Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Thematic Summarization Dashboard for Navigating Student Reflections at Scale

Yuya ASANO[a*], Sreecharan SANKARANARAYANAN[b], Majd SAKR[b] & Christopher BOGART[b]

[a]*Intelligent Systems Program, University of Pittsburgh, USA*
[b]*School of Computer Science, Carnegie Mellon University, USA*
*yua17@pitt.edu

Abstract: Instructors often ask students to reflect on projects or tasks because it has been shown to be effective for learning. Instructors also use these reflections to improve future offerings of a course. Sifting through reflections manually, however, is both time-consuming and inefficient, especially for large courses. This paper describes a method for organizing student reflections by named entities (i.e., topics of interest) and instructor-defined "themes" to produce summaries that better meet the needs of instructors. Named entities are first extracted from the reflection corpus. Upon choosing one named entity to explore, sentences mentioning that entity are collated from across student reflections. The selected sentences are then classified into instructor-defined themes. Instructors can choose to re-define themes as necessary with support from the system in the form of prevalence statistics and theme-definition suggestions. Finally, a summary of student reflections for each theme is provided. This process and the resulting summaries were evaluated in a semi-structured Wizard of Oz interview study with the teaching assistants of a 160-student graduate-level course on Cloud Computing offered online to the students at Carnegie Mellon University. Results from quantitative Likert-scale analyses and qualitative coding show that teaching assistants preferred our topic and theme-focused summaries over general summaries generated from a random subset of student reflections. Deployment in the form of an instructor-facing dashboard and improvement to the system to allow for uncommonly expressed content to be better discoverable through the dashboard are planned for future work.

Keywords: Instructor dashboard, student reflections, natural language processing, summarization, semi-structured interview, qualitative coding

1. Introduction

Written student reflections either about all or part of a course are a common way of enhancing student learning (Baird et al., 1991; Lee, & Hutchison, 1998; Menekse et al., 2011). For instructors, however, these reflections are rich resources that help them understand what students liked/disliked or found easy/difficult about aspects of the course and improve future iterations of the course. Even though this practice is relatively common, technology-supported ways of sifting through large amounts of unstructured student data have not been effectively addressed in prior work. Manually sifting through this feedback is not just time-consuming but can be subject to instructor bias (Mosteller, 1989). A data-driven way to efficiently navigate written reflections of students, therefore, is an important problem to address.

In order to address this problem, we prototyped an instructor-facing dashboard that provides summaries of student reflections organized by named entities (i.e., topics of interest) and instructor-defined "themes." With named entity recognition and theme labeling, we try to mimic the widely used process of developing codes and sorting them into categories to analyze qualitative data (Erlingsson, & Brysiewicz, 2017). The dashboard first sifts through student reflections to identify the most prevalent named entities (Ex: "Java," "MongoDB," "Scala," and "Spark"). Instructors can choose a named entity to explore, resulting in sentences mentioning this named entity being collated from across student

reflections. These sentences are displayed on the dashboard, along with summaries organized by user-defined themes and prevalence statistics about the percentage of student reflections classified into the theme. The entire process can be repeated with different named entities as well as instructor-defined themes. The process, as well as the resulting summaries, were evaluated using semi-structured interviews with the teaching assistants (TAs) of a large online graduate-level course on Cloud Computing[1] offered to 160 students at Carnegie Mellon University. Each TA was responsible for a project unit that students wrote reflections on. They were asked to evaluate the process of defining their own themes and the summaries produced from them. Analysis of Likert-scale questions and qualitative coding of interview transcripts show that the summaries generated by our system are more helpful for instructors in understanding the process students followed when doing assignments than summaries of randomly sampled reflections. We discuss the design implications for summarization tools that better satisfy the needs of instructors.

2. Related Work

There is a significant body of research showing that eliciting student reflections on assignments and lectures helps consolidate learning and improve outcomes (Baird et al., 1991; Lee & Hutchison, 1998). Following this research, several courses embed reflection exercises students can participate in at regular intervals during the course such as the ones by Menekse et al. (2011) and Fan et al. (2017). These reflections not only improve students' understanding of the subject matter but also provide instructors with feedback on their teaching and students' learning. Instructors use these reflections to understand students' experiences in learning and facilitate changes in those experiences (Baird et al., 1991). However, manually coding and summarizing raw written reflections are too laborious for instructors (Mosteller, 1989), and they rarely receive enough support from their institutions to maintain the cycle of analyzing reflections and taking action (Harvey, 2003). Therefore, efficiently parsing and analyzing the content of these student reflections becomes an important problem to solve.

Prior attempts at doing this have been seen, for example, in the work of Fan et al. (2017)'s CourseMIRROR app, where significant work was done not only in exploring the educational benefits of reflection but in extracting insights from the resulting corpus of text. Their app summarizes the reflections they collect not only for instructors to use for course improvement but for students to think about the lecture from multiple perspectives. Their summaries are lists of semantically clustered phrases representing answers to a question about what concepts were "confusing or needed more detail" in a lecture.

Outside of the context of education, there are various algorithms to generate summaries automatically, too. There are two types of summaries: extractive summary and abstractive summary, but we focus on the latter because student reflections are inherently diverse. In the abstract summarization task, neural models have been shown to outperform others (Rush et al., 2015). Even though these models initially targeted relatively short text, researchers have proposed models that can handle a large amount of text such as scientific papers (Beltagy et al., 2020; Cohan et al., 2018; Zaheer et al., 2020).

Their techniques, however, do not attempt to summarize broader themes embedded in responses to more open-ended questions. Yao et al. (2017) give an overview of summarization techniques, saying that most draw on three components: sentence scoring for importance, sentence selection (for coherence, redundancy, and length of final summary), and sentence reformulation (modifying selected sentences into a coherent summary). This technique, we believe, can be used to build a tool to help instructors more efficiently sift through a large number of student reflections. Our work not only presents a first cut summarization of student reflections by theme but also provides instructors with control over choosing topics and their own themes to explore the student reflection data better while enjoying recent advances in neural models of summarization. Our approach utilizes named entity recognition to automatically extract topics of interest from a student reflection corpus and word embedding to help instructors improve their own themes.

[1] http://www.cs.cmu.edu/~msakr/15619-s21/

3. Course Context and Data

This study was conducted in a completely online semester-long graduate-level course on Cloud Computing offered to the students at Carnegie Mellon University and its campuses in Pittsburgh, Silicon Valley, and Rwanda. As a project-based course, a major component of the class is the completion of 10-12 programming projects of significant complexity, using libraries, resources, languages, and tools provided through commercial cloud providers. Students submit solutions to an auto-grading service as many times as they like before the deadline. The auto grader evaluates source code properties, behavior, and performance of students' code and provides feedback allowing students to iterate and improve their solutions.

The course provides significant scaffolding explanations and videos with the projects, separate single-topic primers demonstrating the deployment and use of required technologies, an online textbook teaching underlying concepts, quizzes, small-group activities, and a reflection and discussion forum.

This study focuses on the analysis of the reflection/discussion forum. After each project deadline, students are required to post a reflection paragraph, prompted by the following question:

> *Consider the following topics when creating your post, however, you should never share any code snippets in your reflection:*
> - *Describe your approach to solving each task in this project. Explain alternative approaches that you decided not to take and why.*
> - *Describe any interesting problems that you had overcome while completing this project.*
> - *If you were going to do the project over again, how would you do it differently, and why?*
>
> *After completing this task, confirm that your Reflection Score has been automatically updated on the scoreboard before the project deadline.*

The forum's primary intent is to spark reflection and self-explanation. However, it has also been useful as a way of gathering feedback about the project for iterative improvement of the curriculum from semester to semester.

Students are then asked to reply to three other students' reflections. The reflections are only a small part of the project grade, and points are assigned if the student writes anything at all. However, perhaps because the reflections are seen and discussed by other students, students typically reflect substantially on the project.

A large team of TAs helps operate the course; one TA is assigned responsibility for deploying, supporting, and evaluating each project during the term. After the project has been completed and fully graded, the responsible TA presents to the TA group and instructor an overall evaluation of the project, including a summary of students' responses to a post-project survey, student reflections and discussions, and TAs' experiences with students seeking help in the office hours. When analyzing the written student reflections, TAs read all reflections and select a few representative ones to present to the group. They are asked to identify issues raised through the reflections that should be addressed for future offerings of the project.

4. Methods

The dashboard we have designed helps instructors navigate from an overview to thematic summaries and fine details to understand students' thoughts and opinions on the course and improve it. An instructor would use the tool to first investigate at a high-level what topics (such as tools, services, or languages) were most discussed by students and then drill down into each topic to see several thematic summaries of what students said about it, such as difficulty or usefulness. Since instructors may be interested in different aspects of students' perspectives, depending on the instructor's knowledge and

concerns about the project, we allow them to create and modify their own themes. Finally, instructors may drill down further to see samples of text unclassified into any themes or browse the raw reflections directly. By investigating topics broadly, then diving into details, instructors can "take the temperature" of the class's opinions of the project, topic by topic, and dive into the details to understand the reasons so that they can make informed recommendations for future improvement.

In this paper, we focus on describing and evaluating the thematic summarization aspect of the dashboard: a mixed-initiative machine learning algorithm for automatically organizing and summarizing student reflections. We perform a two-level classification of each sentence by named entities (section 4.1) and user-defined "themes" (section 4.2). Summaries are then produced for each of these fine-grained categories by the summarizer (section 4.3).

4.1 Named Entity Recognition

To identify the topics of interest in student reflections, we performed named entity recognition using the Python package spaCy (Honnibal et al., 2021). According to the Seventh Message Understanding Conference, "Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts" (Chinchor, 1998), but, in our context, we treat them as names of tools, techniques, and terminologies students learn in class. Therefore, when extracting the named entities, we focused only on five types in OntoNotes 5 (ORG, PRODUCT, EVENT, WORK_OF_ART, and LANGUAGE) (Weischedel et al., 2013). Named entity recognition allows instructors to quickly grasp which tools, techniques, and terminologies students talk about in their reflections, which is one of the crucial steps in qualitative analysis.

After obtaining a list of named entities for a particular class module, we classify each sentence of reflections into the entities. We do not classify entire reflections as a unit because students often talk about completely different topics within one reflection. For example, a student wrote, "*Learned from last several projects, I started relative[ly] e[ar]ly this time. H[B]ase is more expensive than I expected. In this project, I was able to explore the u[sa]ge of Hibernate Application, RDBMS and NoSQL databases. I get to understand we have to choose specific techniques based on the user scenario."* They talked about HBase in the second sentence, but the preceding and following sentences were not directly related to HBase. Thus, only the second sentence should be considered when we summarize reflections about HBase.

Before going to the thematic labeling in section 4.2, an instructor has to select one of the entities extracted from student reflections to explore further. To help them decide which entity to choose, our system shows them a histogram of the number of appearances of the entities. An example of one such histogram is shown in Figure 1. Instructors can repeat the entire process, as necessary, with different entities to explore each time a selection is made.

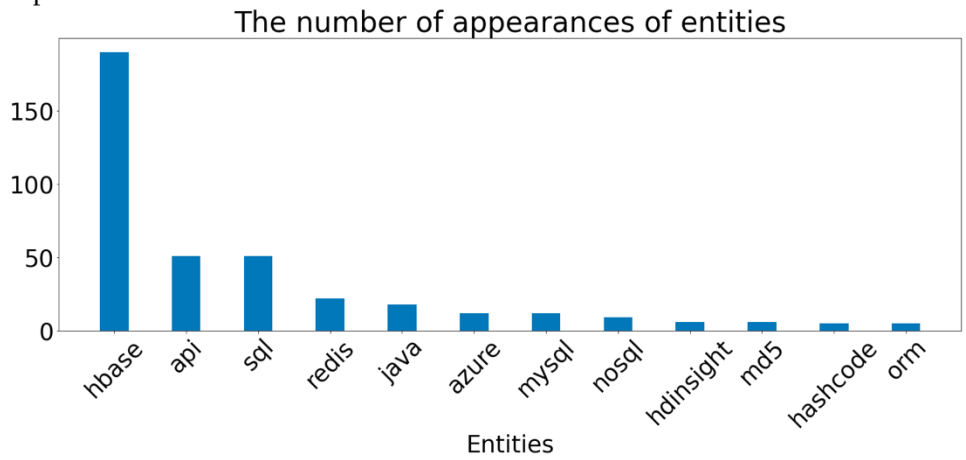

Figure 1. A Sample Histogram of Named Entities.

4.2 Thematic Labeling

After an instructor chooses the named entity that they want to focus on, our system asks them to define a "theme" about the entity. A theme captures how students perceive or feel about the entity selected in the earlier step and is defined by a list of semantically coherent words. For example, the theme "Difficulty" tries to capture if students see the entity as easy or difficult, and it could have the following keywords: "straightforward," "easy," "difficult," "challenging," and "struggle." Next, our system classifies the sentences in student reflections that include the keywords into the user-defined theme. For example, if the instructor had chosen HBase in the previous step, the classified sentences would be treated as "difficulty in HBase." At the same time, the system suggests words in the reflections that have similar meanings to the keywords in the theme, based on the cosine similarities derived from GloVe word embeddings (Pennington & Manning, 2014), by displaying example sentences in reflections that have similar words, as illustrated in Figure 2. Our system also tells the percentage of students who talked about the theme. The instructor may then revise their theme according to the suggestions and the ratio and then run our system again to iteratively improve the theme. In addition, unclassified sentences are also made available to help the TA or instructor later discover other themes related to them.

```
Proportions of the themes:
Difficulty: 0.3252032520325203

Examples of words and sentences picked by embeddings for theme Difficulty
"especially"
Redis required a bit of reading but the WORST was HBASE tasks, especially the row key design.
It is especially true for the HBase task, since I never used it before.
I stuck on the HBase task for a long time, especially in the row key design.

"simple"
In both MySql and HBase questions, I used the simplest scripts to accomplish the requirements.
I also tested queries in MySQL shell, but I seldom did this in HBase, for a simple reason that
HBase shell is more difficult to interact.
but I am totally new to HBase and it seems to me that using the Java API one has to write a lot
of code to do some simple filtering.

"complicated"
The hbase api in general was extremely complicated.
HBase is pretty complex comparing to sql databases, especially the query syntax is a little
complicated.
I think Hbase is more complicated, especially when we are using java API.
```

Figure 2. A Screenshot of Suggestions for Similar Words by Our System. In this example, it suggested adding the words "especially," "simple," and "complicated" to the theme "Difficulty." As a Wizard of Oz study (see Section 5.1), we asked TAs to pick words from the suggestions and manually added to their list of keywords.

4.3 Summarization Technique

We used the Longformer Encoder-Decoder (Beltagy et al., 2020) to summarize the classified reflections. The Longformer was pre-trained to generate abstracts from the papers published in PubMed, using the dataset provided by Cohan et al. (2018). For each named entity and theme, we collected all of the students' sentences classified into the entity and theme and concatenated them into a single string. This string was the input to Longformer to be summarized.

5. Evaluation

5.1 Participants and Method

We interviewed eleven graduate students who had served as TAs (we call them TA 1 to 11) of the course as described in Section 3. Each TA owned one or two projects in the course in which they had already manually summarized student reflections. We performed a Wizard of Oz walkthrough of a prototype system with each TA, implemented in Jupyter notebook. The Jupyter notebook was operated by the researcher but made visible to the TA over a Zoom video call. In each interview, researchers explained the workflow in section 4 and walked the participant through an interaction with the system, interleaving the requested actions below with explanations and instructions:

- **Choosing a topic**: Participants were first shown a bar graph of the most common named entities found in reflections from their project and were asked to choose an entity to explore further.
- **Choosing theme words**: The researcher then explained the system's concept of *themes* and asked the participant to choose a set of keywords representing a theme.
- **Improving theme words**: The researcher types the theme words into Jupyter notebook python variables, runs the cell, and asks the participant to examine the output (suggested other words that may fit the theme) and revise their set of theme words. Participants iterated this step until they were satisfied with the theme and wanted to go on.
- **Generating a summary**: The researcher triggered the Jupyter cell creating a summary from the final chosen theme.
- **Comparing with a random summary**: The researcher triggers a final cell that shows a Longformer summary of random sentences from reflections, without classification for comparison. We did not use all reflections of a project because there was a maximum number of tokens Longformer could handle at one time (Beltagy et al., 2020).

After TAs interacted with our system and read summaries, we asked the questions in Table 1. Each interview took about 30 to 45 minutes.

Table 1. *Questions asked to TAs during the Evaluation*

Questions	
Q1	(After finishing defining a theme) Did you have any difficulty interacting with our system? Why?
Q2	(After showing our summary) Which parts of the summary are useful to a TA? Why?
Q3	Which parts of the summaries are not useful to a TA? Why?
Q4	How useful would this summary be to a future TA? Rate it on a 7-scale Likert scale (1 is not useful at all, and 7 is very useful).
Q5	What elements did you see in the actual reflections that you wish were included in the summaries?
Q6	(After showing another summary from random reflections) How does it compare to the summary above? Rate it on a 7-scale Likert scale again.

Interviews were performed by one researcher, and most were attended by at least one other researcher. Interviews were transcribed by two researchers and qualitatively coded by a single researcher, identifying 10 themes across the 11 interviews, presented in Table 2; these were discussed and revised with two other researchers who had attended sessions. The most prominent themes are further discussed in the following section.

Table 2. *List of Themes Identified from Interviews with TAs*

Themes	Description
Good summary	Our summary was good.
Unuseful summary	Some parts of our summary were not useful.
Not saying concrete challenges	Our summary did not say concrete challenges students faced.
Something missing	Our summary missed something.
Missing new points of view	Our dashboard may prevent instructors from discovering new perspectives.
Better than random sample	Our summary was better than that of randomly sampled reflections.
Summary of random samples is better	A summary of randomly sampled reflections was better than ours.
Difficulty in thematic labeling	Thematic labeling was easy or difficult.

Intermediate outputs are useful	A list of reflections presented during thematic labeling is useful.
Suggestions for future	TAs made suggestions for new functionalities of our dashboard.

5.2 Results

Overall, TAs told us that our thematic summaries matched their expectations about students' experience. In some cases, they told us our summaries clearly articulated the steps students had followed (Q2). For example, TA 4 stated the evidence that students were able to go through the documentation of syntax and code snippets was useful, highlighting the part of our summary that said Neo4J's documents "are useful and we can quickly pick them up by trying out in the shell." TA 11 told us that the thematic summary showed students' workflow of designing data structure of key-value pairs by using Hadoop map reduction.

When we compared ratings of thematic summaries (Q4, Table 1) with random sample summaries (Q6, Table 1), the average rating of thematic summaries was higher, as shown in Table 3. TAs said that the summaries of randomly sampled reflections did not tell them any new information. For example, TA 8, who rated the thematic summary higher, said the random sample summary discussed Piazza, a Q&A forum used by the course; however, this was course infrastructure, not a topic taught in the class; TA 8 did not find this summary useful because she already knew students relied on it. TA 11, who also preferred the thematic summary, said the random sample summary sounded good but only related opinions and facts the TA already knew:

> *At first glance, [the summary] looks more helpful, but it mostly reinforced assumptions about the students: first time using Hadoop. I already picked that up in office hours. The rest of the pieces are more just summarizing the background about MapReduce*.

Table 3. *The Ratings of our Summaries and Summaries of Randomly Sampled Reflections from Q4 and Q6 of Table 1*

	Our Summaries	Random Summaries
Average Ratings	4.93	3.36
Standard Deviation	1.08	1.85

However, more than half of the TAs thought the summaries of random samples were still useful (Q6) because they included the learning objectives and showed the general consensus of the whole class, and four TAs (TA 2, 5, 6, and 9) rated the random summaries higher than the thematic summaries. For example, one of the learning objectives of the project TA 2 and 6 owned was to differentiate between Spark and MapReduce. Both TAs agreed that the summaries of randomly sampled reflections showed that students had learned this objective, even though instructors often struggled with helping students differentiate between these tools.

In addition, five TAs pointed out that our summaries did not address concrete problems faced by students and the causes of those problems (Q3 and Q4). TA 5 said,

> *"The challenges he faced while implementing the program" [are] something we should solve. ... I'd like to know what challenge that was and if it was something we ignored or we intended to do.*

TA 3, whose rating of the thematic summary was below the average, told us that knowing *why* those problems had happened and *how* students had solved them was important because this could tell TAs what to highlight in the explanations of the projects and where to offer more help to students.

Another concern a few TAs had was that they could only see what they had already expected to see before reading reflections because of the keyword-based theme labeling process. TA 2 worried about failing to catch unexpected things in student reflections. He said he would like to know whether students' struggle was their fault or instructors' fault. TA 4 added that he would be less likely to miss out on anything if he went through every reflection manually. Although our dashboard prototype does save a file of unclassified reflections, TAs did not have a chance to look at these in our study due to the limited interview time.

In terms of the usability of our system (Q1), TAs found it difficult to define themes at first. For instance,

TA 9 told us:

> *I was thinking that, for TAs, it's better that you have some hints about what categories could be and for each category what keywords are very likely to be there. I think as a new user to this system, it should take me some time to accommodate. I'd have to try different categories until I realize that I can get some information using such categories but not the others*.

Nevertheless, there is some evidence that this theme definition process can be learned; all three TAs (TA 5, 6, and 7) who had a chance to define two themes rated the second summaries higher or as high as their first try (from 3, 4, and 6 to 5.5, 6, and 6, respectively). Moreover, some TAs stated that they learned useful information during the theme-development process itself; being shown the example sentences and the percentage of the students would be useful because these can reduce the burden of going through unorganized reflections;

> *Basically, every time someone talks about PageRank it's showing here, so I guess that's a good thing. For example, when I'm reading the hundred ... student reviews. You had to do ... that grouping [by] yourself. So, everyone is like 'Okay, these students talk about PageRank.' ... Then you move on to the next and like 'Oh, the students [are] complaining about Scala. Okay,' and that. Having to do that context switch with every student and having to prepare yourself can be tiring* (TA 2).

TA 10 added that it would also be helpful for instructors to know how many reflections are classified into the themes defined by them.

6. Discussion

In this paper, we proposed a novel way to summarize student reflections through two-level classification by named entities and user-defined themes, which is a set of keywords, and compared it with summaries of randomly sampled reflections generated by the same summarizer. Our study with past TAs revealed that our thematic summaries were more useful and better at describing the process students took. In addition, we have found the classification helps TAs reduce their cognitive burden because they can focus on one topic and theme at a time, rather than having to switch contexts constantly if they were to simply read one unrelated student reflection after another. However, this does not suggest that our summary of a selected topic and theme can completely replace summaries of the whole corpus because the latter can show the general consensus of the class. Some TAs indicated such consensus, or even the most commonly used words in all reflections, would be useful when defining themes for our summaries, a task many of them found difficult at first. For example, TA2 said:

> *[It] would be interesting to see the summary [of all reflections] first and then drill down by specific words because I'd be interested to see what's the general consensus with PageRank. ... Imagine, I wasn't expecting complaints about PageRank so [would] be like "oh let's drill down and find those words; why are people complaining, or what they are saying."*

TA 3 suggested showing the most commonly used tokens on the dashboard so that instructors can tell what themes they should define. These testimonies imply that a high-level picture of the entire corpus would help them to make a better selection of keywords in theme labeling (section 4.2).

6.1 Design Implications

Although TAs had a generally positive reaction to the tool, their feedback and our experience building the system suggests several general suggestions for tools such as ours that help build thematic summaries of student reflections:

- **Systems for characterizing student reflections should include thematic summarization** since it appears to help instructors consider topics and themes one at a time rather than context switch between them or conflate issues among them.
- **Thematic summarization should be considered complementary** to tools that allow full browsing and perhaps broad summarization. Instructors need a broad view in order to select

and refine a reasonable theme, as well as to browse and check that they have not built biases into the theme they have selected.

- **Thematic summaries should capture a spectrum of responses**, not a single polar opinion. For example, it appears to be more useful and reliable to summarize students' statements about the spectrum from ease and difficulty together, rather than separately trying to summarize statements that a task is easy from ones that it is difficult.
- **Original reflections behind a summary should be easily accessible** so that instructors can satisfy their curiosity about the reasons and stories behind the statements students make.
- **Unclassifiable responses should also be made visible in some way.** Some TAs said that they sought out unique explanations by individual students that described particular issues, not widely encountered, but worth fixing in course materials. These are easily missed by topic-matching or clustering techniques.

Other possible improvements for future enhancement of the tool include comparing student reflections across multiple semesters as projects evolve and listing the most common negative comments about a tool or service.

6.2 Limitations

There are some limitations to our approach and study. First, our keyword-based method is ambiguous with respect to negation. For example, suppose an instructor defines the theme "Difficult" to gather reflections saying that project tasks associated with a certain entity are difficult. Then, if a student writes the entity "was not difficult," this sentence will be classified into the theme "difficult" even though they mean the opposite. Users can mitigate this effect by including both polarities in their themes. Second, the number of participants (eleven) was too small to conduct statistical testing. This is because we targeted a graduate-level course that typically enrolls fewer students than undergraduate-level courses, hence fewer TAs. Finally, our method can generate only what TAs expect because it asks them to define themes by themselves. Even though our system stores unclassified reflections in a CSV file to help them explore new insights, we could not test its usability because its effective use requires them to define multiple themes. This was not possible in 45-minute-long interviews.

7. Conclusion

Our approach to filtering text for summarization with the interactive entity and keyword selection was considered to be more useful than simply summarizing samples of student reflections in the interviews with TAs and therefore seems to be useful even in its current form. Instructors benefit from our system even without any modifications, but future research can improve it by developing ways to help them pick keywords for themes and discover new themes and tweaking the dashboard to reflect concrete problems faced by students more.

Acknowledgements

We would like to thank TAs who participated in our interviews and students in the cloud computing course who agreed to have their reflections analyzed for this research. This work was funded in part by a grant from Microsoft.

References

Baird, J. R., Fensham, P. J., Gunstone, R. F., & White, R. T. (1991). The importance of reflection in improving science teaching and learning. *Journal of research in Science Teaching*, *28*(2), 163-182.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Chinchor, N. A. (1998). *Overview of muc-7/met-2*. SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/n18-2097

Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, *7*(3), 93-99.

Fan, X., Luo, W., Menekse, M., Litman, D., & Wang, J. (2017). Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (pp. 363-374).

Harvey, L. (2003). Student feedback [1]. *Quality in higher education*, *9*(1), 3-20.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2021). spaCy: Industrial-strength Natural Language Processing in Python (Version v3.0.5) [Software]. http://doi.org/10.5281/zenodo.4593273

Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, *4*(3), 187–210. https://doi.org/10.1037/1076-898X.4.3.187

Menekse, M., Stump, G., Krause, S., & Chi, M. (2011). The effectiveness of students' daily reflections on learning in engineering context. In *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--19002

Mosteller, F. (1989). The 'muddiest point in the lecture' as a feedback device. *On Teaching and Learning: The Journal of the Harvard-Danforth Center*, *3*, 10-21.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Weischedel, R., Palmer, M., Marcus, M., Hovy, M., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). OntoNotes Release 5.0 [Dataset]. *Linguistic Data Consortium.* https://doi.org/10.35111/XMHB-2B84

Yao, J. G., Wan, X., & Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems*, *53*(2), 297-336. https://doi.org/10.1007/s10115-017-1042-4

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020, July). Big Bird: Transformers for Longer Sequences. In *NeurIPS*.