Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Do my students understand? Automated identification of doubts from informal reflections

Siaw Ling LO*, Kar Way TAN & Eng Lieh OUH
School of Information Systems, Singapore Management University, Singapore
*sllo@smu.edu.sg

Abstract: Traditionally, teaching is usually one directional where the instructor imparts knowledge and there is minimal interaction between learners and instructor. With the focus on learner-centered pedagogy, it can be a challenge to provide timely and relevant guidance to individual learners according to their levels of understanding. One of the options available is to collect reflections from learners after each lesson to extract relevant feedback so that doubts or questions can be addressed in a timely manner. In this paper, we derived an approach to automate the identification of doubts from students' informal reflections through features analysis, word representation and machine learning. Using reflections as a feedback mechanism and aligning it to the weekly course content can pave the way to a promising approach for learner-centered teaching and personalized learning.

Keywords: Doubt identification, learner-centered pedagogy, text analytics

1. Introduction

One of the main goals of an instructor is to ensure that most students, if not all, are able to understand and articulate the key concepts of each lesson clearly. It is challenging to verify that students understand the class materials in an informal way rather than an assessment. An approach is to have students write a reflection after each lesson. Since the reflection is a free-form text response, it is not uncommon to find combinations of articulation of learning points, questions and statements reflecting doubts related to the topic of the week. For example, students may include the following phrases, "I am still confuse about…", "It would be good if you can go through [a topic] again", "I am quite unsure when [an example] is a sample or a population".

We define doubt as a statement, which can potentially be a question or simply a statement that requires more clarification of a given topic. A doubt can be different from a question since it may not be expressed in the form of 5W1H (who, what, where, when, which, how) or end with a question mark. With the amount of informal free text collected from each class, it is essential to find an automated method to effectively extract questions or doubts so that the key concepts can be clarified in a timely manner.

In this study, we explored the effectiveness of feature analysis, word representation and machine learning approaches (such as text classification and sentiment analysis) in identifying doubts expressed in free-form students' reflection. We collected reflection data from two classes taught by two different instructors. Instructor 1 had two modes of data collection: free-form reflections and a self-assessed level of understanding captured in a Likert scale rating of 1-5 (with 1 being the lowest and 5, the highest). Instructor 2 collected only free-form reflections.

This study consisted of two parts. The first part focused on mining doubts from free-form students' reflections by automatically identifying doubts based on analyzing various text features and their role in constructing a machine learning model. Essentially, this is a binary classification problem, which differentiates reflections that contain and do not contain doubts, through assessment of suitable features and word representations. Since doubts can potentially be a question, question patterns were identified as one group of important features. Another feature was the sentiment of the reflections since sentiment analysis has been used extensively in feedback analysis studies (Dhanalakshmi, Bino, &

Saravanan, 2016; Gottipati, Shankararaman, & Gan, 2017). Thus, it is of interest to investigate if the sentiment of students' reflection plays a role in identifying doubt. The second part of the study evaluated the possibility of using the self-assessed level of understanding Likert scale rating to identify students who may need more help and also to assist in doubt detection.

To evaluate the effectiveness of our model, our selected top models trained using the data of Instructor 1 were applied to the set of reflections collected by Instructor 2. This second data set was collected without the corresponding self-assessed rating. The purpose of this verification was to evaluate the performance of applying our model to another qualitative data set since it is a common practice for instructors to collect qualitative feedback after a lesson or course without an accompanying rating scale. Our results showed that our proposed model can successfully extract doubts from students' reflections and interestingly, most of the contents did not contain the standard question patterns such as question mark or 5W1H. The findings exemplify the need to differentiate doubts from questions. We believe that identification of doubt contributes towards providing relevant feedback in a learner-centered environment that tailors to the needs of each individual learner.

The contributions of our work can be summarised in three folds. Firstly, we propose an automated doubt identification model using neural embedding word representation that shows promising results based on real-world student reflections. Secondly, we observed from our results that the model built with features inclusive of positive sentiment performed the best. Specifically, the model was constructed using reflections annotated with positive sentiment without doubts and reflections labelled as doubt. However, sentiment analysis alone was not sufficient to detect and identify doubts. Finally, our analysis suggests that self-assessed ratings correlate with doubts and the rating can be used in a learner-centered pedagogy to extract reflections and contents that can be useful to instructors.

In the next section, we will discuss some related work in detecting questions and feedback analysis. This is followed by the scope of data, methods used, our findings and results in Sections 3, 4 and 5, respectively. In Section 6, we discuss our observations of the findings and future plans before conclusions are drawn in Section 7.

2. Background and Related Work

2.1 Detecting Questions in Online Content

Question identification/detection serves many purposes but is very challenging for online content. Online questions are usually long and informal, and standard features such as question mark or 5W1H words are likely to be absent. Both rule-based and learning-based are common approaches to address this challenge. Rule-based approach such as the paper proposed by Efron and Winget (2010) designed several rules from heuristics or observations to check whether a tweet is a question or not. Learning-based approach proposed by Cong, Wang, Lin, Song, and Sun (2008) involves sequential patterns-based classification method to detect questions in a forum thread, and a graph-based propagation method to detect answers for questions in the same thread. Another learning-based approach explored question characteristics in community question answering services, and proposed an automated approach to detect question sentences based on lexical and syntactic features (Wang & Chua, 2010; Efron & Winget, 2010). Since this study focused on identifying doubts and its relevant features that can impact the accuracy of a model, the features identified by the rule-based approach proposed by Efron and Winget (2010) was adopted to assess whether common questions lexical and syntactic features (e.g., question mark, 5W1H, question patterns) can be used to identify doubts in reflections.

2.2 Analysis of Student Feedback

Analyzing student feedback can help to improve student's learning experience. A large part of feedback comes in the form of textual comments which pose a challenge in quantifying and deriving insights. Gottipati et al. (2017) have presented a conceptual framework for student feedback analysis that provides the necessary structure for implementing a prototype tool for mining student comments and highlights the method to extract the relevant topics, sentiments and suggestions from student feedback. Shankararaman, Gottipati, Lin, and Gan (2017) have provided an automated solution for analyzing

comments, specifically extracting implicit suggestions which are expressed as wishes or improvements from the students' qualitative feedback. Dhanalakshmi et al. (2016) have explored opinion mining using supervised learning algorithms to find the polarity of the student feedback based on predefined features of teaching and learning. Opinion mining, especially in the aspect of sentiment analysis and polarity study, has been the cornerstone of student feedback analysis and thus, it was of interest to explore if sentiment analysis played an important role in identifying doubts from the weekly reflections. However, we would like to highlight that this study is different from the common end-of-term student feedback analysis. The end-of-term student feedback focuses on providing qualitative analysis on the course delivery or the instructor while this study works on weekly reflections and feedback so that timely clarification of key concepts can be offered to students before introducing new knowledge or concepts.

3. Scope

3.1 Data Description and Collection

The data for this study was the course: Analytics Foundation, at the Singapore Management University. This is an undergraduate level foundation class involving data analytics for students from various disciplines (Business, Accountancy, Social Science, Economics, Law and Computing). The course required students to understand the algorithms underlying machine learning models, tune the parameters and apply the models to various problem contexts. Since the students had varied backgrounds, they encountered different challenges in understanding algorithms and the application of the learning models.

The course was conducted in a seminar-style learning environment with about 45 students in each class, over 3 hours of class engagement per week. Two course instructors collected weekly reflections as part of students' learning journal. Reflections collected by Instructor 1 consisted of two elements while Instructor 2 collected only one part. The details of the reflections data can be found in Table 1. The intent of the reflections was to provide weekly feedback to the instructors on the level of understanding in each class. Reflections were collected across 10 instructional weeks over a 16-week semester.

The reflection data collected under Instructor 1 served as the core data for our study. This dataset consisted of both free-form text and self-assessed rating. The free-form text from Instructor 2 (covering only one week of data) was used as a verification.

Table 7. Details of reflection data collected

	Instructor 1	Instructor 2
Part 1	Free-form text for entering the key learning points from each lesson.	Free-form text for entering constructive feedback to improve the lesson.
Part 2	An objective question for students to self-assess their level of understanding. Five levels of understanding in the form of Likert scale from 1-5 was collected: 1. Disagree – I am totally lost and I need a consultation 2. Somewhat disagree – I am lost (and I am paying attention) but I'll catch up on my own 3. Not applicable- I was not in class for a valid reason 4. Agree – Understand quite a lot 5. Strongly Agree – Understand almost fully	NIL
Number of students	44 [from one class] (Total reflections collected: 375)	89 [from two classes] (Total reflections collected: 71)

Weekly reflections were collected from students after each lesson and students were encouraged to provide constructive free-form feedback comprising specific learning points for that lesson. The students had up to 5 days to complete the reflections. Some students completed the reflection immediately after class while some students preferred to revise the course content before completing the reflections.

4. Automated Doubt Identification Approach

The focus of this research was to establish an approach to provide timely feedback to the students based on their reflections. Our automated doubt identification approach analyzed individual reflections to extract questions and doubts, thus providing a means to make decisions on the course of action for learner-centered learning. For example, a student whose reflection was identified to contain doubt may receive additional guidance on the given topic. A summary of our approach is depicted in Figure 1.

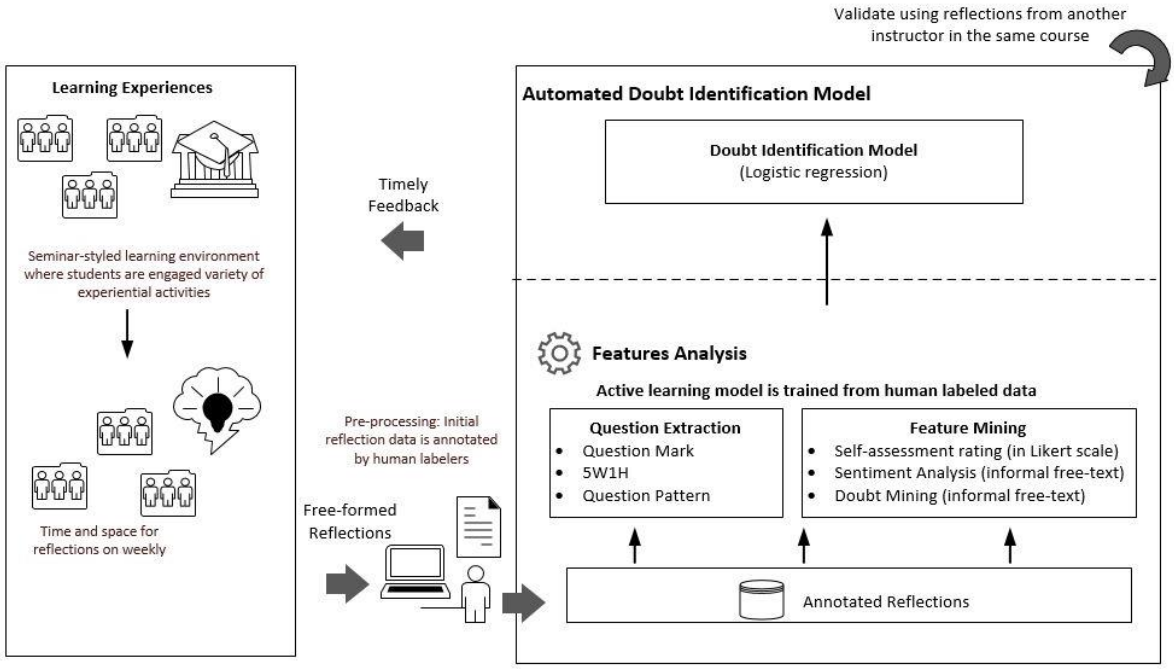

Figure 22 Automated Doubt Identification Approach

The reflections were collected from students in the same course across classes run by two different instructors. To understand if the students learnt from the experiential activities in a seminar-styled learning environment, reflections were collected at the end of each class (during or after the class via a survey). The data from various classes was combined and anonymized before the pre-processing stage. In the pre-processing stage, the initial reflection data was annotated by two human labelers based on presence and absence of doubt and different sentiments. Feature analysis was done on the annotated data, which included two components, namely Question Extraction and Feature Mining. The various features constructed were then assessed and used as the input of a Logistic Regression (LR) model. The LR based automated doubt identification model was validated with a second set of data from Instructor 2 so as to evaluate if the model can be applied to reflections of other students within the same course. Using our automated doubt identification model, instructors can take appropriate decisions and actions, such as timely feedback to help learners who may require more attention and devise additional activities for students who may be more advanced. The details of the various components are described in the following sections.

4.1 Data Preparation and Annotation

Since reflections are usually informal, it was important to clean the data prior to any data analysis. Each reflection was pre-processed to lower case with numbers and stop words removed. However, punctuation removal was selective because question mark was used as a feature in extracting potential questions.

In order to investigate the effect of using sentiment analysis to identify doubts raised by students, the data was annotated in two exercises. The first annotation exercise focused on identifying doubts by labelling via a 'yes' and 'no' label. The second was the annotation of sentiment with 'positive', 'negative' and 'neutral' labels. Since reflections can be an objective expression of the lesson learnt, neutral sentiment was commonly found in the data. In order to ensure the consistency and quality of the annotation, the following questions were derived to assist in determining the Doubt and Sentiment annotations.

For Doubt annotation, a reflection was labelled as 'yes' if one of the following conditions was fulfilled, otherwise, it would be 'no':

- Does the reflection ask for clarification on any of the topics?
- Does the reflection ask for additional information not previously covered in class?

For Sentiment annotation, each reflection was labelled based on the sentiment identified from one of the most relevant questions below. If none of the positive or negative sentiment was found, the reflection was annotated as 'neutral'.

- Does the reflection make remarks about positive/negative teaching (e.g., pace)?
- Does the reflection express any positive/negative feedback about the instructor (e.g., clarity)?
- Does the reflection express any positive/negative sentiment on the topic for the week (e.g., manageable, difficult to grasp)

Two annotators who were familiar with the reflection process annotated individually on the data. Additional review was done on all the records with inconsistent label before finalizing the two sets of annotations. These two sets of annotated data were used as ground-truth datasets in assessing the performance of the doubt identification models.

4.2 Question Features Extraction

Next, we evaluated if the common rule-based lexical and syntactic question features could be used to build an automated doubt identification model. The following features were considered:

i. Question mark (QM)
ii. 5W1H method
iii. Rule-based question patterns (QP) (Efron & Winget, 2010)

The general rule of using QM and 5W1H is to detect questions by finding question marks at the end of sentences and 5W1H at the beginning of the sentence. However, we observed that this rule was not necessarily applicable in this study due to the informal nature of the reflections. Students used free-form expressions and questions which may not adhere to the format of a standard sentence or question structure. Thus, each QM and 5W1H was used as an individual feature regardless of their position in the sentence. On the other hand, the question patterns proposed by Efron & Winget, (2010) were expanded into phrases and each phrase became a feature, e.g., "I try to find" or "need to know". These phrases were generated based on the pattern "(pronoun)* [try, like, need] to [find, know]" where * sign is a wildcard, signaling 0 or more instances and verbs in brackets ([]) are treated as single words.

4.3 Features Mining

The two methods adopted to identify potential features for identifying doubts are student self-assessed rating and sentiment analysis. Since a Likert scale of 1 – 5 (with 1 being the lowest rating and 5 being the highest) were submitted by students together with their reflection to indicate their level of understanding. It was reasonable to assume that doubts are more likely found in reflections of the lower rating than the higher rating. Hence, this rating was extracted to analyze if doubts could be mined from the corresponding rating, with emphasis on the lower rating.

Sentiment analysis is commonly used in feedback analysis to extract feedback that is of value to improve course delivery and student experience. Since sentiments can provide insights to how well a student perceives the learning experience in class, we were interested in assessing if sentiment analysis could be leveraged for doubt identification. In this study, manual annotation of sentiment was done for a detailed analysis of the informal reflection data. In addition, an off-the-shelf sentiment analysis tool, TextBlob1, was used to assess if the sentiment identified by the software could be an identifier for uncovering the underlying doubts in the reflection. In particular, polarity score greater than zero indicated positive sentiment and polarity score lower than zero was considered as negative sentiment. The software was implemented, without adaption to the domain as we were interested in exploring if an off-the-shelf tool could be used to identify doubts.

4.4 Doubt Identification

In order to automate the doubt identification, machine learning algorithms built with various features and word representations were evaluated. Even though there are myriad of learning algorithms available, this study focused on the analysis and comparison of the features rather than the different types of learning algorithms. Logistic Regression (LR) was used in this study since preliminary study on Naïve Bayes did not yield better results from the data.

In LR models, the probability of observing a discrete label, y (1, 0) for a given input feature, x and the purpose of the binary classification is to find a parameter θ, such that $z = \theta T\ x$, where $\theta \in R^d$ and d is the dimensionality of input features x and z is a real number.

Logistic sigmoid function was then applied to z to fit to the range of [0,1] so that it can be interpreted as a probability of predicted output.

Given training data $\{x_i, y_i\}_{i=1}^{n}$, parameter θ can be found by minimizing the following where λ is a regularization parameter:

$$\min_{\theta} \sum_{i=1}^{n} \log(1 + \exp(-y_i\, \theta^T \mathbf{x}_i)) + \lambda \|\theta\|^2$$

The annotated data was split into 70-30 percent for training and testing purposes. Grid search was applied on a three-fold cross validation to find the best parameter of the training data. Result reported was based on the LR model run with the best parameter on the testing data. Two types of word representation methods were adopted. The first being the vector space model using unigram and bigram on both term frequency (TF) and Term frequency-inversed document frequency (TF-IDF) measures. The second was a neural embedding method using doc2vec's distributed bag of words model (Lau & Baldwin, 2016) with a dimension size of 100 and a minimum word frequency of 2. Doc2vec word representation has shown promising performance in various natural language processing (NLP) tasks so it is of interest to assess its performance against the commonly used vector space model in our context.

4.5 Evaluation Metric

Typical accuracy metrics used for statistical analysis of binary classification, which considers the true positive and true negative, have known issues in terms of reflecting the performance of a classifier (Sokolova, Japkowicz, & Szpakowicz, 2006). Therefore, we used F-measure or F1 score as the metric when assessing the performance of the various approaches proposed in addition to the correct assignment or accuracy percentage of positive and negative datasets. F1 score is the harmonic mean of both precision and recall where precision is defined as the ratio of true positive found from the predicted positive while recall is the ratio of true positive identified from the actual positive.

5. Experiments and Results

5.1 Analysis of annotated data

[1] https://textblob.readthedocs.io/en/dev/

Out of the 375 reflections extracted, 295 were annotated as 'no' or not containing doubts while 80 were labelled as 'yes'. On the other hand, 100 were annotated as 'positive' or reflections with positive sentiment; 31 and 244 were labelled as 'negative' and 'neutral' respectively.

With the annotated data of doubt, Figure 2a shows that self-assessed rating can potentially be an indicator in identifying doubts since the percentage of reflections containing doubt decreases with increase in self-assessed rating. However, it does not necessarily imply that higher ratings (e.g., rating 4 and 5) do not contain doubts in the reflection. Similarly, in Figure 2b, the higher ratings have higher percentages of annotated positive sentiments. Conversely, lower ratings have higher percentage of reflections annotated with negative sentiment. Both ratings and sentiment analysis show their potential as features for identifying doubt in reflection statements.

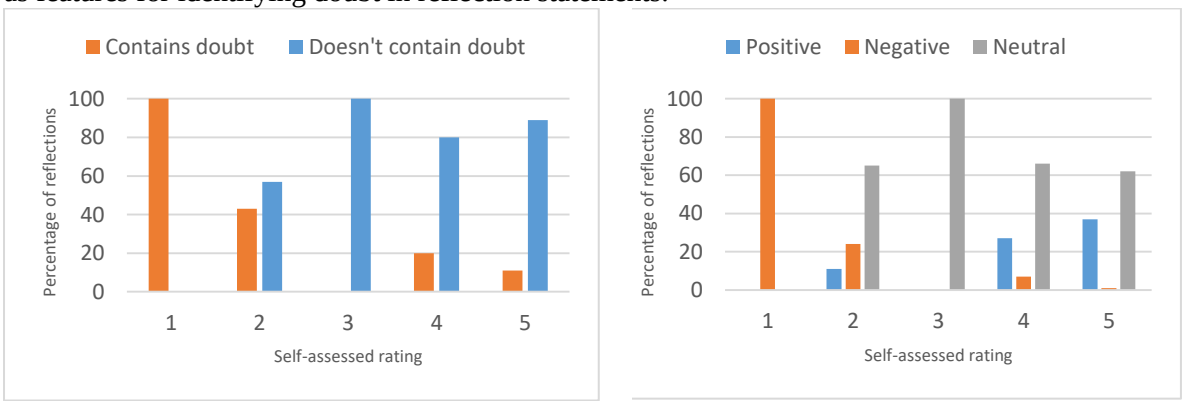

Figure 23. a) Percentage distribution of annotated doubt with respect to self-assessed rating and b) Percentage distribution of annotated sentiment with respect to self-assessed rating

We took further steps to assess if sentiment analysis could solely be used to identify doubts. Figure 3 shows that sentiment results should not be used directly to identify doubts as doubts can also be found in reflections with positive and neutral sentiments. These are two examples of actual reflections which students wrote (informally): (1) "`Clustering interesting leh. HAHAHA with the steps and powerpoint animations, very clear can understand. latent variable (g) and error -- g is unobserverd which is not reflecting in the SAS result or data so where does this infor appear ? but error is produced after the data is being analysed right?`"; and (2) "`I briefly learnt how K-means clustering work and how to interpret the results of a K-means clustering in SAS EG. I felt that I may need a brief revision on SSE.`" The first reflection is an example of a reflection with positive sentiment containing doubt while the second reflection shows a neutral sentiment with doubt.

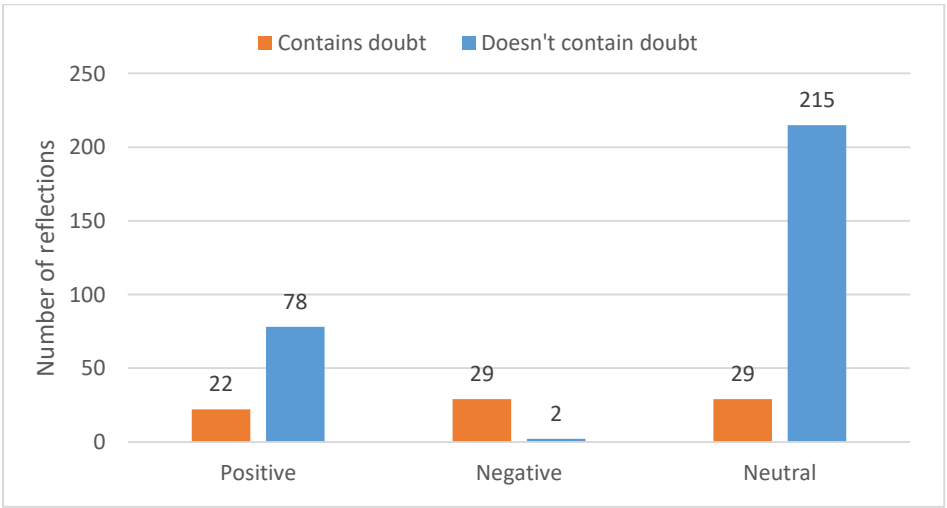

Figure 3. Number of reflections with doubts identified with respect to the three types of sentiments annotated

It is also reasonable to state an observation that reflections with negative sentiment are likely to contain doubts. However, statement of doubt can be found in all types of sentiments (refer to Figure 3). Therefore, it is not sufficient to identify doubts from sentiment analysis. It is essential to treat doubt identification problem separately from sentiment analysis.

5.2 Performance results of various features

With the annotated data, various LR models with different features were constructed for doubt identification. However, in view that the annotated data was an imbalanced dataset (with 295 reflections labelled as 'no' and 80 as 'yes'), which might affect the accuracy of the model, random resampling via replacement of the smaller dataset was performed. The performance of the various features is presented in Table 2. The results (Model 1-4 in Table 2) shows that resampling of data achieved a higher F1 score. In other words, the imbalance sample size affects the performance of the LR model.

Next, we trained our model with a set of question features, that is, QM, 5W1H and QP (refer to Model 5 and 6 in Table 2) and similar results were found for both models. Further analysis was done on the question features and it was found that QM was detected in only 12% of the annotated reflections while 5W1H and QP patterns were found in 59% and 1% of the data respectively. Since QP was not commonly found, it is understandable why the model yielded the same result regardless of whether QP was used as a feature in the model. The question features, in fact was found to be one of the lowest performing features for identifying doubt. In short, reflection statements with questions do not necessarily reflect doubts.

In Model 7, polarity score assigned by TextBlob was used for detecting doubts. Specifically, a reflection with polarity score lower than zero is considered to contain doubt and a reflection with polarity score greater than zero does not. The result in Table 2 shows that the off-the-shelf sentiment analysis tool does not perform well in identifying doubts.

Table 2. Performance metric of LR models using various features

Model	Features*	Precision	Recall	F1 score
1	All data unigram features	0.70	0.29	0.41
2	All data unigram features with resampling	0.61	0.46	0.52
3	All data unigram, bigram features	0.86	0.25	0.39
4	All data unigram and bigram features with resampling	0.64	0.38	0.47
5	QM and 5W1H with resampling	0.29	0.38	0.33
6	QM, 5W1H and QP with resampling	0.29	0.38	0.33
7	TextBlob polarity score	0.24	0.51	0.33
8	Selected data with unigram features	0.70	0.67	0.68
9	Selected data with unigram, bigram features	0.83	0.62	0.71
10	Selected data with doc2vec embedding	0.76	0.75	0.75

*The result of Model 1-9 is based on the TF as the feature vector space. The result from TF-IDF is omitted since it is consistently lower than the above.

In view of the poorer-than-expected performance results of the features used in Model 1-7, it is plausible that none of the features can effectively separate the classes and thus we fine-tuned our feature selection method to select suitable features that can improve the model. Based on Figure 3, it is less likely for a reflection with positive sentiment to contain doubt, so a new dataset consisting of (a) all the annotated 'yes' reflections (80); (b) reflections with positive sentiment and 'no' doubt label (78) were extracted to be the selected training data. This training data was more distinctive in statements with doubt and those without doubts. The results (using unigram and both unigram and bigram as features) are listed as Model 8 and 9 in Table 2. The results show that selected data with unigram and bigram features achieve better F1 score compared to earlier models. One possible reason for the improvement of the result is due to the fact that reflections with positive sentiment contain features that can be used to clearly differentiate the identification of doubts. Indeed, with further analysis of the top features of the LR model, words such as "unsure, lost, confuse, don really" are extracted under the "yes" label (reflections with doubt identified) and "interesting, useful, clearer understanding, better" are found

under the features labelled as "no" (reflections that do not contain doubt). With the encouraging results from Model 8 and 9, doc2vec embedding was constructed using the selected data. The LR model with the embedding representation (Model 10) resulted in a highest F1 score of 0.75 among the tested models. Interestingly, both precision and recall values of Model 10 were of the same range, without being bias to any of the metrics. Therefore, we considered Model 10 as a better model compared to the best-performing vector space model (i.e., Model 9). We attribute the better performance of neural embedding methods to its ability to capture the semantic of words, which can be hard to represent using the vector space model.

5.3 Validation of model against qualitative survey reflection data without self-assessed rating

We validated our model using another qualitative reflection data collected by Instructor 2. This dataset had 71 reflections covering the topic from one lesson and there was no self-assessed rating to act as a reference. The three best models, namely Model 8, 9 and 10, from Table 2 were tasked to extract as many reflections containing doubts as possible. Based on the result stated in Table 3, LR model built using the doc2vec embedding performed the best with F1 score of 0.74. The other LR models (Model 8 and 9) built using TF vector space model resulted in lower F1 score of 0.57 and 0.59 respectively. This implies that neural embedding method is a better representation in identifying doubts from the informal reflections. The following are some of the reflections extracted that contains doubts:

- `Need more help to determine the best fit line.`
- `I think the way p-value works should be revisited and explained more cle arly because it was unclear during the class.`

These reflections do not contain any known question features and hence model built purely with question features will not be able to identify these students requiring further attention.

Table 3. Result on validation data

Model	Features	Precision	Recall	F1 score
8	Selected data with unigram features	0.50	0.65	0.57
9	Selected data with unigram, bigram features	0.53	0.65	0.59
10	Selected data with doc2vec embedding	0.74	0.75	0.74

6. Discussions and Future Work

The LR model built using the selected data of annotated doubt and sentiment has shown a promising result in automating the identification of reflections containing doubt. The main reason that it was performing better than others is predominately the selection of the training data and the importance of identifying suitable features. One of the findings of this study shows that reflections with positive sentiment are less likely to contain doubts (but not the absence of it) and this feature can be used for constructing a better doubt identification model. In view of the importance of detecting positive sentiment, TextBlob was tasked with the evaluation. Out of the 100 reflections with positive sentiment, TextBlob only managed to identify 11 of them as positive. The rest were labelled as negative. With a F1 score of 0.2, it is of concern to use any of the off-the-shelf sentiment analysis tools to aid in identifying doubts. Since the accuracy of a sentiment analysis is heavily dependent on domain knowledge and data, it is essential to construct a sentiment analysis that is adapted to the domain for better accuracy.

We plan to expand the doubt identification study to more classes and courses to improve the model and analyze if it is feasible to identify phrases or doubt embedded patterns that are unique in the expression of informal reflections, for example, "I'm confused", "need more help" etc. Besides that, various other machine learning algorithms such as Support Vector Machine, will be adopted on top of LR to assess the effect of different algorithms in doubt identification.

We noticed that student's expressions from the reflections were more open, casual and in fact truthful. This is partly because the reflections were collected individually within a learning management platform where the statements were read only by the instructors. This behavior may not exhibit if reflections were collected in open discussion forums or formal feedback surveys such as course

evaluation collected by the institution where students may be more conscious about what they share and may be more reserved in expressing doubts or seeking help.

As shown in Table 2 and 3, LR model with doc2vec embedding performed the best based on the testing dataset of Instructor 1 and the qualitative reflection data of Instructor 2. One observation was the balanced values of precision and recall. This is encouraging since it showed that neural embedding is able to learn the intrinsic expression of informal reflections and can be used to improve the accuracy for doubt identification. With the promising results from various deep learning approaches in NLP, our future plan is to assess if pre-trained models coupled with transfer learning (Chronopoulou, Baziotis, & Potamianos, 2019) can be used to automate identification of relevant reflections with higher accuracy. The aim is to develop a generic model that can address doubt identification from other subjects and domains by allowing target task adaptation to the model. Besides that, the ultimate goal is to extract the topics or concepts from the reflections that required more attention. Once the doubts can be extracted, the next step is to perform topic modelling study to identify difficult or commonly misunderstood concepts for further clarification in class. We believe that timely feedback and doubt clarification play important roles in the success of learning.

7. Conclusion

In this study, we have analyzed the nature of informal reflections and investigated various features and word representation methods that can aid in identifying doubts from reflections. Our results show that selecting suitable features are important and reflections with positive sentiment do play a role in constructing a better model. In addition, using neural embedding as the word representation method has shown to achieve the best performance among our data sets. Analysis of the reflection data suggests that the self-assessed level of understanding Likert scale rating can be adopted together with the proposed automated approach to enable learner-centered methodology to improve students' understanding. With the ability to automate doubt identification, topics or concepts that may be difficult or misleading can be extracted more efficiently for providing timely feedback, doubt clarification and improve learning experience.

References

Chronopoulou, A., Baziotis, C., & Potamianos, A. (2019, June). An embarrassingly simple approach for transfer learning from pretrained language models. *Proceedings of NAACL-HLT*, 2089-2095

Cong, G., Wang, L., Lin, C. Y., Song, Y. I., & Sun, Y. (2008, July). Finding question-answer pairs from online forums. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR 08*, 467-474.

Danielsiek, H., Paul, W., & Vahrenhold, J. (2012, February). Detecting and understanding students' misconceptions related to algorithms and data structures. *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education – SIGCSE 12*, 21-26.

Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016, March). Opinion mining from student feedback data using supervised learning algorithms. *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, 1-5.

Efron, M., & Winget, M. (2010, February). Questions are content: A taxonomy of questions in a microblogging environment. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1-10.

Gottipati, S., Shankararaman, V., & Gan, S. (2017, October). A conceptual framework for analyzing students' feedback. *2017 IEEE Frontiers in Education Conference (FIE)*, 1-8.

Lau, J. H. & Baldwin, T. (2016, August). An empirical evaluation of doc2vec with practical insights into document embedding generation. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78-86

Shankararaman, V., Gottipati, S., Lin, J. R. & Gan, S. (2017, December). Extracting implicit suggestions from students' comments – A text analytics approach. *Proceedings of 25th International Conference on Computers in Education.* 261-269.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence*, 1015–1021.

Wang, K., & Chua, T. S. (2010, August). Exploiting salient patterns for question detection and question retrieval in community-based question answering. *Proceedings of the 23rd International Conference on Computational Linguistics*, 1155-1163.