Research and Practice in
Technology Enhanced Learning

RESEARCH Open Access

Automated doubt identification from informal reflections through hybrid sentic patterns and machine learning approach

Siaw Ling Lo[*], Kar Way Tan and Eng Lieh Ouh

[*] Correspondence: sllo@smu.edu.sg
School of Information Systems,
Singapore Management University,
Singapore, Singapore

Abstract

Do my students understand? The question that lingers in every instructor's mind after each lesson. With the focus on learner-centered pedagogy, is it feasible 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 a hybrid approach that leverages a novel Doubt Sentic Pattern Detection (SPD) algorithm and a machine learning model to automate the identification of doubts from students' informal reflections. The encouraging results clearly show that the hybrid approach has the potential to be adopted in the real-world doubt detection. Using reflections as a feedback mechanism and automated doubt detection can pave the way to a promising approach for learner-centered teaching and personalized learning.

Keywords: Doubt identification, Sentic computing, Learner-centered pedagogy, Text analytics

Introduction

Traditionally, teaching is usually one directional where the instructor imparts knowledge with minimal interaction between learners and instructors. As a result, it is challenging to identify topics or concepts that may need more clarification after each lesson. Even though the main goal 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, there is no straightforward way to verify that students understand the class materials except through an assessment. Since an assessment is prepared by the instructor, there is an understanding gap between the instructor who has some presumptions of how a topic is to be understood and what the student actually understand. One possible solution is to have students write a reflection after each lesson so that any doubt and misconception can be detected in a timely manner. It is found that reflection is particularly important in science and technology education since students' initial understanding may not be in accordance with the scientific explanations (Kori, Pedaste, Leijen, & Mäeots, 2014). In other words, reflections have the potential to be used as a tool to assess if students are able to connect their knowledge with the new concepts.

Most studies on reflections focus on how reflections can be used as a learning tool for students (Kori et al., 2014; Veine et al., 2020). This study aims to develop an approach to identify important details, such as doubts, so that reflections are not merely a learning tool but also an informal assessment tool for instructors. The reflections were collected after every lesson and students were encouraged to share their learning experience in 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”.

In this paper, 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. This enables two educational strategies: (1) an agile curriculum adjustment strategy and (2) an adaptive and personalized learning strategy. Sharp and Lang (2018) proposed a conceptual framework for integrating agile principles in teaching and learning. Besides using agile pedagogy for teaching agile methodology, the study presented additional pedagogical intervention such as active learning and reflection journals to improve students’ learning experience. According to the authors, “given that instructors face large amounts of uncertainty regarding the needs and capabilities of the students prior to or at the beginning of a course, it appears that agile principles may be useful in the course development process”. In this study, we use informal weekly reflections to agilely adapt the course content as part of the first proposed strategy. Another related work, Ozdemir, Opseth, and Taylor (2019) proposed to use a learning analytics system that aligned course objectives with assessments to help students reflect on their areas for improvement. The study showed the potential of leveraging the insights from reflections to improve and enhance on the curriculum, be it delivery, assessment design, or the depth of content. We have leveraged on doubt(s) identified in the weekly reflections to design an assessment and learning tool that provides personalized learning to each student. We believe that both strategies, agile and adaptable course delivery combined with the provision of personalized learning have the potential to provide timely and relevant guidance to students, hence enhancing their learning experience.

The first strategy of agile curriculum adjustment gives instructors an option to adapt teaching materials dynamically based on students’ reflections after each lesson. With the doubt identified, a list of topics that require further explanation or clarification can be extracted. Additional materials can be designed to cater to the students, addressing the misconceptions before the start of the next lesson. In other words, by knowing the topics with doubts, instructors can adjust the teaching materials with agility to help students understand the intended learning objectives. This allows timely intervention to the learning and understanding of the students. One implementation method is through interactive questionnaire tools, such as Kahoot, Wooclap, or any audienceengagement software. Content or topics that students have expressed doubts can be formulated into questions, for students to assess where they stand in terms of understanding the content, and for instructors to assess students' understanding. Based on the results obtained, the instructor can clarify the gap in concepts and explain the differences. The impact is two-fold. Firstly, with targeted explanation on the concepts of the previous lesson, currently enrolled students can better understand the new topics (that is usually built on top of the previous lesson). Secondly, suitable content and new materials can be developed to supplement the current curriculum for continuous improvement, benefitting future cohorts of students.

The second strategy is to develop an adaptive tool to aid in students' learning and understanding. In view that it is challenging to provide tailored guidance to individual students after each lesson, the proposed automated doubt identification approach can be used to design an adaptive and personalized learning tool. Multiple-choice questions or true/false questions that are aligned with each topic can be prepared and served to the students according to the doubt(s) detected in the reflection. This is not merely an assessment tool with questions to assess students' understanding but an adaptive tool that is able to personalize and address the needs of the individual student.

An earlier study (Lo, Tan, & Ouh, 2019) on doubt identification focused on evaluating various features such as question features, the role of sentiment analysis, and the different type of word representations including vector space model using unigram and bigram on both term frequency (TF) and term frequency–inversed document frequency (TF-IDF) measures. The results showed that selecting suitable features are important, and reflections with positive sentiment do play a role in constructing a better machine learning model. In addition, using neural embedding as the word representation method has shown to achieve the best performance among our datasets. In this study, we have done an extensive study using a much larger testing dataset—708 records instead of the earlier 71 records. Considering that it is likely for instructors to collect free-form reflections through different questionnaire survey formats, two common formats were included. The purpose is to implement a working system that is able to generalize two different formats of questionnaire survey. In addition, due to the small training data, neural embedding word representation and pre-trained models were adopted in this study. Since the pre-trained models are trained in massive corpus such as Wikipedia or news, the resulted embedding can capture word meanings in a statistically robust manner that small-sample data have no access and unable to represent. Various machine learning (ML) models, in particular, deep neural network (DNN) architectures were used in combination with the word embedding. The results of ML models showed that various configurations have consistently lower doubt identification. In view of that, detailed study of doubt sentic patterns was included in this study. Sentic patterns (Poria, Cambria, Winterstein, & Huang, 2014) consider the dependency relation of common-sense reasoning and concepts that are found in natural language. It has been shown that a better understanding of the contextual role of each concept within a sentence can improve polarity detection markedly. Five types of doubt sentic patterns were identified, and a proposed Doubt Sentic Pattern Detection (SPD) algorithm that incorporated the doubt sentic patterns and English polarity sentic patterns was implemented. The use of sentic patterns enables a generalized analysis that is not limited to the subject domain studied since it focuses on conceptual phrase structures that help to detect doubt-containing sentence. The hybrid *"sentic patterns and ML"* approach has shown promising results in the automated doubt identification.

To evaluate the effectiveness of our approach, our selected top ML models were evaluated using data from two instructors. Reflections of Instructor 1 collected in one questionnaire format were used for training before applying to the set of reflections collected by Instructor 2. This second dataset was collected using another questionnaire format. The purpose of this verification was to evaluate the performance of applying our model to another qualitative dataset since it is not uncommon for instructors to collect qualitative feedback after a lesson or course with more than one type of questionnaire formats. Our results showed that our proposed hybrid approach can successfully extract doubts from students' reflections. Interestingly, standard question patterns such as question mark and 5W1H do not work well in doubt identification. The findings exemplify the need to differentiate doubts from questions. This is partly due to the informal nature of reflections, and students are encouraged to express themselves freely and often in spoken language structure rather than following a formal structure. 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 summarized in three folds. Firstly, we propose a set of doubt sentic patterns and a Doubt SPD algorithm that shows promising results based on real-world student reflections. Secondly, we observed from our results that the ML models built using pre-trained word embedding can be used with the small training data to identify reflections with doubt that is not covered by the doubt sentic patterns. Finally, our analysis suggests that a hybrid approach has the potential to be used as an automated doubt identification to extract reflections and contents that can be useful to instructors in a learner-centered pedagogy.

In the next section, we will discuss some related work in detecting questions, sentic patterns, and feedback analysis. This is followed by the scope of data, methods used, and our findings and results in the Scope of data, Automated doubt identification approach, and Experiments and results sections, respectively. In the Discussions and future work section, we discuss our observations of the findings and future plans before conclusions are drawn in the Conclusion section.

Background and related work

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 pattern-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 (Efron & Winget, 2010; Wang & Chua, 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 the common questions' lexical and syntactic features (e.g., question mark, 5W1H, question patterns) can be used to identify doubts in reflections.

Analysis of sentic patterns

While lexicon-based and machine learning approaches or a combination of these approaches have been used for sentiment and polarity analyses, concept-based techniques are gaining popularity due to their ability to detect subtly expressed sentiments (Cambria & Hussain, 2015). In particular, sentic patterns (Poria et al., 2014) have shown that a better understanding of the contextual role of each concept within a sentence can improve polarity detection. However, currently known sentic patterns are mainly English polarity patterns that include handling of English negation and adversative terms. On top of that, several other studies that made use of sentic patterns include a collection of concept disambiguation algorithms implemented in Chinese language (Xia, Li, Cambria, & Hussain, 2014) and a multilingual polarity detection approach (Lo, Cambria, Chiong, & Cornforth, 2016). In view that polarity or sentiment analysis of the reflections might be important for the doubt identification, English polarity sentic patterns were used in this study. To the best of our knowledge, no known doubt sentic patterns have been reported in any literature. Therefore, this study aims to derive a set of novel doubt sentic patterns that leverage polarity sentic patterns for automated doubt identification.

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, Shankararaman, and Gan (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, Bino, and Saravanan (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.

It is of interest to analyze if there is any prior research in the area of doubt identification. Based on the literature review, the closest is misconception analysis. Gusukuma, Bart, Kafura, and Ernst (2018) has proposed a Misconception-Driven Feedback (MDF) to detect mistakes through program analysis that provides evidence for a set of misconceptions defined by the instructor. It is measured via a summative test on multiple choice quizzes and programming problems. While there is a positive impact of the MDF on student learning and performance, it is not applicable to non-programming courses. On the other hand, Danielsiek, Paul, and Vahrenhold (2012) has developed a concept inventory for algorithms and data structures with the aim to detect students' misconceptions related to the subject. Based on expert interviews and the analysis of 400 examinations, they were able to identify several core topics, which are prone to error. Our proposed doubt identification approach focuses on free-text reflections that are not limited to computing or programming subjects. In addition, since it is not required for instructors to define a set of fixed misconceptions or develop a concept inventory, it has the potential to extract topics or concepts that may require more clarifications but are previously unknown by instructors.

Scope of data

Data description

The data used for this study was the course Analytics Foundation, offered at the Singapore Management University. This is an undergraduate foundation level 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.

Data collection

The course was conducted in a seminar-style learning environment with about 45 students in each class, over 3 h of class engagement per week. Two course instructors collected weekly reflections as part of students' learning journal. Reflections collected by Instructor 1 consisted of one free-text question, while the format used by Instructor 2 consisted of two free-text questions. 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 training data for our study. Since there are two free-form questions from Instructor 2, the contents from the two questions were concatenated as one single text for analysis. The data from Instructor 2 was used as the testing data on the various ML models and approaches proposed. It is important to highlight that the questionnaire formats used by the two data are different, and the main reason of not enforcing the same format is because this

Table 1 Details of reflection data collected

	Instructor 1 (training data)	Instructor 2 (testing data)
Free-form questionnaire survey	Consist of one question	Consist of two questions
Question	Question: Reflect upon the learning point that you have learned in class this week.	Question 1: What do you like best in this session? Question 2: Please provide constructive feedback to improve the session.
Number of students	44 [from one class] (Total reflections collected: 375)	89 [from two classes] (Total reflections collected: 708)

study aims to portray the real-world scenario where it is not uncommon to find different instructors having different formats to collect reflections.

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 3 to 5 days (depending on the instructor) 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.

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 Fig. 1.

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 were combined and anonymized before the pre-processing stage. In the pre-processing stage,

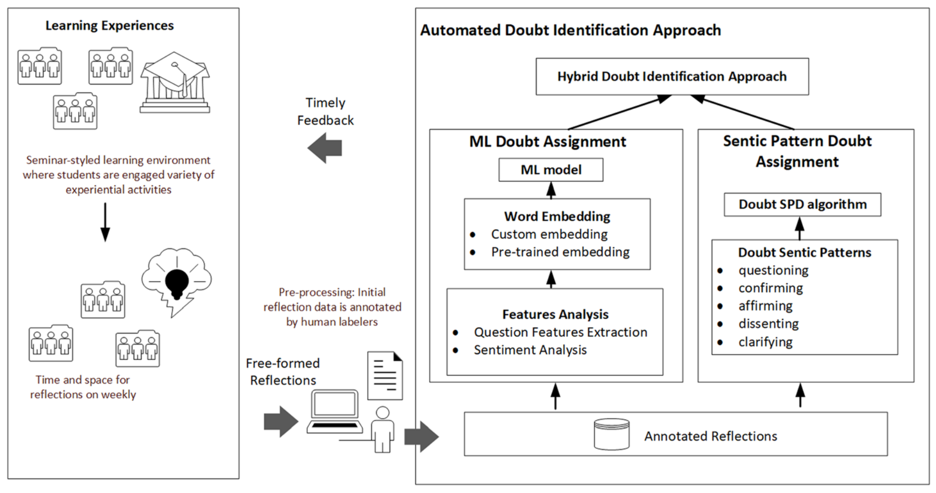

Learning Experiences
Seminar-styled learning environment where students are engaged variety of experiential activities
Time and space for reflections on weekly
Timely Feedback
Pre-processing: Initial reflection data is annotated by human labelers
Free-formed Reflections
Automated Doubt Identification Approach
Hybrid Doubt Identification Approach
ML Doubt Assignment
ML model
Word Embedding
Custom embedding
Pre-trained embedding
Features Analysis
Question Features Extraction
Sentiment Analysis
Sentic Pattern Doubt Assignment
Doubt SPD algorithm
Doubt Sentic Patterns
questioning
confirming
affirming
dissenting
clarifying
Annotated Reflections

Fig 1 Automated doubt identification approach

the initial reflection data was annotated by two human labelers based on presence and absence of doubt and different sentiments. Two doubt assignment approaches were implemented, namely ML Doubt Assignment and Sentic Pattern Doubt Assignment. The ML Doubt Assignment made use of a ML model for assignment. Before a ML model was selected, feature analysis that consists of question feature extraction and sentiment analysis was done on the annotated data. The various features constructed were then assessed before using either of the two types of word-embedding methods to represent the training data. The ML algorithms were trained using the training data to build a model for further evaluation on the testing data. On the other hand, the Sentic Pattern Doubt Assignment leveraged on the different sentic patterns and the Doubt SPD algorithm for doubt assignment. Finally, the Hybrid Doubt Identification Approach considered the results from both types of doubt assignments with the doubt identification of sentic patterns taking precedence over the ML models in the presence of sentic patterns. However, if none of the sentic patterns was found in the reflection, the result from ML Doubt Assignment will be taken as the final result. Using our automated doubt identification approach, 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.

Data preparation and annotation

Since reflections are usually informal, it was important to clean the data prior to any data analysis. Each reflection was preprocessed to lower case with emoticons such as “=)” replaced with a smiley_face placeholder. However, punctuation removal was selective because question mark was used as a feature in extracting potential questions. Contraction handling is implemented to address the shortened version of words and syllables such as “i’ll” for “i will”, “isn’t” for “is not”, “can’t” for “can not”, and “cause” for “because”.

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 ‘y’ and ‘n’ 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 ‘y’ if one of the following conditions was fulfilled, otherwise, it would be ‘n’:

- 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 sentiments were 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 on the data individually. The inter-annotator agreement rates are 0.87 and 0.72 (based on Cohen's kappa calculation) for training and testing data, respectively. McHugh (2012) suggested that Cohen's kappa coefficient value between 0.60 to 0.79 can be interpreted as having moderate level of agreement while value of 0.80–0.90 indicates a strong level of agreement between the annotators. The difference in inter-annotator agreement rate is likely attributed to the different questionnaire formats with testing data having a more complicated two-question format and hence incurring different interpretations from different annotators. The final agreed annotated dataset was curated through the review from one of the instructors of the course and used as the ground-truth dataset in assessing the performance of the doubt identification models.

Machine Learning (ML) Doubt Assignment

There are three main components under ML doubt assignment, namely feature analysis, word embedding, and creation of the ML model. In order to build a ML model to identify doubts from students' reflections, it is necessary to extract relevant features from the annotated data. In particular, question features and reflection s sentiments were analyzed in this study. Since a doubt is considered as needing to seek clarification, it is likely that students may ask specific questions in the reflection, and the reflection content may consist of negative sentiment. The detailed question feature analysis and sentiment analysis can be found in the subsections below. Based on the earlier study (Lo et al., 2019), word embedding using neural model such as doc2vec has achieved a much better performance compared with the traditional word presentation methods of vector space model such as TF or TF-IDF. Two main word-embedding methods are implemented and analyzed to assess if embedding can help to address the small data issue. We also evaluated multiple ML models, and detailed constructions were covered in the Machine learning models section.

Question feature extraction

For the question feature analysis, we evaluated if the common rule-based lexical and syntactic question features were important to build an automated doubt identification approach. 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, it is 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 positions in the sentence. On the other hand, the question patterns proposed by Efron and 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 zero or more instances and verbs in brackets ([]) are treated as single words.

Sentiment analysis

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, TextBlob[1], 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 adaptation to the domain as we were interested in exploring if an off-the-shelf tool could be used to identify doubts.

Word embedding

In view of the limited annotated data available for analysis, both custom-made embedding and pre-trained embedding were studied. The first is using the annotated data to create an embedding from scratch, which includes identifying suitable hyper-parameters and creating an embedding that can be used to represent the data. The second is essentially a transfer learning approach that makes use of a pre-trained embedding to represent the training data.

Doc2vec (Le & Mikolov, 2014) is a custom embedding method, and since it has shown promising results in an earlier study (Lo et al., 2019), it was adopted in this study. Distributed bag of words model is used as the training algorithm with minimum word frequency of two. Embedding vector sizes of 50, 100, 150, 200, and 250 with an epoch size of 30 were used for each training.

Various pre-trained embedding models that are trained using large corpus such as Wikipedia or news corpus have been released for many NLP downstream tasks. Two of the popular pre-trained models are GloVe (Pennington, Socher, & Manning, 2014) and word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). GloVe, which stands for "Global Vectors for word representation" is trained on Wikipedia 2014 and consists of 1.9 million vocabularies. Various embedding dimensions of 50, 100, 200, and 300 are available. On the other hand, word2vec pre-trained model includes word vectors of 3 million words and phrases that are trained on roughly 100 billion words from a Google

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

news dataset and stored using a 300-dimension embedding. Both models were used in this study as a mean to overcome the small training data and make use of the trained representation to better understand the context of the reflections.

Machine learning models

In order to automate the doubt identification, ML algorithms built with various word embeddings were evaluated. Logistic Regression (LR), which was used in the previous study has achieved one of the best results with doc2vec embedding. In this study, three DNN architectures are used. They are Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and recurrent convolutional network (CNN-LSTM).

LR is a binary classification algorithm that learns the relationship of the input features and predicts the probability of an output. The logistic sigmoid function is then used to fit the predicted output to the range of [0,1] for binary classification. Since LR is a supervised learning algorithm, the annotated data was split into 70-30% for training and testing purposes. Grid search was applied on a three-fold cross validation to find the best parameter for 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 TF and TF-IDF measures. The second was a word-embedding method using doc2vec's distributed bag of words model. The details of doc2vec have been described in the Word Embedding section.

The DNN architectures were implemented using the Keras library (https://keras.io/). Specifically, CNN is an implementation of one layer of Keras one-dimensional (1D) convolutional network (convnet) with ReLu as the activation function followed by 1D max pooling and a hidden layer; LSTM consists of two layers of Keras LSTM and CNN-LSTM is a sequential model of a 1D convnet followed by 1D max pooling and a layer of LSTM. Simple 1D architectures were used because of the small data size. In short, the input is the word embedding layer while the output layer is a Sigmoid layer with binary cross-entropy as loss function and ADAM as the optimizer algorithm. The hidden layers are the corresponding different flavors of neural network architectures, for example, the CNN.

Similar to LR, the annotated data was split to 70-30% for training the neural networks. However, due to the many hyper-parameters involved such as learning rate, number of filters and kernel size for CNN, and number of LSTM units, a fine tuner[2] with different ranges of hyper-parameters were engaged to every setup of word embedding and ML model. Specifically, learning rate = {1e-2, 1e-3, 1e-4}; number of filters = (min = 8, max = 64, step = 8); kernel size = (min = 3, max = 7, step = 2) and number of LTSM units = (min = 30, max = 70, step = 10). Early stopping mechanism with minimum validation loss as the criterion was in place to find the best performing model from all the epoch runs. The testing result was obtained from the average of 10 runs using the best performing model.

[2] https://github.com/keras-team/keras-tuner

Sentic Pattern Doubt Assignment

The proposed novel doubt sentic patterns were extracted based on expert knowledge and in-depth analysis of the doubt-labelled training data. In view that doubt is a statement, which can potentially be a question or simply a statement that requires more clarification of a given topic, statements containing question-like queries are the main targets. In addition to the question features stated in the Question feature extraction section, terms with doubt and clarifying bearing words, and informal phrases such as, "can I ask ..." were included in the sentic pattern discovery. The extracted sentic patterns were incorporated into the Doubt SPD algorithm that took in a reflection and indicated if the content contains any doubt sentic pattern in conjunction with polarity sentic patterns. Reflection with doubt sentic patterns identified is highly likely to contain one or more statements with doubt.

Doubt sentic patterns

Since the content in reflections can be casual, question-like queries may not contain the typical question structure that starts with 5W1H and ends with a question mark. However, in order to discover and extract relevant doubt sentic patterns, 5W1H was used as seeds to extract question-like queries. This first sentic pattern is named as 'questioning' which contains phrases like "why do [pronoun]" and "can [pronoun] [say|use|ask]" where [pronoun] is usually {'i', 'we'} and [say|use|ask] represents one of the actions used. Some of the examples are "why do we need to find expected confidence" and "can I ask what is y2?"

The second sentic pattern is considered as confirming a concept or topic. The 'confirming' sentic pattern includes phrases like "does—mean", "what—differ" where – is an arbitrary number of words with a maximum of three words. In other words, 'confirming' sentic pattern is extracting statements with the identified terms (of the phrases) in close adjacent. For example, "if the assumption has been violated, does that mean the model is useless?", "how similar it is to clustering and what is the different or they are the same"

The next group of sentic patterns are terms instead of phrases. Three types of doubt sentic patterns were found and they are 'affirming', 'clarifying', and 'dissenting'. Affirming sentic patterns, as the name suggests, is identifying doubt-related terms or expressions such as "confused", "lost", and "unsure". Clarifying sentic patterns indicate that there are topics or content that need more explanations or clarifications, so terms like "explain" and "clarify" were included. The last sentic pattern is 'dissenting', and it is essentially the opposite of having doubt and hence we have included terms such as "understand" and "know". Besides identifying the candidate terms from reflections, the list was expanded using thesaurus to extract relevant synonyms. The list of words and phrases associated with each sentic pattern can be found in Table 2. Stemmed words using snowball stemmer is kept as the pattern to address the various forms of the root word. For example, "confus" is a stemmed word which originated from confuse, confused, and confusing.

English sentic patterns

It is not sufficient to analyze only the doubt sentic patterns for doubt identification from the informal reflections. The main reason is because there are polarity-reversing

Table 2 Doubt sentic patterns type and its associated sentic patterns

Doubt sentic pattern type	Sentic patterns*
Questioning	Why [do|does] [pronoun], how to, how [do|does] [pronoun], how [pronoun] [say|use|ask], [can|could] [pronoun] [say|use|ask], [pronoun] thought, is there anyway else, [will|would] it be right to [say|use|ask]
Confirming	Does—mean, will—mean, how—differ, what—differ, is—same
Affirming	Confus (confuse), lost, troubl (trouble), unclear, complex, unsur (unsure), wrong, doubt, cheem^
Clarifying	Explan (explanation), explain, refresh, recap, wonder, clarifi (clarify), curious, revis (revision)
Dissenting	Sure, understand, understood, abl (able), know

*Stemmed word is used, and word in () represents the intended candidate term
^Localized expression denoting "complex"

rules such as negation and adversative terms that need to be considered in the English language (Lo et al., 2016). English negation terms such as "not", "couldn't", "shouldn't" should be extracted and hence the contraction handling described in the Data Preparation and Annotation section was implemented. In short, if a polarity term was found after a negation, the polarity of the term was reversed. Besides negation handling, reflections containing adversative terms such as "but" were further processed to ensure the correct polarity was assigned. Specifically, if an adversative term was detected, the reflection was separated into two parts based on the adversative term. Only the polarity of the part following the adversative term was considered. For example, considering the following reflection:

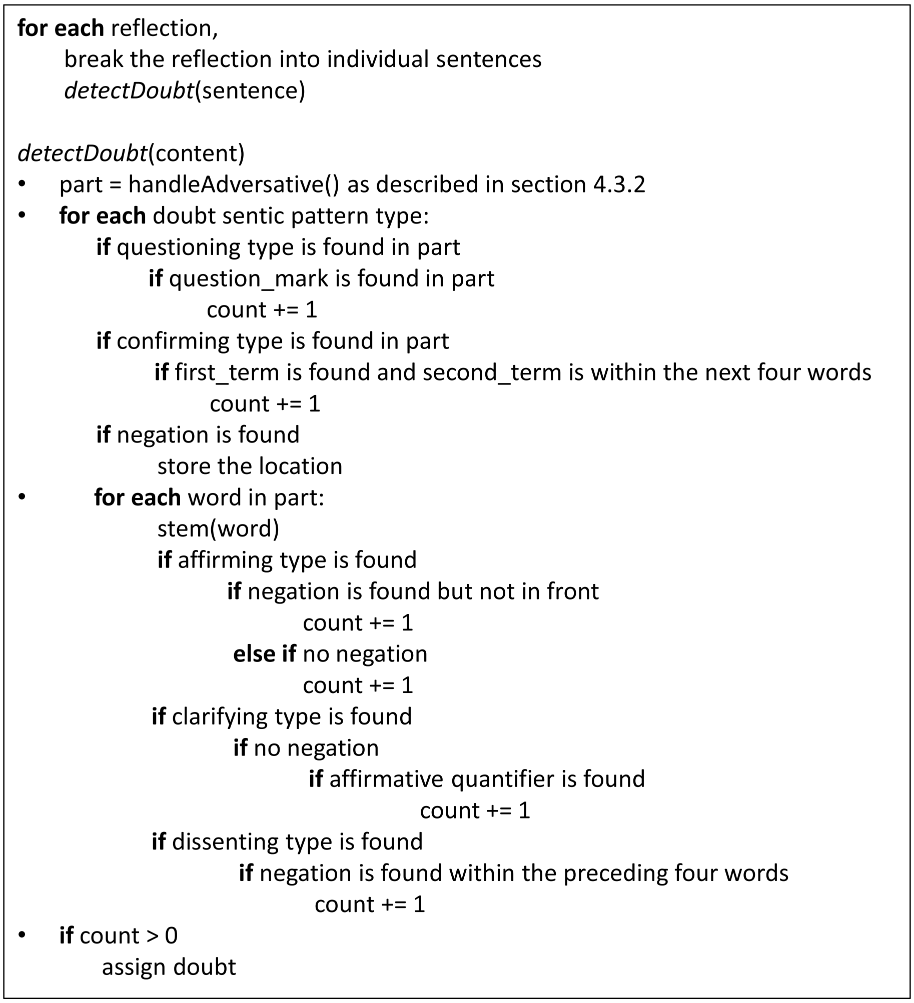
> "week 4 session is overall very great but I might need some recaps about cluster mean and sse and I do not really understand the part about initial seeds... "

In this example, the correct polarity is detected at the part following "but" or the second part of the reflection, which is "I might need some recaps about cluster mean and sse and I do not really understand the part about initial seeds", and thus the reflection is assigned as containing a doubt statement.

Doubt Sentic Pattern Detection (Doubt SPD) algorithm

In order to leverage the various sentic patterns derived in this study, a Doubt SPD algorithm was implemented to integrate the patterns for doubt assignment. Figure 2 shows the details of the algorithm.

The Doubt SPD algorithm takes into consideration the context of different doubt sentic patterns and incorporates the English sentic pattern to ensure doubt-containing statements can be identified. In particular, the algorithm checks for the negation position and assess each sentic pattern term with respect to its type. For example, even if a negation is found but if it is not in the preceding words of an 'affirming' sentic pattern, sentence with affirming type of sentic pattern can still contain doubt. For example, "We were quite lost and do not know what to do with the data given." or "this is confusing as many of us have not used it before". In contrast, a clarifying sentic pattern is sensitive to negation so the algorithm will only consider assigning doubt when no negation is found. In addition, the algorithm also assesses if an affirmative quantifier such



```
for each reflection,
    break the reflection into individual sentences
    detectDoubt(sentence)

detectDoubt(content)
•   part = handleAdversative() as described in section 4.3.2
•   for each doubt sentic pattern type:
        if questioning type is found in part
            if question_mark is found in part
                count += 1
        if confirming type is found in part
            if first_term is found and second_term is within the next four words
                count += 1
        if negation is found
            store the location
•   for each word in part:
        stem(word)
        if affirming type is found
            if negation is found but not in front
                count += 1
            else if no negation
                count += 1
        if clarifying type is found
            if no negation
                if affirmative quantifier is found
                    count += 1
        if dissenting type is found
            if negation is found within the preceding four words
                count += 1
•   if count > 0
        assign doubt
```

Fig. 2 The Doubt SPD algorithm

as "more" is detected before a doubt assignment. For example, "please provide more explanation" or "it will help if prof can explain more". On the other hand, 'dissenting' sentic patterns (e.g., understand, know) requires negation to be found in the preceding words before the statement can be assigned as containing doubt. One example is, "I do not fully understand how the equation helps the analysis".

Hybrid Doubt Identification Approach

Since the Doubt SPD algorithm relies very much on the sentic patterns, and due to the limited resources available (e.g., lexicon) for doubt detection, it is possible that a reflection would not be assigned as containing doubt. In such a situation, the ML Doubt Assignment approach described in the Machine Learning Doubt Assignment section will be used to complement the Sentic Pattern Doubt Assignment. This hybrid approach is able to leverage the strength of the knowledge-based doubt assignment via sentic pattern detection and tap on the classification

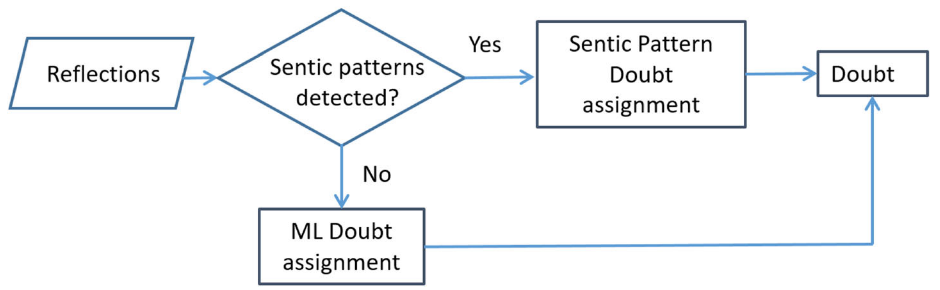


Fig. 3 Hybrid doubt identification approach using sentic patterns and ML model

ability of a machine learning algorithm at the same time. The overall approach is depicted in Fig. 3.

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. 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.

In order to analyze the result in more detail, the precision, recall, and F1 score for both doubt- and without-doubt-labelled datasets are included in this study. Since this is essentially a doubt classification problem, the doubt- and without-doubt-labelled data is denoted as ‘y’ and ‘n’, respectively. Macro F1 score, which is the arithmetic mean of the per-class F1 scores, is used in this paper to summarize the per-class performance. Based on Opitz and Burst (2019), macro F1 is designed with the intention to assign equal weight to frequent and infrequent classes and hence it can be more robust in handling imbalanced dataset.

Experiments and results

Feature analysis and construction of the selected training data

Out of the 375 reflections extracted from training data, 295 were annotated as ‘n’ or not containing doubts while 80 were labelled as ‘y’. 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, two feature analysis were studied, and they are sentiment analysis and question features.

Based on an earlier study (Lo et al., 2019), sentiment analysis has been shown to play a role in identifying doubt in reflection statements. Even though sentiment of the reflections is an important feature, sentiment results should not be used solely to identify doubts as doubts can also be found in reflections with positive and neutral sentiments (Fig. 4). These are two examples of actual reflections that 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

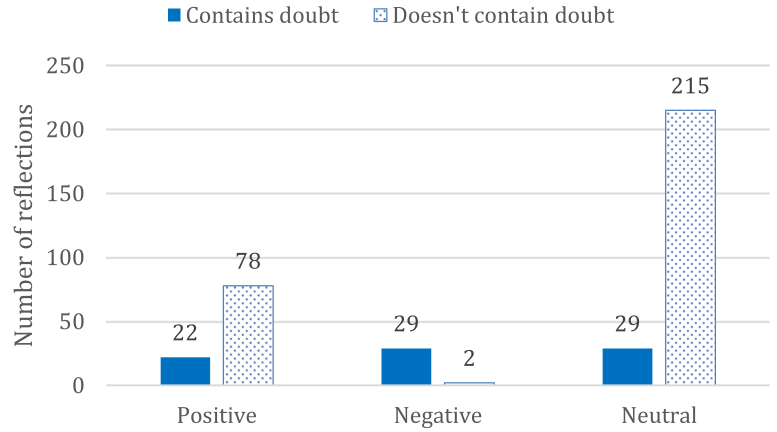


Fig. 4 Number of reflections with doubts identified with respect to the three types of sentiments annotated

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.

It is also reasonable to state an observation that reflections with negative sentiment are likely to contain doubts. However, statements of doubt can be found in all types of sentiments (refer to Fig. 4). Therefore, it is not sufficient to identify doubts from sentiment analysis. It is essential to treat doubt identification problems separately from sentiment analysis.

In order to study the effect of different features, various LR models were constructed for doubt identification. However, in view that the annotated data was an imbalanced dataset (with 295 reflections labelled as 'n' and 80 as 'y'), 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 3.

The results (Models 3–4 in Table 3) show that similar results were found for models trained with question features, that is, QM, 5W1H, and QP. Further analysis was done on the question features, and it was found that QM was detected in only 12% of the

Table 3 Performance metric of LR models using various features

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

*The result of Models 1–7 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 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. Since it is a reflection, it is likely for students to express their thought process via questions; one of the examples found is "whats the purpose of apriori? It is to …". In short, reflection statements with questions do not necessarily reflect doubts.

In Model 5, the 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 3 shows that the off-the-shelf sentiment analysis tool does not perform well in identifying doubts.

In view of the poorer-than-expected performance results of the features used in Models 1–5, 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 Fig. 4, it is less likely for a reflection with positive sentiment to contain doubt, so a new dataset consisting of (a) all the annotated 'y' reflections (80); (b) reflections with positive sentiment and 'n' 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 Models 6 and 7 in Table 2. The results show that selected data with unigram and bigram features achieve better F1 scores compared with earlier models (Models 1–2). One possible reason for the improvement of the result is due to reflections with positive sentiment containing 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 'y' label (reflections with doubt identified) and "interesting, useful, clearer understanding, better" are found under the features labelled as 'n' (reflections that do not contain doubt). With the encouraging results from Models 6 and 7, doc2vec embedding was constructed using the selected data. The LR model with the embedding representation (Model 8) resulted in the highest F1 score of 0.75 among the tested models. Interestingly, both precision and recall values of Model 8 were of the same range, without being biased to any of the metrics. Therefore, we considered Model 8 as a better model compared with the best-performing vector space model (i.e., Model 7). We attribute the better performance of word-embedding methods to its ability to capture the semantic of words, which can be hard to represent using the vector space model.

Results of ML Doubt Assignment

With the better performance shown using word embedding in Table 3, it is of interest to use pre-trained embedding to assess if the model created can achieve a better performance compared with the custom trained doc2vec using selected data. As mentioned earlier, two pre-trained embedding models, GloVe and Word2Vec, were used in this study. Even though pre-trained models are trained using vast corpus, there is no guarantee that all the words found in the training data can be found in the embedding vocabulary. In other words, there is a need to minimize the out-of-vocabulary (OOV) words from our data so that the embedding used can represent the data better. An analysis on calculating the percentage of OOV words was done, and it was found that GloVe has 1.98% OOV, while Word2Vec has 10.54%. As a result, GloVe pre-trained with a 100-dimension embedding was adopted in this study.

In order to decide which ML model to use, four ML algorithms were evaluated with two types of word embedding, which are doc2vec and GloVe embedding. Specifically, eight classifiers were generated: doc2vec_LR, doc2vec_CNN, doc2vec_CNN-LSTM, doc2vec_LSTM, glove_LR, glove_CNN, glove_CNN-LSTM, and glove_LSTM. doc2vec denotes models using doc2vec embedding on the selected data, while glove denotes models using GloVe pre-trained embedding. The overall performance can be found in Table 4. It is observed that GloVe pre-trained embedding using LSTM as a classifier achieved the best performance followed by doc2vec embedding using CNN-LSTM. While LR with doc2vec embedding performed well, it did not do as well using the GloVe pre-trained embedding.

Since this study is about identifying reflection with doubt, the ability of the classifier to correctly extract the relevant reflection is important. All the eight classifiers were evaluated on testing data that consists of 708 records with 537 labelled as ‘n’ or containing no doubt and 171 labelled as ‘y’ or containing doubt. The detailed analysis of both ‘n’ and ‘y’ testing data can be found in Fig. 5. It can be clearly observed that all classifiers performed well on the ‘n’ dataset but consistently did poorly on the ‘y’ data. The reference line on Fig. 5 is the lowest F1 score of 0.29. The best-performing classifier is doc2vec_CNNLSTM-y with a F1 score of 0.43.

Analysis of doubt sentic pattern

As mentioned earlier, a sentic pattern is known to help in understanding the contextual role of a natural language concept. In view that Doubt ML models did not perform as well on the ‘y’ testing data or the data containing doubt, this study aims to analyze sentic patterns that are specific to doubt identification. Five different sentic patterns were extracted, and each played a different role in identifying doubt from reflections. In order to evaluate the impact of each of the sentic patterns, a distribution study (refer to Table 5) has been done on both training and testing data (a total of 1083 records with 832 labelled as ‘n’ and 251 labelled as ‘y’). ‘questioning’ sentic pattern is the most differentiating pattern which is mostly found in ‘y’ labelled data only. The rest of sentic patterns are also found mainly in ‘y’ labelled with ‘clarifying’ sentic pattern being found

Table 4 Overall performance of various ML models using different word embedding on training data (*P, R, F denotes Precision, Recall, and F1 score, respectively*)

	LR			**CNN**			**CNN-LSTM**			**LSTM**		
Word Embedding	P	R	F	P	R	F	P	R	F	P	R	F
Selected data with doc2vec embedding	0.76	0.75	0.75	0.62	0.61	0.61	**0.8**	**0.78**	**0.78**	0.69	0.64	0.61
Selected data with glove pre-trained word embedding	0.46	0.46	0.46	0.60	0.60	0.59	0.74	0.73	0.73	**0.83**	**0.82**	**0.82**

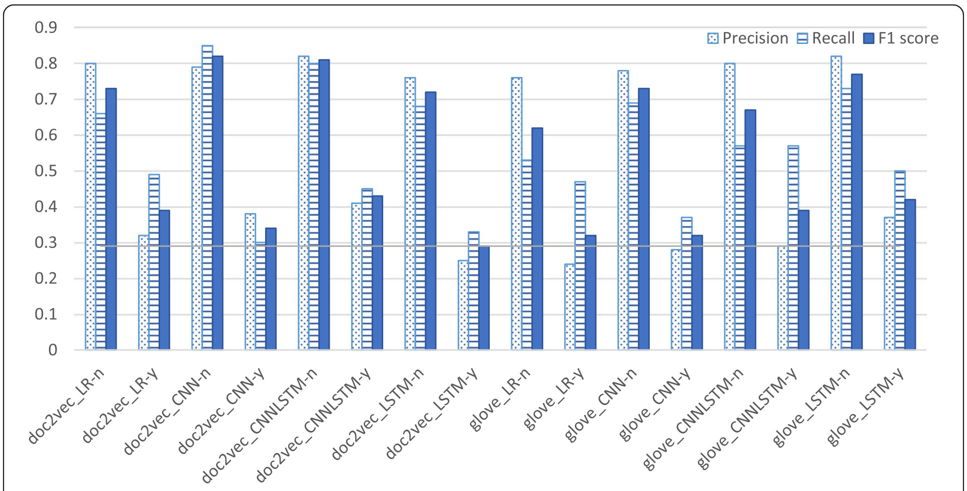


Fig. 5 Performance of various ML models on 'n' and 'y' testing data

the most in the 'n'-labelled data. This is not too surprising since it is common for students to reflect on the lesson learnt, for example, "video chosen was great to explain concept", "I like the revision at the start of the lesson".

With the understanding of the distribution of sentic patterns on the data, it is of interest to assess the performance of the Doubt SPD algorithm described in the Doubt sentic pattern section on the testing data. In order to ascertain the impact of each sentic pattern, various experiments were conducted, and results can be found in Table 6. As observed from the table, 'dissenting' and 'confirming' sentic patterns both performed well. However, the performance metrics on the 'y' testing data is not as good with the best performance having a F1 score of 0.34. A set of assessments using the combination of sentic patterns was done for pattern #s 6–10. Pattern 6 uses the top two scoring patterns of 'affirming' and 'dissenting', while patterns 7 and 8 use the top three scoring with 'questioning' included in 7 and 'clarifying' in 8, respectively. Even though both 'questioning' and 'clarifying' patterns have the same score individually, 'questioning' pattern achieves a higher F1 score and hence may imply it plays a more important role. Pattern 9 shows the results of the top four scoring patterns, and Pattern 10 makes use of all the sentic patterns, and it has shown to achieve the best result. In short, all the sentic patterns play a role in identifying doubt from reflections.

Results of hybrid doubt identification approach

With the promising result of sentic patterns with Doubt SPD algorithm, the next experiment is to evaluate if sentic patterns can help to improve the identification of doubt from

Table 5 Doubt sentic pattern distribution on both training and testing data *('n' is data labelled as having no doubt and 'y' as having doubt)*

Pattern	'n' labelled data	'y' labelled data	Difference
Questioning	0.2%	15.1%	75.5
Confirming	3%	11.6%	3.87
Affirming	5%	27.1%	5.42
Clarifying	4%	10.8%	2.7
Dissenting	5%	27.5%	5.5

Table 6 Results of various sentic patterns with performance metrics for both identifying doubt and overall doubt classification (*P, R, F denotes Precision, Recall, and F1 score, respectively with P-y indicating Precision in identifying doubt or label 'y' during testing*)

Pattern #	Sentic pattern + Doubt SPD algorithm	P-y	R-y	F-y	P	R	F
1	Questioning	0.95	0.11	0.20	0.86	0.55	0.54
2	Confirming	0.67	0.08	0.15	0.72	0.53	0.51
3	Affirming	0.54	0.22	0.32	0.67	0.58	0.59
4	Clarifying	0.45	0.15	0.22	0.61	0.54	0.54
5	Dissenting	0.48	0.27	0.34	0.64	0.59	0.60
6	3+5	0.49	0.57	0.52	0.67	0.69	0.68
7	1+3+5	0.54	0.53	0.54	0.70	0.69	0.69
8	3+4+5	0.51	0.47	0.49	0.67	0.66	0.67
9	1+3+4+5	0.51	0.63	0.57	0.69	0.72	0.70
10	1+2+3+4+5	0.51	0.68	**0.58**	0.70	0.74	**0.71**

testing data for the Doubt ML models. The top two performing Doubt ML models from the Results of ML Doubt Assignment section were selected to implement the hybrid approach. As shown in Table 7, the proposed hybrid approach helped to improve the extraction of doubt from reflection with an over 10-point improvement on the F1 score for the 'y' testing data. In other words, Sentic Pattern Doubt Assignment enhanced the identification of reflection with doubt through the analyzing of the sentence structure and context and increased the recall metric markedly for both the 'y' and overall testing data.

A further analysis on the contribution of ML models on the hybrid approach by extracting reflections that were not identified by Sentic Pattern Doubt Assignment shows that this complementary method has its merit since there are some informal expressions with query intention that are not covered by sentic patterns but is recognized by ML models. The following are some examples:

- "perhaps we could go through a little more on which means of analysis is best and is the most commonly used?"
- "are the results still considered invalid because ultimately there is no convergence?"

Even though it seems that QM does play an important part, our analysis on QM and 5W1H in Table 2 showed that it cannot be used effectively as a feature in a classifier. QM should be used in conjunction with other sentic patterns, and from our analysis; it is a key determinant for the 'questioning' sentic pattern. Without QM, 17% of

Table 7 Results of various approaches considering performance for both identifying doubt and overall doubt classification (*P, R, F denotes Precision, Recall, and F1 score, respectively, with P-y indicating Precision in identifying doubt or label 'y' during testing*)

Approach	P-y	R-y	F-y	P	R	F
doc2vec_CNN-LSTM	0.41	0.45	0.43	0.62	0.62	0.62
glove_LSTM	0.37	0.50	0.42	0.59	0.61	0.60
Hybrid (doc2vec_CNN-LSTM + Sentic pattern)	0.42	0.75	**0.54**	0.66	0.71	0.65
Hybrid (glove_LSTM + Sentic pattern)	0.40	0.82	**0.54**	0.66	0.71	0.63

reflections that do not have doubt contain the ‘questioning’ sentic pattern instead of just 0.2% as reported in Table 4. In other words, ‘questioning’ sentic pattern would not be so specific without the presence of QM in the same sentence.

Discussions and future work

Limitations and promises of the proposed approach

The proposed novel doubt sentic patterns have shown promising results in identifying doubt from reflection. The different questionnaire survey formats did not impact the Sentic Pattern Doubt Assignment as much as the ML Doubt Assignment. For example, LR model, which performed well in the earlier study, was not able to do as well in the larger and different testing data (Fig. 5). On the other hand, when the sentic pattern with the Doubt SPD algorithm was used to classify both the training and testing data, the overall F1 score is 0.73, which is better than the results reporting using just the testing data alone. Even though this study used the subject content from the analytics domain, sentic pattern and the Doubt SPD algorithm has the potential to generalize on all types of data since it does not require any domain knowledge to be encoded. Changes in questionnaire format also will not affect the predictive capability. We ran a preliminary test on a set of reflections from a course in a different domain—Digital Business. The results showed that the proposed approach is able to identify doubts in a similar manner. Some of the statements from the doubt extraction include, “clearer explanation of how digital differs from…”, “I am not sure about omnichannel strategy”. The statements reflect that doubts were identified despite of the domain. In addition, since the sentic patterns have five different types, the results obtained are explainable and can be used by the instructors to understand the topics or concepts that need further clarification.

Although Doubt ML models did not outperform in this study, it is important to highlight that training data used is very small. In fact, it is just 80 ‘y’- and 78 ‘n’-labelled data. The main reason for not increasing the training data is to explore if word embedding can be used to address the small data challenge that is very common in the real world. From the experiment result, pre-trained word embedding is a very promising approach when only small data is available. There are many new pre-trained embedding approaches that have been released recently, for example BERT (Devlin, Chang, Lee, & Toutanova, 2019), which is a bidirectional language representation that has achieved state-of-the-art results for many NLP tasks. Since this study aims to build a model that is capable of generalizing well on different domains and different questionnaire formats, its focus is on the effect of word embedding on small data and also the effect of doubt sentic pattern. Considering that most deep learning approaches are resource intensive and some even need a special hardware (for example, a GPU) to complete the training, having a knowledge-based approach via sentic patterns can essentially enable a development of a simpler classifier that can help to identify doubt from reflections.

As shown in Tables 2 and 6, doc2vec embedding performed the best. However, it can be challenging to select suitable parameters to build a custom doc2vec embedding model that is robust enough to represent the dataset. It is of interest to note that the results from the pre-trained embedding are comparable with doc2vec, even though there is no custom training done. This is partly because the pre-trained embedding was trained using huge amounts of data, and it is able to encode the text semantic more accurately for the construction of a NLP model. However, in this study, the ML models built using the pre-trained embedding still have many rooms for improvement. One possible reason is because the pre-trained word vectors may not fit the domain-specific text, and the training data is not large enough to get accurate results. Adapting pre-trained word embedding through retrofitting (Faruqui et al., 2015) and fine-tuning the embedding layer of the neural network is one option that can be used to improve the result. In other words, the performance of ML Doubt Assignment can be further improved with techniques such as using a better pre-trained embedding and fine-tuning the embedding layer. We did not focus on achieving the best ML models, instead, we proposed a simpler complementary approach through sentic patterns that can be implemented with a simple algorithm and did not require any special resource for the automated doubt identification.

The hybrid approach as proposed in this study is able to improve the less-than-satisfying result of a particular class. Nevertheless, it may not be suitable for all types of classification problems. The main reason that the hybrid approach helped in the doubt identification is because a comprehensive and obvious set of doubt sentic patterns can be extracted and used successfully to differentiate doubt-containing reflections from the non-doubt-containing reflections. If extracting sentic patterns is not feasible, it may be worthwhile to invest effort in training a ML model. As mentioned in the Results of Hybrid Doubt Identification Approach section, ML models have its merit to extract doubt contents that are not recognizable by the Sentic Pattern Doubt Assignment, and hence using this hybrid approach can be a promising option to identify more and accurate reflections with doubt. In addition, the automated process allows scalability and efficiency for large-scale implementation across other courses.

Future work includes batch learning of new datasets, exploring other text-embedding methods such as sentence embedding or context-sensitive pre-train models such as BERT (Devlin et al., 2019) to build a more robust ML model. This coupled with sentic patterns and Doubt SPD algorithm has the potential to automate the identification of relevant reflections with a 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.

Reflection as a tool for instructors

It was observed 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.

With the automated doubt identification, instructors can now extract topics or concepts from the reflections that required more attention for each student. As mentioned earlier in the Introduction section, new educational strategy such as the agile curriculum adjustment strategy provides an option for instructors to adjust the teaching materials dynamically before each new lesson to address doubts or misconceptions found in reflections. Furthermore, the insights uncovered can play an important role in the learner-centered pedagogy since the instructor is now equipped with the knowledge of which concepts should be explained in more detail and is able to provide a specific example with respect to the content shared in the reflection. Targeted personalized learning is now feasible even in a large group teaching setting. Furthermore, by analyzing reflections and identifying doubts, the instructor can design a selective and more specific formative assessment with the purpose of clarifying doubts and misconceptions for a selected group of students or an individual student.

We are currently implementing an adaptive learning tool that measures the alignment of reflections and learning objectives. The purpose is to create an alignment score on learning objectives based on content shared in reflections. A set of well-aligned learning objectives can imply that students are aware of the importance of the topics taught. However, if they are coupled with a high number of doubts identified, the instructor should be concerned that many students may have difficulties understanding the topics or concepts associated with the learning objectives. The ultimate goal of the system is to develop an adaptive learning mechanism that is capable to serve specific practice questions to the topics related to the doubt identified. This feature can be used to help an individual student dispel his or her misconception or wrong understanding of the topic in an ongoing basis. We believe that timely feedback and doubt clarification play important roles for the students to achieve their learning outcomes.

Conclusion

In this study, we have proposed an automated doubt identification approach that helps to extract useful insights to aid instructors in assessing students' understanding of a new topic or concept taught. In particular, we have analyzed the nature of informal reflections and developed novel doubt sentic patterns with a Doubt SPD algorithm that can identify doubts from reflections. The proposed hybrid approach that leverages both the strength of sentic patterns and ML model is a promising and complementary approach that is able to identify reflections with doubt content. With automated doubt identification, topics or concepts that may be challenging to students can be extracted more efficiently to provide timely feedback, doubt clarification, and improved learning experiences.

Acknowledgements

The authors would like to thank Chong Yun Yu, James Lim Jia Yang, and Aaron Teo Wee Kiat for their annotation efforts.

Authors' contributions

All authors contributed to the research idea and approach. Both SLL and KWT contributed towards data annotation and collection. SLL was the main contributor to data analysis and implementation. All authors contributed towards manuscript writing. All authors read and approved the final manuscript.

Funding

This study was funded by Singapore Management University's Technology-enhanced Learning Grant in 2020.

Availability of data and materials

The research data is restricted to only the research team and its affiliates.

Competing interests

The authors declare that they have no competing interests.

Received: 3 July 2020 Accepted: 28 January 2021
Published online: 08 February 2021

References

Cambria, E., & Hussain, A. (2015). *Sentic computing: A common-sense-based framework for concept-level sentiment analysis*, (vol. 1). Springer International Publishing. https://www.springer.com/gp/book/9783319236537.

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

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186.

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

Efron, M., & Winget, M. (2010). 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.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 1606–1615).

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

Gusukuma, L., Bart, A. C., Kafura, D., & Ernst, J. (2018). Misconception-driven feedback: results from an experimental study. In *Proceedings of 2018 International Computing Education Research Conference*, (pp. 160–168).

Kori, K., Pedaste, M., Leijen, Ä., & Mäeots, M. (2014). Supporting reflection in technology-enhanced learning. *Educational Research Review*, *11*, 45–55.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, (pp. 1188–1196).

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016). A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, *105*, 236–247.

Lo, S. L., Tan, K. W., & Ouh, E. L. (2019). Do my students understand? Automated identification of doubts from informal reflections. In *Proceedings of the 27th International Conference on Computers in Education, I*, (pp. 252–262).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing System*, (pp. 3111–3119).

Opitz, J., & Burst, S. (2019). Macro F1 and Macro F1.arXiv preprint arXiv:1911.03347.

Ozdemir, D., Opseth, H. M., & Taylor, H. (2019). Leveraging learning analytics for student reflection and course evaluation. *Journal of Applied Research in Higher Education*, *12*(1), 27–37.

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*, (pp. 1532–1543).

Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*, 45–63.

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

Sharp, J. H., & Lang, G. (2018). Agile in teaching and learning: Conceptual framework and research agenda. *Journal of Information Systems Education*, *29*(2):45–52.

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

Veine, S., Anderson, M. K., Andersen, N. H., Espenes, T. C., Søyland, T. B., Wallin, P., & Reams, J. (2020). Reflection as a core student learning activity in higher education—insights from nearly two decades of academic development. *International Journal for Academic Development*, *25*(2), 147–161.

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

Xia, Y., Li, X., Cambria, E., & Hussain, A. (2014). A localization toolkit for SenticNet. In *Proceedings of IEEE International Conference on Data Mining Workshops*, (pp. 403–408).

Publisher's Note