course and session learning outcomes?

B. Methodology

We first describe the course details and the informal-short-reflections settings for each weekly session.

1) Course details

For our study, we choose a second year undergraduate Information Systems course where one of the authors of the paper is the course coordinator. Three instructors delivered this course to 367 students split into nine sections with each section having around 40 students. The topics covered in this course are related to business process management, which includes process modelling, process analysis, recommendations for improvements, designing IT solution architecture, and integrating digital technology into the business process.

2) Settings for Capturing the Reflection

We designed a weekly, after lesson reflections activity using Google forms. The students are required to complete the reflections at the end of the class or before attending the next class. The students' submission statistics are automatically generated by Google forms. The reflections are collected for nine sessions across the entire term. Table 1 shows the reflection questions that the students are required to answer after each session and the corresponding data statistics. The format of the questions is designed such that the students may answer the questions using short texts and in an informal casual manner. During the course delivery, the instructor manually analyses the textual content to gain insights. The main goal was to identify the topics that the students found challenging. Using this insight, the instructor can then revise the concepts in the next session or provide additional learning material to the students.

Table 1. Directed self-reflection writings –weekly

Question	Data type	# records
Timestamp	Date time	3099
Student name	String	3099
Class Section (Indicates section name and time)	String	3099
List one topic of the class that you enjoyed	String	3092
What was the most challenging topic of the class?	String	3081
What is the overall learning experience today?	Likert1-5	3099
What are your suggestions to improve the class?	String	1277

The average length for the textual input by a student is eight words and thirty-six characters which is like a microblog post or Tweet. Manually analysing the weekly reflections was tedious and painstaking. Based on the descriptive analysis, we observed several challenges related to analysing the text in the reflections. We shall describe these challenges with examples in the following sections together with the solution model designed to overcome these challenges. There are two considerations when designing the solution. First, it should be flexible to handle any topic of interest and second, it should manage the problems associated with short texts.

IV. Solution Model

In view of the limitations that arise due to the short texts in our dataset, we utilize different text mining methodologies to generate the optimal topic model for this dataset. The same methodology was also executed on the following three aspects in the reflection dataset: "Enjoyable", "Challenging" and "Suggestions". However, for brevity, in this paper we will only be discussing the results for one of the aspects; "Challenging".

To make a comparison between the outputs of the different methods used, we will choose a baseline model and compare the other models' performance to it. Figure 1 depicts the solution model design.

A. Data Pre-processing

We observed that informal reflections have noise such as spelling errors, improper sentence structures, abbreviations, and contractions and we apply NLP techniques to handle the noise. To handle spellings, we used a spellchecker, and to handle the contraction and acronyms, we developed a dictionary. Some examples of the terms include "pls", "ty", "I'll" and "don't", which are the short forms for the words "please", "thank you", "I will" and "do not". Since stop words such as "the"," and", "to" and "of" will affect the performance of topic models, we removed them using a dictionary of stop words. Further, to improve the quality of the topics, we generate bigrams to handle words such as "value chain" and "root cause". We removed the bigrams with a threshold of five, since less frequent phrases may create noise in the data [26]. Further pre-processing is applied using the domain knowledge for the course. Firstly, noise data like "N.A" and "None" were filtered as it does not contain any meaningful context related to the actual topic. Secondly, words such as "as is" and "to be" which are important phrases in the domain have been converted to "asis" and "tobe" to prevent them from being removed when filtering out the stop words from the dataset. Thirdly, domain specific short forms, for example, "rcr" or "rcrs" were expanded to "root cause recommendation"; "plm" to "product lifecycle management". Lastly, frequent words and phrases are standardized to aid in topic mining and clustering. Specifically, words such as "internet of thing" to "iot"; "work flow" is converted to "workflow".

B. Content mining – Topics discovery

To discover the topics of interest from the reflections, we used topic modelling techniques from the field of text mining research. LDA models are widely used in extracting topics from textual content [8]. However, LDA models have limitations when applied on short texts. Therefore, we employed LDA as the baseline model to compare with the DMM. Topics can also be extracted using clustering algorithms. We explain the details and settings for each model in this sub-section.

1) Topic Models

a) Baseline LDA– unigram

In the baseline LDA-unigram algorithm, the input corpus is the data with unigrams generated from the data pre-processing stage. LDA was developed by Blei et al. [27] as a generative probabilistic modelling approach to reveal hidden semantic structures in a collection of textual documents. The basic idea is that each self-reflection text exhibits a mixture of latent topics wherein each topic is characterized by a distribution over the words. In our study, we implemented python based LDA mallet code. Figure 2(a) shows the LDA model.

The mixture of topics for documents is indicated by θ and the topic mixture of words is given by ϕ. α and β are the priors and in our settings, we assume few topics to each document and hence α is set low.

b) LDA - bigrams

LDA bigrams are like LDA unigrams. In this approach, we combine the bigrams to the corpus and this forms the input to the LDA model. As a result, the bag of words is both unigrams and bigrams. All the settings remain the same as LDA-unigram.

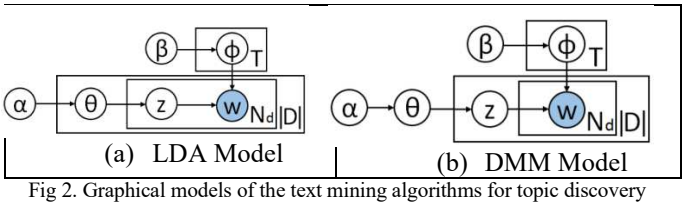

Fig 2. Graphical models of the text mining algorithms for topic discovery

c) GSDMM - bigrams

Dirichlet Multinomial Mixture (DMM) model assumes that each short document has only one topic and works as the basis of our model [11]. GSDMM approach proposed by Yin and Wang [28], can tackle the issue of sparse data in short texts, and able to converge quickly. Yin and Wang demonstrated the algorithm using an analogy of a Movie Group Process. Following is the description of this analogy: There is a class filled with students and each student has a list of his or her favourite movies. Each student is randomly assigned to tables at the start of the class. The professor repeatedly reads the class list. Each time the student is called, the student will select a new table that satisfies the two conditions: the new table should have more students than his current table or the table has students with similar lists of movies. As this process continues, some tables will eventually vanish, and others will have students that share similar interests.

To state it formally in terms of our problem, the "students" here refer to the documents, which are the weekly informal-short-reflections submitted by each student, and the "list of movies" are the words in the reflection documents. The "tables" that the documents are assigned to are the topics of interest that the students write their reflections on. As indicated in Figure 2(b), the main difference is that the topic distribution, θ, is assumed for the entire corpus, unlike LDA model, where it is assumed for each document in the corpus. Gibbs sampling provides the estimation of all the variables and thus generates the matrices for topics-to-words.

Since there is a need to pre-define a T or number of topics, based on the weekly content for the course, T is set to10. The quantitative studies are based on these ten topics and compared against other models.

2) Word2Vec- Clustering Models

Word2Vec is a word embedding representation composed of two types of architecture, Continuous Bag of Words (CBOW) and Continuous Skip-gram Model (skip-gram). CBOW model learns word embedding by predicting words based on the surrounding words, otherwise known as the context. Whereas Skip-gram works in a reverse way where it learns word embedding by predicting the context given the input word. An additional strategy, negative sampling was also introduced. Instead of learning the observed data as positive examples, the strategy focuses on assigning random words from the corpus as negative examples.

The Word2Vec model could potentially be useful for our dataset since words/phrases with seemingly different meanings are associated with the same concepts taught in the module. For example, phrases like “swim-lane” and “workflow” are associated with the topic, “Workflow”. Similarly, “application model” and "solution overview model’ with the topic, “Concept Solution Blueprint”. After training the model with the dataset, the Word2Vec model could pick up such relationships between these phrases and convert them to similar vector representations.

After obtaining the individual word vectors, the average of all word vectors in each sentence is taken to create a summary vector. Clustering algorithms such as K-Means and Agglomerative clustering are then used to cluster all the summary vectors

a) K-Means

K-Means [29] is a simple unsupervised learning algorithm that can be used to solve clustering problems. It starts with defining a few clusters required, for example K cluster. K points or centroids are placed into the space represented by feature vectors. After which, each object is assigned to the group that has the closest centroid through Euclidean distance calculation. When all the objects have been assigned, the positions of the k centroids are recalculated. The assignment process is repeated until there is no change in the position of centroids. K-mean++ implementation from “scikit-learn package” is used with maximum of 500 iterations.

b) Agglomerative

K-Means is based on a partitional clustering approach while Agglomerative clustering, is a hierarchical approach [30]. Agglomerative hierarchical clustering produces a nested sequence of partitions with the top being a single cluster and the bottom, singletons clusters. The creation of subsequent clusters is based on bottom-up comparison of each singleton cluster. At each step, the most similar or closest pair of clusters are merged. This merging process continues until a condition is met, for example, a predefined number of clusters have been formed. “Scikit-learn package” of agglomerative clustering by Ward’s method is adopted in this study.

Like LDA models, since there is a need to pre-define a K or number of clusters, K=10 is selected to align with the weekly content of course. To study the sub-topics, a range of values is tested in this study and K=20 is selected through analysis of Silhouette scoring method. We explain the details in the section V.

C. Evaluations

Coherence scores [31] and perplexity are common quantitative metrics applied for LDA model evaluations. We did not choose to use coherence score and perplexity parameters as in the principal evaluation method. This was because the metrics mainly depend on the co-occurrences of the text or wordnet and this was identified from our preliminary analysis as one of the limitations of our dataset. Hence, we evaluated the quality of the models using human evaluations, both qualitative and quantitative methods. Once the best topic model is identified, we applied the accuracy, F-Score, precision, and recall, to compare the best model with the Word2Vec based clustering models. The details are provided in the next section.

D. Visualizations

To provide insights into the self-reflections, a dashboard is generated from the outputs of the best content-mining model. The insights help the instructor to analyse the learning outcomes and explore the statistics related to the students’ learning skills.

V. Experiments and Evaluations

In this section, we first describe the parameter settings for the topic mining models described earlier. We then present the evaluations followed by a discussion.

A. Parameter settings

For all the models, we set the number of topics to ten. We also studied the coherence score model to gauge the number of topics. For the LDA model, the default settings for hyper-parameters were used. For DMM, “α” influences the probability of a document being assigned to an empty cluster. In our experiments, “α” performed in a stable manner for the testing datasets across all values $0<\alpha<1$. However, the performance decreased slightly when the “α” got larger. Hence, we selected a default value of $\alpha=0.1$. We use the Word2Vec model available in the gensim package and skip-gram with epoch of 100 since it has been commonly known to have a better performance than CBOW. As opposed to the default dimension size of 300, the dimension has been reduced to 20 due to the small vocabulary size of our dataset, which was 1257.

B. Ground truth

Based on each weekly session learning outcomes, the labels for the week are produced by the instructor as depicted in Table 2. Week 8 is the break week and week 13 is the presentation week. From Table 2, we observe that the labels for week 2 and 3, and week 9 and 11 are similar. This indicates that for quantitative evaluations, the number of topics and clusters can be set at 10.

Table 2. Labels based on weekly learning outcomes

Week	Labels
1	‘business process management’, ‘modelling’
2	'modelling', ‘activity’, 'workflow'
3	‘modelling’, 'activity','workflow'
4	'static','activity', 'signavio'
5	'dynamic','activity','signavio'
6	'it requirement','activity','modelling'
7	'solution architecture', 'activity','modelling'
9	'business innovation', 'process innovation'
10	'business innovation', 'process innovation'
11	'process architecture','activity','modelling'
13	'activity'

C. Evaluations

In this sub-section, we describe both quantitative and qualitative evaluations.

1) Quantitative evaluations

a) Comparison of Topic Models

This section answers our first research question, RQ1. Using the human labelling where they provide the rating from 1 to 3

for each topic, we calculate the purity of each topic and identify the incoherent topics. A topic is identified as incoherent if there are more than four incoherent words. To calculate the purity p,

$$p = \sum_{i=1}^{k} \frac{p_i}{k}$$

where k is the number of topics (k=10), and $p_i = \frac{a}{w}$, where a, is the number of coherent words and w is the total top words (w=10). The quantitative results are depicted in Table 3.

Table 3. Topic models - performance comparison

Score	LDA- unigrams	LDA–bigrams	GSDMM-bigrams
Purity Score	0.62	0.65	0.84
Incoherent score	4	3	0

From Table 3, we observe that GSDMM-bigrams outperforms the LDA models with the highest average topic purity score (0.84) and the lowest number of incoherent topics (0). Table 4 depicts output from GSDMM showing the 10 clusters of topics with their respective top 10 words. The human labels are given based on the top words and these labels are used for quantitative evaluations to compare against the best Word2Vec-cluster model.

Table 4. Topic clusters of GSDMM

Cluster	Top 10 topics words	Human labels
0	process, collaboration, workflow, activity, confusing, time, business, little, need, question	modelling
1	innovation, case, business, process, iot, trend, techn ology, chain, presentation, coming	business innovation
2	class, lab, presenting, part, everything, hard, today, slide, think, follow	activity
3	application, modelling, business_canvas, solution_overview, signavio, iot, iot_architecture, triangle, function, gassmann_magic	process innovation
4	use_case, function, package, solution_overview, teaching_case, workflow, diagram, collaboration, application, activity	IT requirement
5	system, resource, workflow, swim_lane, activity, process, automated, swimlanes, interactive, internal_external	modelling
6	signavio, lab, analysis, dynamic_analysis, report, analysis_technique, process, case_study, tobe, cost	dynamic
7	root_cause, issue, recommendation, cause_description, cause, impact, description, finding, process, problem	static
8	process, business, workflow, iot, innovation, modelling, management, modeling, diagram, package	modelling
9	process_orientation, functional, functional_versus, process, business, location, functional_orientation, organisation, alignment, setup	business process managemen t

To evaluate the model, we need to label each reflection based on the reflection-topics distribution. Example labels are shown in Table 5.

Table 5: Self-Reflection labelling using GSDMM outputs

Self-Reflection (Snippet)	GSDMM (Cluster)	GSDMM (Human labels)
"understanding the various organisation models"	9	business process management
"describing components of a business process	0,9	modelling, business process management

Note that topic models allow multiple labels to be allocated to each document, unlike clustering algorithms. Hence, we assign multiple topics to the same sentence with the probability score threshold set to 20%.

b) Comparison of topics models and word2vec-clustering models

Recall that the evaluations between topic models and clustering models are based on accuracy, precision, recall, and F-Score. Table 6 shows the comparison among the best topic model (GSDMM), K-Means, and Agglomerative models.

Table 6. Performance of GSDMM, Word2Vec+K-Means, Word2Vec+Agglomerative based on evaluation metrics

Metric	GSDMM	Word2Vec+K-Means	Word2Vec+ Agglomerative
Accuracy	0.64	**0.78**	0.67
Average Recall	0.64	**0.78**	0.67
Average Precision	0.65	**0.80**	0.66
F1 Score	0.61	**0.76**	0.61

From Table 6, we observe that Word2Vec models have better performance and K-Means has the highest accuracy. Like GSDMM, the human labels are given to K-means clusters based on the top words. The sentences are also labelled based on human assigned labels. Top cluster words and the corresponding human labels can be found in Table 7. On inspection, we observe that the top words in each cluster adequately represent the course topics. Compared to GSDMM outputs, Word2vec K-Means generate coherent and words which are relevant to course content. For example, words in the topic, "business innovation" such as "canvas" (business model canvas) and "gasman_magic" (gasman BPM triangle) are important theoretical aspects of business innovation lecture slides.

Table 7: Results from Word2Vec+K-Means and human assigned label

Cluster	Top 10 cluster words	Human labels
0	use_case, function, package, activity, new, existing, diagram, workflow, modified, type	IT requirement
1	analysis, cost, tobe, resource, process, modelling, report, system, signavio, risk	dynamic
2	signavio, lab, class, teaching_case, dynamic_analysis, part, using_signavio, presenting, everything, time	activity
3	innovation, business, business_canvas, case, triangle, process, gassmann_magic, trend, template, technology	business innovation
4	root_cause, issue, recommendation, cause_description, cause, impact, description, finding, problem, determining	static
5	application, solution_overview, iot, modelling, solution, coming, finding, modeling, process, function	solution architecture
6	process, case_study, question, collaboration, business, iot_architecture, analysis_technique, need, presentation, understand	activity
7	process, business, management, modelling, iot, modeling, component, categorizing, attribute, executive	modelling
8	process_orientation, functional, functional_versus, functional_orientation, location, organisation, process, organization, setup, role	business process management
9	workflow, process, collaboration, diagram, drawing, business, package, activity, still, little	modelling

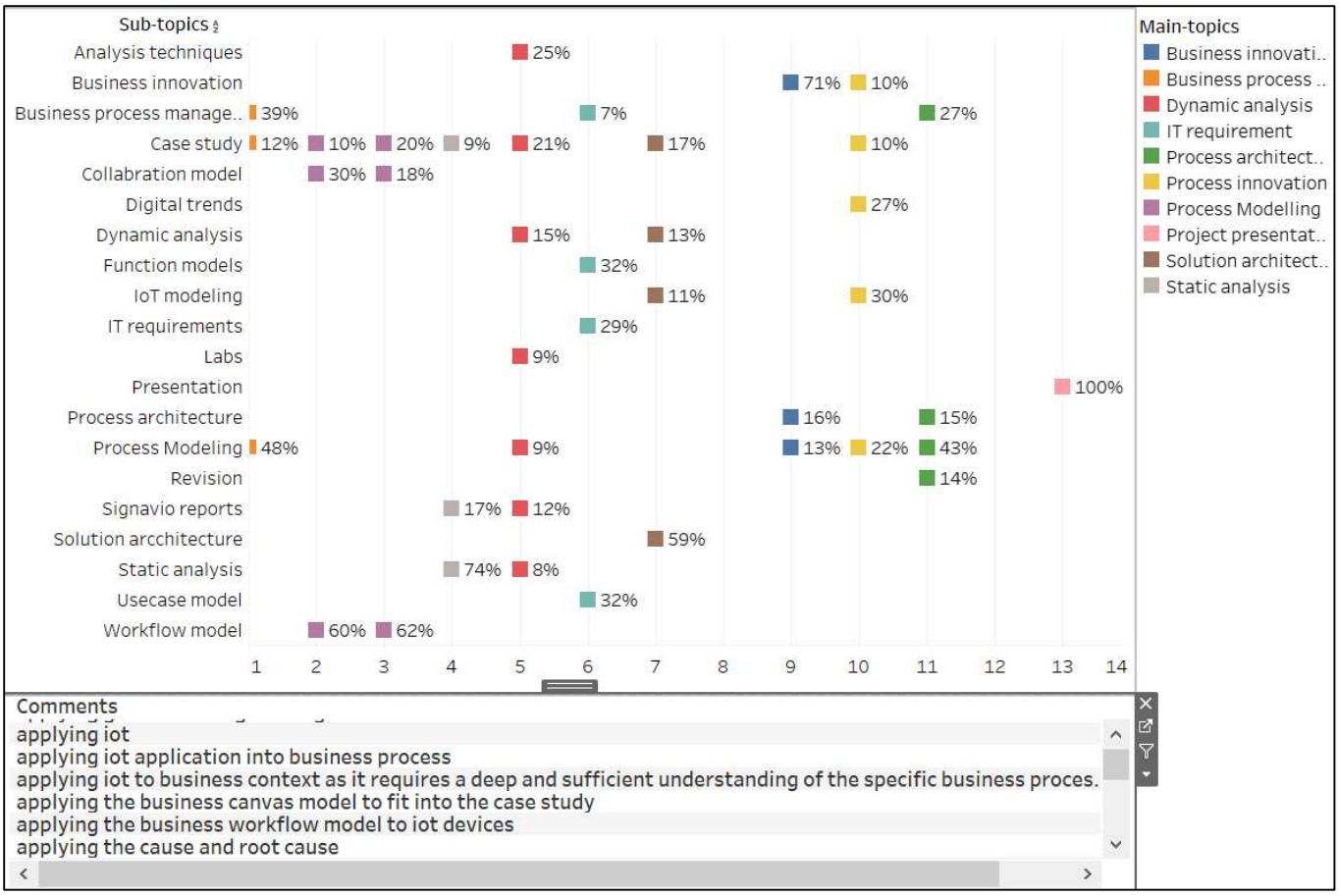

Fig. 4. Dashboard – Report on weekly challenging sub-topics and the corresponding student ratings

Further, we conducted a detailed analysis by week using the confusion matrix of Word2Vec+K-Means as depicted in Figure 3.

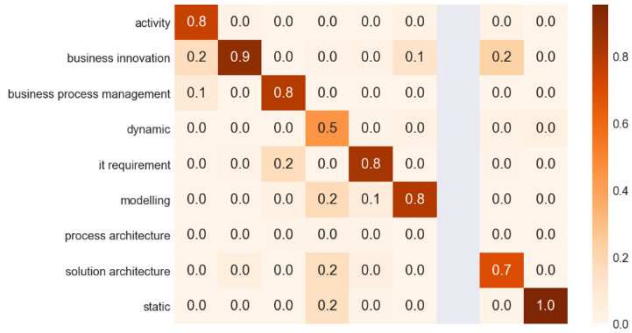

Fig 3. Confusion matrix to depict the accuracy (rounded) by topics.

From Figure 3, we observe that all the topics, except "dynamic" have an accuracy of greater than equal to 70%. The topic, "dynamic" has low accuracy of 50% (0.5). Further analysis revealed that the topic, "dynamic", has words that overlap with other topics resulting in lower performance. Additionally, the topic "process architecture" is not analysed since the label is missing from the k-means clusters (Table 7). Based on the results presented, the answer to RQ1 is that machine learning algorithms are useful in detecting the course topics from self-reflections. Overall, K-Means performed with high accuracy.

2) Dashboard for insights

From the previous analysis, we observe that K-Means performs the best. To aid the instructor, we provide a dashboard, see Figure 4. In Figure 4, we observe multiple sub-topics for each week that the students have rated as challenging. For each sub-topic, the dashboard shows the percentage of students who found it challenging. Additionally, the dashboard also shows the relevant comments from the students' self-reflections. Recall that to generate the sub-topics, we choose K to be 20 based on Silhouette scores. Therefore, we generated 20 sub-topics, labelled them, and fed them as inputs to the dashboard. The colours in the graph indicate the different main-topics.

3) Qualitative evaluations

This section answers our second research question, RQ2. Using this visualization, the instructor can identify the most challenging topics and the corresponding self-reflection comments regarding these topics. To answer RQ2, we show an example of how the instructor can use this dashboard along with Table 8 to intervene in the student learning process.

Table 8: Weekly learning outcomes from lecture and corresponding challenging sub-topics from self-reflections

Week	LOs from Lecture	LOs from the dashboard and % of students find it challenging
1	• Understand the importance of business and IT alignment • Explain the importance of the need for a business process management methodology (BPM) • Explain the phases of Business Process Management • List process modelling activities and its steps.	Case study 12% BPM – 39% Process Modelling- 48%
5	• Understand the methodology for performing Dynamic Analysis of a business process • Apply systematic methodology in a scenario. • Perform Dynamic Analysis using Signavio. • Compare the as-is and to-be models using static and dynamic reports	Analysis techniques -25%, Case study –23%. Dynamic analysis – 15% , Labs – 9%, Process modelling – 9%, Signavio reports- 12%, Static analysis - 8%
9	• Understand Business Innovation and its role in business survival • Define 3 types of innovation and describe their roles in business survival. • Apply the innovation approaches in the given scenarios. • Discuss innovation in given business process models	Business innovation – 71% Process architecture – 16% Process modelling – 13%

Table 8 shows the LOs of the week and the corresponding students' ratings in terms of challenging topics. This table is created by extracting the weekly LO's defined by the instructor and mapping them to the weekly challenging sub-topics in Fig 4. We show only selected sample weeks for analysis. The weekly sub-topics obtained from the self-reflections help the instructor to know which concepts are more challenging and hence the instructor can adapt the pedagogy to help further enhance understanding of these concepts. The instructor may plan accordingly for the subsequent week's session by repeating some of the content or by providing more examples to help students to understand, critique, and apply the concepts, and thus move to higher cognitive levels.

For example, from Table 8, we observe that 48% of students did not understand "process modelling" concept in week 1 which is one of the LOs. This is a significant number and requires immediate action by the faculty. Hence, the faculty team discussed this issue and created additional process modelling exercises. For the following week, the schedule is adjusted to add these exercises to the class and students expressed positive feedback under the enjoyable topics (not shown in the graphs).

D. Discussions

From our experiments and finding, we observe that, for the given problem and data set, Word2Vec+K-Means is the best performing approach among all three approaches. Even though GSDMM, with the single-topic assumption, can model our data well but it still suffers from the data sparsity issue. Specifically, since DMM uses word occurrences to determine similarity to predict the topic they belong to, words that are semantically related are unable to be captured due to low co-occurrences. This limitation is addressed using Word2Vec, through its internal neural representation to learn the word embedding which uses a similar vector to represent semantically similar words.

One disadvantage of K-Means clustering is the need to pre-define a K or number of clusters. This value can be influenced by the content, especially when it is applied to different domain or courses. In order to find the K value, an automated approach such as Markov stopping moment can be included to find the optimal clustering [15]. The accuracy of the models can be improved by studying sequential learning models and BERT networks, which we leave as future work.

Though, through qualitative evaluation the instructor can identify the challenging topics and implement the appropriate intervention in the subsequent class session. In its current form, the key limitation of our work is that the impact of the intervention is not evident. We leave it to future work to improve the dashboard analytics to add the impact through the listing of enjoyable topics and suggestions from students.

VI. CONCLUSION

In this paper, we propose a solution for collecting and analysing end of session student self-reflections that are informal and short in nature. The main contribution of this paper is the novel solution model that combines topic extraction to identify concepts that students find challenging and then mapping these topics to the weekly learning outcomes. The solution model is then evaluated by studying its implementation in an undergraduate course by comparing three different content mining models.

The main limitation of our work stems from the ability to generate topics from a data set of informal short reflections, where there is a sparsity of words. In the approach adopted by Li et al.[32], they extended the traditional DMM to include word embedding to capture semantic relatedness between words. The algorithm samples a topic from a document, and words highly related are selected. Auxiliary word embedding is used to promote semantically related words, which will form better topics as words that are highly semantically related but have fewer co-occurrences will still be grouped together. Hence, we can leverage word embedding in this manner to solve the sparsity issue. We are currently enhancing the dashboard interface where the statistics for each topic is integrated with Bloom's verbs and displayed in a user-friendly manner. A recommendation system for the instructor to highlight the most critical course concepts will also be one of our future tasks.

REFERENCES

[1] C. Park, "Engaging students in the learning process: The learning journal," *J. Geogr. High. Educ.*, vol. 27, no. 2, 2003, doi:

10.1080/03098260305675.
[2] K. A. Noels, R. Clément, and L. G. Pelletier, “Perceptions of teachers’ communicative style and students’ intrinsic and extrinsic motivation,” *Mod. Lang. J.*, vol. 83, no. 1, 1999, doi: 10.1111/0026-7902.00003.
[3] K. Thorpe, “Reflective learning journals: From concept to practice,” *Reflective Pract.*, vol. 5, no. 3, 2004, doi: 10.1080/1462394042000270655.
[4] S. Gottipati, R. Barros, K. J. Shim, "Mining Informal and Short Weekly Student Self-Reflections for Improving Student Learning Experience" (2021). AMCIS 2021 Proceedings. 21. https://aisel.aisnet.org/amcis2021/is_education/sig_education/21
[5] F. Marefat, “The Impact of Diary Analysis on Teaching/Learning Writing,” *RELC J.*, vol. 33, no. 1, 2002, doi: 10.1177/003368820203300106.
[6] G. Cheng and J. Chau, “Digital video for fostering self-reflection in an ePortfolio environment,” *Learn. Media Technol.*, vol. 34, no. 4, 2009, doi: 10.1080/17439880903338614.
[7] S. Gottipati, V. Shankararaman, and S. Gan, “A conceptual framework for analyzing students’ feedback,” in *Proceedings - Frontiers in Education Conference, FIE*, 2017, vol. 2017-October, doi: 10.1109/FIE.2017.8190703.
[8] S. Gottipati, V. Shankararaman, and J. R. Lin, “Latent dirichlet allocation for textual student feedback analysis,” in Proceedings of the 26th International Conference on Computers in Education ICCE 2018. APSCE, 2018.
[9] T. D. Ullmann, “Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches,” *Int. J. Artif. Intell. Educ.*, vol. 29, no. 2, 2019, doi: 10.1007/s40593-019-00174-2.
[10] M. Liu, S. B. Shum, E. Mantzourani, and C. Lucas, “Evaluating machine learning approaches to classify pharmacy students’ reflective statements,” in *Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science*, 2019, vol. 11625 LNAI, pp. 220–230, doi: 10.1007/978-3-030-23204-7_19.
[11] J. Yin and J. Wang, “A Dirichlet multinomial mixture model-based approach for short text clustering,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 233–242, doi: 10.1145/2623330.2623715.
[12] R. Walker, “GSDMM: Short text clustering,” *GitHub*, 2017. https://github.com/rwalk/gsdmm.
[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations ofwords and phrases and their compositionality,” 2013.
[14] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, Z. A. A. Alyasseri, and S. N. Makhadmeh, “A novel hybrid multi-verse optimizer with K-means for text documents clustering,” *Neural Comput. Appl.*, vol. 32, no. 23, 2020, doi: 10.1007/s00521-020-04945-0.
[15] S. S. Bodrunova, A. V. Orekhov, I. S. Blekanov, N. S. Lyudkevich, and N. A. Tarasov, “Topic detection based on sentence embeddings and agglomerative clustering with markov moment,” *Futur. Internet*, vol. 12, no. 9, 2020, doi: 10.3390/fi12090144.
[16] E. G. Bugg and J. Dewey, “How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process,” *Am. J. Psychol.*, vol. 46, no. 3, 1934, doi: 10.2307/1415632.
[17] S. Veine *et al.*, “Reflection as a core student learning activity in higher education - Insights from nearly two decades of academic development,” *Int. J. Acad. Dev.*, vol. 25, no. 2, 2020, doi: 10.1080/1360144X.2019.1659797.
[18] M. Gläser-Zikuda, “Self-Reflecting Methods of Learning Research,” in *Encyclopedia of the Sciences of Learning*, 2012.
[19] J. Amulya, “What is reflective practice?,” *Center for Reflective Community Practice: Massachusetts Institute of Technology*, 2004. http://www.itslifejimbutnotasweknowit.org.uk/files/whatisreflectivepractice.pdf.
[20] J. Dykes and M. Meyer, “Reflection on reflection in design study,” *arXiv*. 2018.
[21] A. Wallman, A. K. Lindblad, S. Hall, A. Lundmark, and L. Ring, “A categorization scheme for assessing pharmacy students’ levels of reflection during internships,” *Am. J. Pharm. Educ.*, vol. 72, no. 1, 2008, doi: 10.5688/aj720105.
[22] D. Kember *et al.*, “Determining the level of reflective thinking from students’ written journals using a coding scheme based on the work of Mezirow,” *Int. J. Lifelong Educ.*, vol. 18, no. 1, 1999, doi: 10.1080/026013799293928.
[23] T. D. Ullmann, F. Wild, and P. Scott, “Comparing automatically detected reflective texts with human judgements,” in *CEUR Workshop Proceedings*, 2012, vol. 931.
[24] A. Gibson, K. Kitto, and P. Bruza, “Towards the Discovery of Learner Metacognition From Reflective Writing,” *J. Learn. Anal.*, vol. 3, no. 2, 2016, doi: 10.18608/jla.2016.32.3.
[25] V. Kovanović *et al.*, “Understand students’ self-Reflections through learning analytics,” 2018, doi: 10.1145/3170358.3170374.
[26] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. 2008.
[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.
[28] J. Yin and J. Wang, “A Dirichlet multinomial mixture model-based approach for short text clustering,” 2014, doi: 10.1145/2623330.2623715.
[29] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1.
[30] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” 2000. doi: 10.1109/ICCCYB.2008.4721382.
[31] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing semantic coherence in topic models,” 2011.
[32] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Enhancing topic modeling for short texts with auxiliary word embeddings,” *ACM Trans. Inf. Syst.*, vol. 36, no. 2, 2017, doi: 10.1145/3091108.