10-2015

# Analyzing Educational Comments for Topics and Sentiments: A Text Analytics Approach

Gokran Ila NITIN
*Singapore Management University*, ingokarn.2011@sis.smu.edu.sg

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

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

**DOI:** https://doi.org/10.1109/FIE.2015.7344296

# Analyzing Educational Comments for Topics and Sentiments: A Text Analytics Approach

Gokarn Ila Nitin
School of Information Systems
Singapore Management University
Singapore
ingokarn.2011@sis.smu.edu.sg

Asst. Prof. Gottipati Swapna
School of Information Systems
Singapore Management University
Singapore
swapnag@smu.edu.sg

Prof. Venky Shankararaman
School of Information Systems
Singapore Management University
Singapore
venks@smu.edu.sg

***Abstract*— Universities collect qualitative and quantitative feedback from students upon course completion in order to improve course quality and students' learning experience. Combining program-wide and module-specific questions, universities collect feedback from students on three main aspects of a course namely, teaching style, content, and learning experience. The feedback is collected through both qualitative comments and quantitative scores. Current methods for analyzing the student course evaluations are manual and majorly focus on quantitative feedback and fall short of an in-depth exploration of qualitative feedback. In this paper, we develop student feedback mining system (SFMS) which applies text analytics and opinion mining approach to provide instructors a quantified and exhaustive analysis of the qualitative feedback from students and avail insights on their teaching practices and this in turn will lead to improved student learning.**

***Keywords—Student feedback, education data mining, topics, sentiments, text analytics, clustering.***

## I. INTRODUCTION

Universities employ various formal and informal methods to collect and analyse feedback from students in order to enhance the quality of teaching and learning. Many institutions have implemented evaluation surveys which combine "program-wide" questions and "module-specific" questions that enable comparisons to be made across the institution whilst allowing flexibility for individual modules [4]. These surveys provide valuable feedback that helps course designers towards improving teaching style, course content and assessment design, and overall student learning [2][3]. The feedback must be analysed and interpreted with great care so that action, and ultimately improvement, can result from feedback process [1].

Students provide feedback in two distinct forms namely quantitative (numerical) ratings for questions and qualitative comments related to teaching, content and learning [5]. The teaching component refers to instructors' interaction, delivery style, ability to motivate students, out of class support, etc. The content refers to course details such as concepts, lecture notes, labs, exams, projects, etc. The learning refers to students learning experience such as understanding concepts, developing skills, applying skills acquired, etc. Analysing and evaluating this qualitative data to help us make better sense of student feedback on instruction and curriculum.

Current methods for analysing student course evaluations are manual and majorly focus on the quantitative feedback [17] [18]. More often, an analysis of student feedback falls short of an in-depth exploration of a qualitative feedback [32], thereby limiting instructors to the numerical scores and a human understanding of a sample of the feedback, which abstracts collective sentiments for individual components of courses. The question is to how to help the faculty to better digest such large amounts of comments and discover the gaps in the course delivery.

Going forward, a more useful approach will be to map the students' qualitative feedback in the form of topics and sentiments towards the three major components namely teaching, content and learning. Figure 1 shows the problem setup. The input data is a set of students' comments given for an information systems curriculum undergraduate course, IS304, process modelling and solution blueprinting.

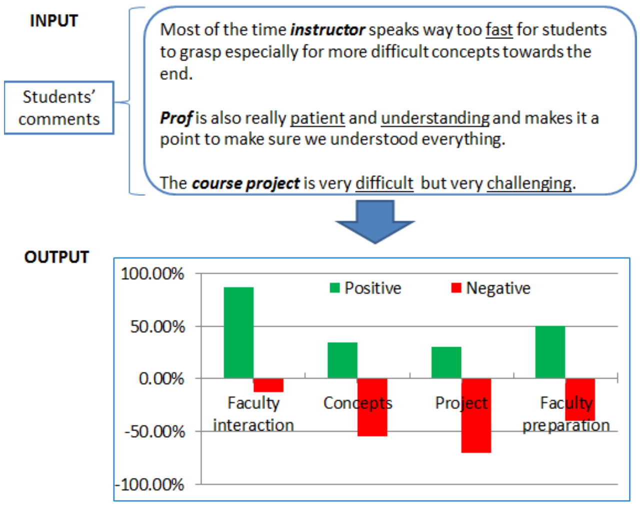


Fig. 1. Sample comments from students for an information systems curriculum undergaduate course. Bolded words are the topics and underlined words are the sentiment words.

With such data, an instructor can only get an overall impression of the course and not the deeper insights. It is infeasible to go over all the comments for deeper analysis. In contrast, an output such as topic based summary on sentiments as shown in Figure 1 provides a detailed analysis. A **topic** refers to an aspect of the course such as concepts, delivery style, understanding, lab, faculty interaction, skills, learning,

Published in Frontiers in Education Conference 2015: Proceedings: El Paso, Texas, October 21-24.
https://doi.org/10.1109/FIE.2015.7344296

etc., and **sentiment** refers to positive or negative experience with the corresponding topic. Figure 1 shows the topics such as "concepts" and "project", and sentiment words such as "patient", "understanding", "challenging", etc. The overall sentiment for *professor* is positive, while the sentiment on *concepts* is negative. Extracting individual topics and sentiments automatically provides instructors and curriculum managers a data-driven approach for improving teaching and learning. Decisions can be made while constructing future cycles of course delivery to maintain or improve components as per feedback and measure the impacts.

In this paper, we provide automated techniques to diagnose textual feedback. The main challenge with this task is the textual nature of comments which are expressed in natural language. Furthermore, the feedback topics and sentiments are embedded within the text. Opinion mining, topic extraction and NLP techniques [8] [15] [16] from the text analytics and linguistics research are widely popular for mining users' comments in social media. Sentiment mining techniques are widely used for product review mining in consumer business world [9] [12]. We leverage these techniques for building the student feedback mining system (SFMS). SFMS applies data mining, text mining and opinion mining techniques on qualitative comments to extract topics and sentiments on courses aiding in generating quantitative visuals to support a deeper analysis.

We evaluated SFMS using student feedback provided by the students for undergraduate core courses taught at the School of Information Systems, Singapore Management University collected for two semesters on seven courses. Information Systems is classified under science and technology education (all engineering courses as classified under this) by Ministry of Education, Singapore. The evaluation is conducted in two phases; quality of the topic extraction and quality of the sentiment extraction. Our experiments show that SFMS system provided meaningful clusters of comments and aspect words for topic extraction task and precision of 80.1% for sentiment extraction task.

The paper will be structured as follows. Section II will review the key background of text analytics techniques and opinion mining problem. Section III will be devoted to literature review and will primarily focus on describing the current research done in the field of student feedback analysis. Section IV describes our system in detail for topic and sentiment extraction from students' comments. Section V describes our dataset and pre-processing of data. In section VI, we focus on experiments, results, discussions and pointing some interesting future directions of our work, and we conclude in section VII.

## II. BACKGROUND

Text mining and natural language processing techniques are useful for opinion mining research. Therefore, we first provide a brief description of few text analytics techniques that are key components to our system, followed by the background of opinion mining research.

### *A. Text Analytics Techniques*

Text analytics or text mining is a knowledge discovery technique that provides computational intelligence [12] [19] through devising of patterns and trends. The techniques comprise of multidisciplinary fields, such as information retrieval, extraction, text analysis, natural language processing, and data mining. Text mining techniques enable to identify similarities between text attributes [12]. Some of the natural language issues that should be considered during text mining are tokenization, stop word lists, etc.

**Stop word removal**: Most frequently used words in English are useless in Text mining. For example "has", "if", "and", "on" etc. Such words are called stop words. Stop words are language specific functional words which carry no information and therefore removed from the documents during data pre-processing stage. Parts of Speech such as pronouns, prepositions, conjunctions are defined in stop word list[1].

**Tokenization:** Tokenization deals with the splitting of text into units during data pre-processing. Text can be tokenized in to paragraphs, sentences, phrases and single words. The delimiters used in this process vary with data sets.

**Stemming:** This method is used to find out the root/stem of a word. Words are stemmed using the Porter Stemming algorithm [31], which returns the root form of a word. For instance, the word "progression" is stemmed as "progress" and "progress*" is formed as part of the query. However, in our preliminary experiments we observed that, stemming impairs our results. Therefore, we do not use stemming.

**Document Representation:** In order to score the similarity between two documents, we need to first adopt a vector space representation of a document where each document is evaluated as a term-frequency (TF) vector [18] and inverse document frequency (IDF) [18]. TF-IDF is a statistical measure or weight often used in information retrieval and text mining to evaluate how important a word or term is to a document in a collection or corpus. Term frequency is the number of occurrence of a term in a document. The information that is captured by term frequency tells how salient a word is within a given document. Document frequency on the other hand can be interpreted as an indicator of informativeness. Inverse document frequency is used to scale down the term frequency of terms with high total number of occurrences in the collection. Both these measures aids in generating the aspects or topics for a comment in our case. One way to combine a word's term frequency and inverse document frequency into single weight is a TF-IDF. Finally, each document in the dataset is represented as a document-term matrix.

**Document similarity score:** The similarity score between two documents determines the co-occurrence of a primary topic between two documents to cluster them together. We compute this score by computing the cosine angle between them [18] which are modeled as vectors in a vector space.

[1] www.ranks.nl/resources/stopwords.html

**Agglomerative Clustering**: Clustering algorithms are exploratory data analysis tools that have proved to be essential for gaining valuable insights on various aspects and relationships of the underlying textual data [16]. Agglomerative algorithms find the clusters by initially assigning each object to its own cluster and then repeatedly merging pairs of clusters until either the desired number of clusters has been obtained or all the objects have been merged into a single cluster leading to a complete agglomerative tree. The key step in these algorithms is the method, also referred to as clustering function, used to identify pairs of clusters to be merged iteratively.

### B. *Opinion Mining*

Opinions are central to almost all human activities and are key influencers of our decision making process. Opinion mining is a well-studied research topic for the past ten years mainly focusing on opinion extraction, sentiment classification, opinion summarization and applications in real world. Opinion mining found its roots in many real-life applications and several application-oriented research studies have been published.

Figure 2 shows the architecture of opinion mining. Opinion mining architecture takes users' comments as inputs to generate sentiment analysis visualizations as outputs that can aid the decision makers in decision making process. The text processing component handles data cleansing and processing issues. In next subsections, we briefly explain the main components of the architecture namely, topic extraction, sentiment classification and opinion summarization.

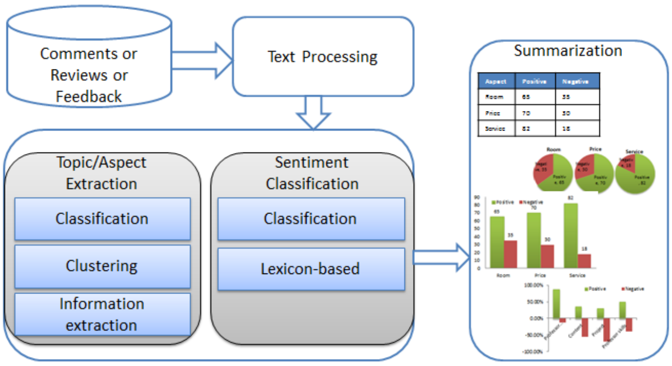


Fig. 2. Opinion mining architecture.

### C. *Topic Extraction*

Opinion extraction aims at automatically finding attitudes or opinions about specific targets, such as named entities, consumer products or public events [8] [15]. An opinion without its target being identified is of limited use [9]. For many applications opinion extraction is insufficient, and a fine-grained opinion mining and analysis such as topic or aspect extraction is highly effective [9] [10]. "The iPhone's call quality is good, but its battery life is short" evaluates two aspects, call quality and battery life. Hu and Liu used a data mining algorithm that finds explicit aspect expressions that are nouns and noun phrases from a large number of reviews in a given domain [9]. Jiang et al. proposed how a dependency parser was used to generate a set of aspect dependent features for classification [15]. Many algorithms based on supervised learning have been proposed in the past for information extraction [14]. Clustering based feature extraction techniques are implemented by some research works [6] [7]. Beil et al. designed clustering technique on the basis of frequent pattern mining [7]. Lu et al. proposed clustering based technique for discovering aspects from users' comments [6]. Inspired by these works, we use agglomerative clustering to group comments into clusters based on their cosine similarity.

### D. *Sentiment Classification*

Sentiment classification aims at classifying the data into positive or negative polarities [12] using supervised methods or unsupervised methods. Similar to opinion extraction, fine grained sentiment analysis is desired as it is highly effective to understand the pulse of the consumers at feature level. The task of sentiment target detection [9] aims at extracting the sentiment targets in the reviews using multiple heuristic techniques. Pang et al. examined several supervised machine learning methods like SVM and Bayes classification for sentiment classification of movie reviews and showed that classifiers performed poorly on sentences as sentences contains less information [12].

Lexicon methods are based on sentiment words and phrases which are instrumental to sentiment analysis for obvious reasons [8]. A list of such words and phrases is called a sentiment lexicon (or opinion lexicon). Over the years, researchers have designed numerous algorithms to compile such lexicons; SentiWordNet [11] and Sentiment lexicon [9]. Our system generates sentiment for each topic using classification approach.

### E. *Opinion Summarization*

Summarization is a study that attempts to generate a concise and digestible summary of a large number of opinions [8]. Current research aims at two types of summarization: aspect-based summarization and non-aspect-based summarization. Aspect-based summarization divides input texts into aspects, which are also called features, and generates summaries of each aspect [9] [13]. A common form of summary is based on aspects and is called aspect-based opinion summary (or feature-based opinion summary) [8] [9].

## III. RELATED WORK

Traditionally, universities collected written feedback from students regarding the course taught and the professor's engagement in order to assist the development of the course through future cycles. Pedagogical theory of student feedback describes the need for interpreting students' perceptions and sentiments for overall teaching evaluation and improvements [33]. Donovan et al. [17] found that online student feedback comments were longer and that they were more formative in nature than the traditional written feedback. Moreover, online feedback received longer and half as many (54% or more) comments as traditional written comments. This highlights the importance of collecting online comments. However, manually reading and analyzing these online comments takes a lot of time and hence the need for an automated feedback system

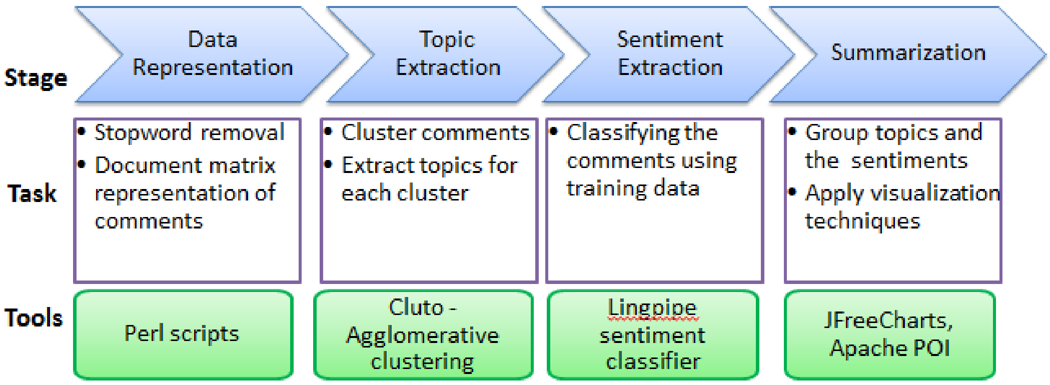


Fig. 3. SFMS system architecture

which automates feedback collection and analysis, allowing a visual analysis of opinions or sentiments on different aspects of the course.

In existing research on educational data mining, the more prominent forms of analysis are with Apriori algorithms, decision trees and clustering algorithms [18] with most research being done on association pattern mining to find links between opinions on courses and professors in order to better cater to students, enhance their grades, prevent drop-out or transfers and improve the overall degree experience. Altrabsheh et al. devised a system to analyze sentiments in real time to provide real-time intervention in the classroom. Their experiments yielded the conclusion that Support Vector Machines and Complement Naïve Bayes produced the most accurate results while learning sentiment [19]. Hajizadeh et al. experimented on student feedback to analyze whether or not a student would retake the course [20] indicating sentient opinions about the course. Rashid et al. used generalized sequential pattern mining and association rule mining with 87% accuracy to analyze opinion words from student feedback while stopping short of sentiment classification upon identification of the opinion words [21]. Gamon et al. took another approach for analyzing sentiment in free flowing text – as is with student feedback as well – by building a system, Pulse, that brought together algorithms that clustered topics and classified sentiments with intuitive visualization to allow a deeper analysis of customer feedback and sentiment on special topics [22].

Qualitative research in education suggests that student feedback is not only important for course improvement; it also allows universities to align their courses to international accreditation standards in the avenue of quality control and assurance of universities [23]. The actual use of the analyzed data is also a subject for research where researchers have tried to tie feedback to changes in teaching, grading and self-evaluation for professors. Yao et al. found that professors do indeed care about student feedback and used discretion in using formative comments in modifying their teaching. Despite variations in the ultimate use of the feedback collected, they found that sentiments towards student feedback ranged from neutral to positive, indicating the usefulness of collecting feedback [24].

Given course codes and students' comments as inputs, the goal of our project is to develop a system that can extract and visualize the topics and sentiments on courses. In the next section, we explain the details of SFMS system.

## IV. STUDENT FEEDBACK MINING SYSTEM (SFMS)

In this section, we describe the architecture of SFMS (depicted in Figure 3), which basically follows the opinion mining architecture shown in Figure 2. The first layer in the architecture shows the main stages of the system. Second layer depicts the key tasks in each stage. The third layer depicts the tools or techniques used to accomplish the tasks in each stage.

The system consists of four main stages as shown in Figure 3. In first stage, a dense matrix of comments is generated after preprocessing the data. In second stage, the comments are clustered based on their primary common topic. In third stage, sentiment of each comment is extracted, and finally in the fourth stage, topics and sentiments are aggregated for comprehensive reporting. SFMS system is developed in Java platform.

The system initiates with the course codes and corresponding comments as inputs and executes all the stages to generate visual analysis reports. For example, "The course project is very difficult but very challenging" is a comment for a course code, IS203. We explain each stage in detail in the subsequent sections.

### *A. Data Representation*

In data representation stage, first, all terms are extracted from input comments using tokenization by space. Second, stop words are removed from each comment using the stopword list. Document representation is shown in Figure 4.

| | Feature | Value | Feature | Value | Feature | Value |
|---|---|---|---|---|---|---|
| 1 | pace | 1.000000 | time | 1.000000 | uploaded | 1.000000 |
| 2 | sometimes | 2.000000 | require | 1.000000 | meet | 1.000000 |
| 3 | students | 1.000000 | sometimes | 1.000000 | factor | 1.000000 |
| 4 | meet | 1.000000 | teams | 1.000000 | factor | 1.000000 |
| 5 | ridiculous | 1.000000 | is480 | 1.000000 | grades | 1.000000 |
| 6 | students | 2.000000 | especially | 1.000000 | understand | 1.000000 |

Fig. 4. Document representation matrix (document – term - value). Each row represents comment. Feature represents word or term and value represents frequency of a term in the document. Due to space constraints only three features for sample comments are depicted.

In order to calculate the similarity between two comments, we need to transform the free-flowing text in the comment into a numerical matrix as shown in Figure 4. Each row denotes a comment, and each cell in that row is occupied by a single word (or feature) from that comment together with its term frequency (log or square root).

Such data representation would assist in judging the importance of each word in a comment and therefore measuring the similarity between comments. To generate the document term matrix, we use perl scripts. Our interface takes the comments and stopword list as inputs and generates document matrix.

*B. Topic Extraction*

A topic is the subject or target of a student's comment. For example, given the comment, "The course project is very difficult but very challenging", "project" is the topic of the comment. In topic extraction phase, the objective is to breakdown all the comments by topics such as teaching, content, learning etc. To achieve this, the first task is to cluster the comments using clustering algorithms and specific clustering functions.

Various clustering criterion measures such as $I_1$, $E_2$, $H_2$ etc., are available for measuring the clustering similarity [29]. These schemes differ on how the similarity between the individual objects in various clusters is combined to determine the similarity between the clusters themselves. Table I provides the notation for the formulae and Table II provides the formulae for selected clustering functions.

TABLE I. NOTATION

| |
|---|
| $S$ : collection of documents |
| $S_1, S_2, \ldots S_k$ : set of document of *k-th* cluster |
| $k$ : number of clusters |
| $n_1, n_2, \ldots n_k$ : number of docs of corresponding clusters |
| $C_r$: centroid vector of r-th cluster |
| $C$: the centroid vector of the entire collection |
| $d_i$, $d_j$: i-th and j-th documents |
| $D$: the composite vector of the entire docs |

TABLE II. MATHEMATICAL FORMULAE FOR CLUSTERING CRITERION [29]

| Criterion | Formula |
|---|---|
| $I_1$ | $Maximize(I_1) = \sum_{r=1}^{k} n_r, \left( \frac{1}{n_r^2} \sum_{d_i,d_j \in S_r} \cos(d_i, d_j) \right)$ |
| $I_2$ | $Maximize(I_2) = \sum_{r=1}^{k} \sum_{d \in S} \cos(d_i, C_r)$ |
| $E_1$ | $Minimize(E_1) = \sum_{r=1}^{k} n_r \cos(C_r, C)$ |
| $H_1$ | $Maximize(H_1) = \frac{I_1}{E_1}$ |
| $H_2$ | $Maximize(H_2) = \frac{I_2}{E_1}$ |

$I_1$: This function tries to maximize the intra cluster similarity between the elements of a cluster.
$I_2$: This function also tries to maximize the intra cluster similarity between the elements of a cluster. The only difference between $I_1$ and $I_2$ is that while calculating I2 we must take the square root of the function.
$E_1$: This function divides the intra-cluster similarity with inter cluster similarity.
$H_1$: This is a hybrid function to maximize $I_1/E_1$.
$H_2$: This is a hybrid function trying to maximize $I_2/E_1$.

To cluster the comments, we use agglomerative clustering algorithm, and the tool we use is Cluto clustering library [25]. Similar comments will be grouped together by the clustering algorithm. The top words in the cluster represent the topic of the cluster. The examples are demonstrated in our experiments section.

Once clusters are generated, the second task is to extract topics for the clusters. In this context, the topics are the high frequency words that appear in each cluster. For example, the words like *project*, *time*, *practice* etc., are some of the high frequency words that represent the cluster with comments related to topic, *project*. However, the label for the cluster should be manually provided to generate a meaningful representation for a cluster. We developed an interface that accepts users' inputs for the cluster labels and the system uses them for subsequent phases.

Mostly student comments refer to a single topic, but there are few instances that they may span across many topics in a single comment. A comment with multiple topics is not a focus of our work and we leave it to future work.

*C. Sentiment Extraction*

Discovering the sentiment of each comment provides the user with an analysis of collective sentiments against each topic or each cluster in the entire collection. Sentiment refers to the positivity or negativity of a given comment. For example, given the comment, "The course project is very difficult but very challenging", the sentiment is "negative".

In this phase, the objective is to find the overall positive or negative sentiment for a given comment. We propose a classification based approach for this task and therefore created a training set for the training the classifier. For this purpose, we use LingPipe Language Identification Classifier [26] which adopts the classification approach to sentiment analysis using a sentence-level logical regression classifier. It deconstructs each comment sentence into n-grams, or number of words evaluated at a time while processing the sentiment of the comment. We have chosen to use bi-grams for sentiment extraction task. We use bi-grams as they aid in processing negating phrases such as "not good".

Using bi-grams, LingPipe evaluates two words at a time before assigning an overall sentiment to the comment. This allows evaluation of double negatives which allow a better evaluation of sentiment. The classifier learns the natural distribution of characters in the language model of a training data set and then assigns a sentiment probability to each evaluated bi-gram according to a probability distribution. Eventually leading to an aggregated final sentiment for each comment evaluated. Agarwal et al. [30] evaluated the use of three categories of sentiment for basic polarity in sentiment – negative, positive and neutral, but found that the results were better with strict polarity between positive and negative only. In our preliminary studies, we observed that the students' comments are mostly negative or positive. We leave neutral component for future exploration.

### *D. Summarization*

Summarization is the final stage where the goal is to provide user friendly summaries of the quantitative results obtained from the previous phases. Once the comments have been clustered into topics and the sentiment for each cluster is known, we categorize the comments by course. Therefore, each course has its own set of *k* clusters with their individual comments annotated with a sentiment. Essentially, the courses serve as high level category or a curriculum level summarization. In contrast, individual comments serve as course level summarization. Visualization charts use the topics, sentiments and course codes as inputs. The charts are generated using JFreeCharts [27] and inserted into Microsoft Excel files that are created and manipulated using Apache POI libraries [28] for enabling users with easy analysis. We adopt the charts similar to feature-based sentiment summaries by Hu et al. [13].

## V. DATASET

We use dataset of feedback comments given by students attending courses offered by the School of Information Systems at Singapore Management University for the academic year 2013-14. These comments are collected at two feedback cycles, midterm and end-term, and span across two semesters. In total, seven courses are evaluated, yielding 5,341 comments for evaluation.

In our data analysis, we noticed that some students provide "NA" or "Nil" comments. In order to avoid noise in the results, we removed comments with less than 10 characters. This allows us to focus on those comments which would yield a constructive view on topics being discussed and their respective sentiments. Finally, we have 3,144 comments for our experiments.

## VI. EXPERIMENTS

We first explain experiment setup followed by the results.

### *A. Experiment Setup*

We developed SFMS as a desktop application with simple graphical interface. Recall that our first stage of SFMS system is to generate document representation in matrix format. To generate the document term matrix, we use *doc2mat* perl scripts provided in the Cluto library [25] and Figure 5 shows the GUI for data representation stage. The UI takes the comments and stop word list as inputs and generates document matrix in a given location. We observed that words such as "students", "course" etc. that occur very frequently in the dataset generate noise and impact the quality of customers. Therefore, these words are added to the current stopword list.

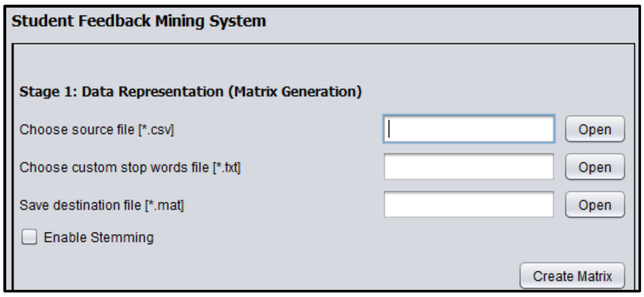


Fig. 5. SFMS System UI: Data Representation Stage.

Cluto API is an easy-to-use platform that combines a variety of different clustering algorithms. We use vcluster (agglomerative) in the toolkit to generate clusters. Cluto provides three row models; log, MAXTF and square root. All our experiments are based on agglomerative clustering with cosine similarity and log model. We set number of clusters to 10 after some preliminary experiments. For sentiment classification, we use Lingpipe [26] which provides a sentence based logistic regression classifier for sentiment classification.

### *B. Topic Extracion Results*

We first present quantitative results on clusters followed by qualitative analysis of topics generated. Recall that Cluto provides multiple clustering functions to determine clusters as described in Section IV. Figure 6 depicts the GUI for cluster generation.

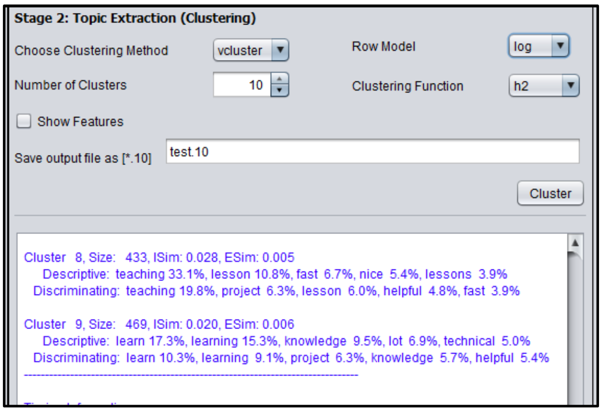


Fig. 6. SFMS System UI: Topic Extraction Stage (Clustering task).

The UI allows users to choose cluster methods, term frequency measures, count of clusters, and clustering functions to generate clusters. Various clustering criterion measures such as $I_1$, $E_2$, $H_2$ etc., are available for measuring the clustering similarity [29] as described in Section IV. SFMS calls Cluto API with the user inputs and performs clustering. For our preliminary experiments, we tested with multiple combinations. However, we present only the results from the selected combinations that provide better performance. We use *Purity* (the higher the better) and *Entropy* (the lower the better) to evaluate the performance of clustering algorithm in topic extraction phase [8]. From our results, we observed that the clustering function, $H_2$ provides Purity of 93.4%, which is slightly higher than other clustering functions and Entropy of 0.214. Therefore, we use $H_2$ for subsequent experiments. We now present the qualitative analysis of topics generated by clustering task. Using $H_2$, the comments are clustered and the top features of each cluster are as shown in Table IV.

TABLE III. TEN CLUSTERS WITH TOP WORDS AND HUMAN ALIAS ($H_2$)

| Cluster # | Top frequency words | Alias |
|---|---|---|
| 0 | approachable, friendly, enthusiastic, consultation, help | Faculty interaction |
| 1 | helpful, feedback, concepts, understanding, encouraging, help | Faculty feedback |
| 2 | patient, knowledgeable, passionate, responsible, fun | Faculty preparation |
| 3 | project, heavy, time, requirements, lot | Project |
| 4 | time, assignment, sql, labs, php | Assignments |
| 5 | challenging, lab , test, project, exercises | Labs |
| 6 | excel, future, skills, real, applicable | Skills |
| 7 | understand, concepts, help, questions, explain | Concepts understanding |
| 8 | teaching, lesson, fast, nice, lessons | Classroom delivery |
| 9 | learn, learning, knowledge, lot, technical | Learning experience |

Each cluster has distinguishing or determining features or words which determine the topic of the cluster. We notice that all clusters are very coherent and meaningful except clusters 4 and 5 which both refer to labs. This is one of the drawbacks of clusters as it is unsupervised. To improve the quality, one approach is to exploit the questions together with the comments and we leave it to future work on improving the quality of the clusters. SFMS system provides an UI to users to provide *Alias* or *labels* for each cluster as shown in Figure 7. Once aliases are provided, the comments are also categorized by course codes. This categorization aids in generating user friendly visual reports.

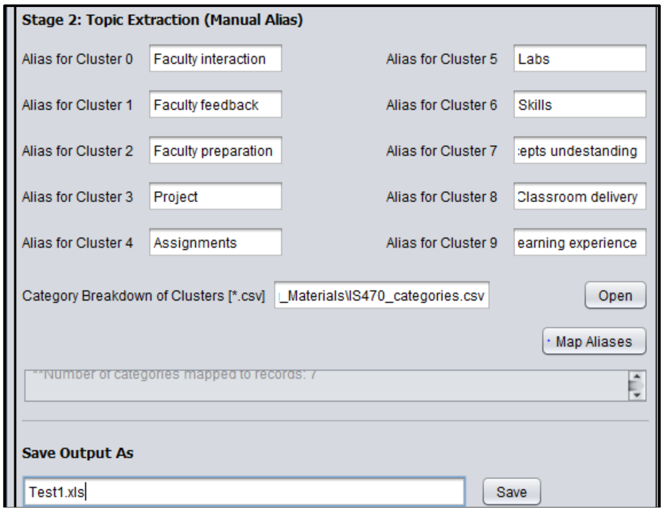


Fig. 7. SFMS System UI: Topic Extraction Stage (Human labeling task).

## C. Sentiment Extraction Results

In the sentiment extraction phase, we use human labeling for training the data and evaluating SFMS. Lingpipe uses logistic regression and we evaluated the effect of domain knowledge on the training of sentiment classifier. Figure 8 shows UI for sentiment classification. The UI allows users to provide training data to SFMS to train the classifier.


Fig. 8. SFMS System UI: Sentiment Extraction Stage.

We used log regression model trained on Internet Movie Database (IMDB) domain and education domain. Sample comments and the corresponding sentiment classification results are depicted in Table II. We observe that training the classifier on education domain gives best results instead of the standard (IMDB) dataset provided by Pang et al [6].

TABLE IV. SAMPLE COMMENTS AND COMPARISION OF BOTH DOMAINS FOR SENTIMENT CLASSIFCATON TASK

| Function | IMDB | Education |
|---|---|---|
| very knowledgeable, patient and easygoingl | -ive | **+ive** |
| sometime he went through the concepts a bit too fast for us to gasp. | +ive | **-ive** |
| always concern for student and willing to help weaker student | -ive | **+ive** |
| Asks challenging questions to get us to think deeper. | **+ive** | -ive |

Table II shows some example comments and the corresponding sentiment labels generated by SFMS when trained on IMDB and education domain. In our analysis, we observed that low precision for IMDB domain is due to false positives. In contrast, the classification approach labeled it as positive. Overall, the sentiment extraction phase with education domain training has a precision of 80.1%, which is significantly higher than IMDB trained classifier.

## D. Feedback Summarization Results

Topics and sentiments generated by previous phases are used for reporting using *JFreeCharts* [27].

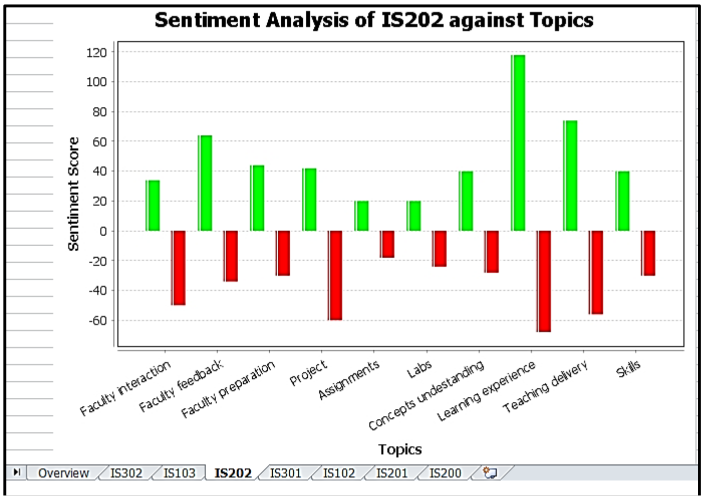


Fig. 9. Integrated visualization of topics and sentiments for seven courses in Information systems curriclum. Excel sheet generated by SFMS.

Figure 9 shows the integrated reporting view of student feedback which can be useful for curriculum designers and management. To categorize the results, course codes are used for generating reports. It has the clusters information as well as reports for every course in our dataset. Figure 10 provides deeper analysis for the course instructors on various aspects of the course.

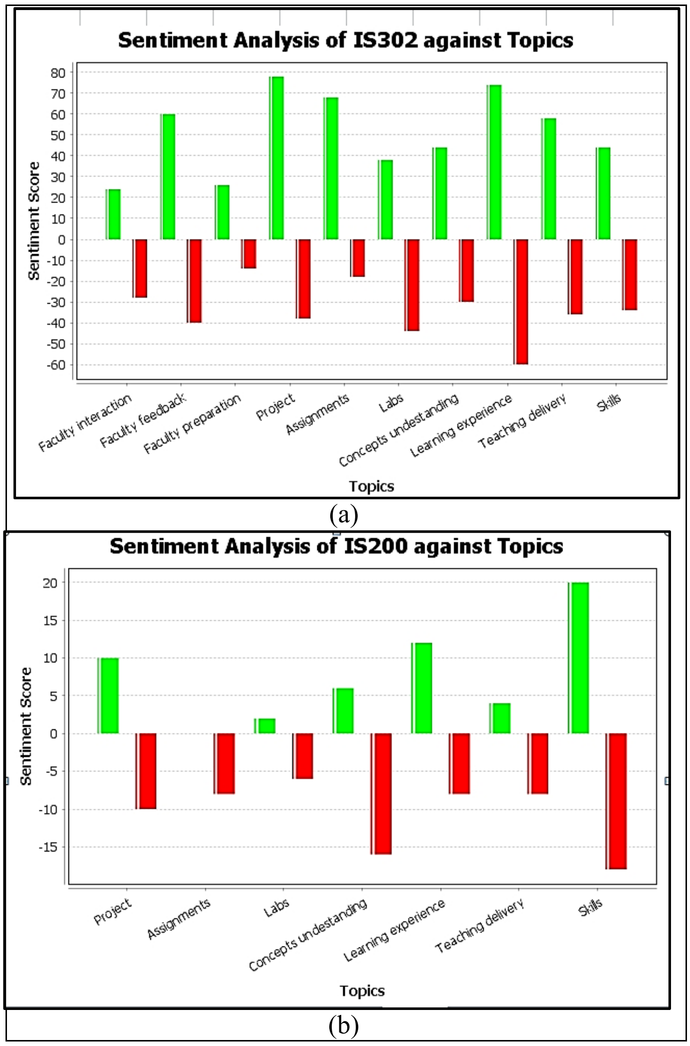


Fig. 10. Visualization of topics and sentiments for Information systems course codes. (a). Course code, IS302 - Information security & trust.(b). Course code, IS200 - Software Foundations.

We observe that the students provides comments on aspects such as project, labs, skills, etc. of IS200 (IS software foundations) course, which is programmatic in nature. However, the faculty feedback and interaction is not of their concern. In contrast, for IS302 (Information security & trust) course which is less programmatic in nature, but with an open research project, the students are concerned with the faculty feedback and interaction. We observe the negative sentiments are also quite high. Therefore, there is a need for the faculty to plan for some changes in the project or consultation sessions. For example, faulty may provide additional consultation hours or online discussion forums. For IS200, faculty may provide some extra tutorials to improve student learning experience.

### E. *Discussions and Future Work*

One major limitation of Cluto is that, each comment can only belong to one cluster. This means that even though a single comment can span multiple topics, it will be clustered under the primary topic – or the topic with most of its discriminating significant words similar to those in the comment, as judged by the clustering function. Topic models such as LDA can be explored to overcome this limitation which we leave it to future work. Similarly, exploring sentence based topic-sentiment is an interesting future work. Currently, SFMS can only take in a single-level categorization for distribution of clusters. Other categories such as term, year, school, faculty, etc., can provide detailed analysis. Future iterations of the development of this system could create dynamic hierarchies of categories that will allow users and analysts to drill down dynamically into topics and sentiments by each level for deeper analysis. Secondly, SFMS offers a sentiment score for each topic being discussed by students but does not go deeper to signify what the actual comments spoke about. Faculty might want to retrieve the comments interactively for further analysis. Lastly, students often leave suggestions for professors regarding delivery, content, interactions and so on. Since these relate strongly to each of the topics being analyzed, each topic and its respective sentiment can also have a highlighted set of suggestive comments to take the system one step forward from descriptive analytics to actionable insights.

## VII. CONCLUSION

In this paper, aspect or topic based sentiment mining techniques are evaluated in order to build a desktop-based solution to analyze topics and their sentiments from student generated feedback in universities. We found that agglomerative clustering with cosine similarity using a hybrid approach generates coherent clusters for topic extraction task. Further, using a logistical regression algorithm, which is trained on education domain, extracts sentiments on comments with higher accuracy compared to the classifier trained on movies domain. Free flowing textual data like student feedback in an education context can be therefore analyzed automatically in order to gather a deeper understanding and facilitate the stakeholders in course improvement cycles.

## REFERENCES


[1] Hounsell, D. (2003) The evaluation of teaching in H. Fry, S. Ketteridge, and S. Marshall. A handbook for teaching and learning in highereducation: enhancing academic practice London: KoganPage

[2] Murray, H. (1997). Does evaluation of teaching lead to improvement of teaching? International Journal for Academic Development, 2(1), 8-23

[3] Elaine Keane & Iain Mac Labhrainn , Obtaining Student Feedback on Teaching & Course Quality , CELT, April 2005

[4] Brennan, J. & Williams, R. (2004) Collecting and Using Student Feedback. A Guide to Good Practice (LTSN, York).

[5] Mark Huxham*, Phyllis Laybourn, Sandra Cairncross, Morag Gray, Norrie Brown, Judy Goldfinch and Shirley Earl (2008). Collecting student feedback: a comparison of questionnaire and other methods. Published in Assessment & Evaluation in Higher Education 2008, 1---12

[6] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In Proceedings WWW pages 131–140, , Spain, 2009, ACM.

[7] F. Beil, M. Ester, X. Xu. Frequent term-based text clustering, ACM KDD Conference, 2002

[8] B. Liu. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, 2nd ed, 2010.

[9] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews (2004). In 19th national conference on Artifical intelligence, AAAI'04.

[10] A.-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. In Proceedings of HLT '05: October 2005.

[11] Esuli, Andrea and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. in Proceedings of Language Resources and Evaluation (LREC-2006). 2006.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In EMNLP '02: Proceedings of the 2002 pages 79–86, July 2002.

[13] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In KDD '04: Proceedings of the tenth ACM SIGKDD, 168–177, 2004.

[14] Sarawagi, Sunita. Information extraction. Foundations and Trends in Databases, 2008. 1(3): p. 261-377.

[15] Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In Proceedings (ACL-2011).

[16] Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

[17] Donovan, J., Mader, C. E., & Shinsky, J. (2010). Constructive Student Feedback: Online vs. Traditional Course Evaluations. *Journal of Interactive Online Learning*, *9*(3), 283-296

[18] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM*, *1*(1), 3-17.

[19] Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. In *Adaptive and Intelligent Systems* (pp. 40-49). Springer.

[20] Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect students' academic performance-Data Mining Approach. *arXiv preprint*.

[21] Rashid, A., Asif, S., Butt, N. A., & Ashraf, I. (2013). Feature Level Opinion Mining of Educational Student Feedback Data using Sequential Pattern Mining and Association Rule Mining. *International Journal of Computer Applications*,*81*(10), 31-38.

[22] Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*(pp. 121-132). Springer Berlin Heidelberg.

[23] Leckey, J., & Neill, N. (2001). Quantifying quality: the importance of student feedback. *Quality in Higher Education*, *7*(1), 19-32.

[24] Yao, Y., & Grady, M. L. (2005). How do faculty make formative use of student evaluation feedback?: A multiple case study. *Journal of Personnel Evaluation in Education*, *18*(2), 107-126.

[25] Karypis, George. "Data Clustering Software.". N.p., n.d. Web. 17 Apr. 2015. http://glaros.dtc.umn.edu/gkhome/views/cluto

[26] Alias-i. "Lingpipe." Lingpipe. Alias-i, 01 Jan. 2003. Web. 17 Apr. 2015. http://alias-i.com/lingpipe/index.html

[27] Object Refinery Limited. "JFreeChart". JFreeChart. Jan 2005. Web. 17 Apr. 2015. http://www.jfree.org/jfreechart/

[28] The Apache Software Foundation. "Apache Poi Project". Apache POI. Jan 2002. Web. 17 Apr. 2015. https://poi.apache.org/

[29] Zhao, Y., Karypis, G., & Du, D. Z. (2005). *Criterion functions for document clustering* (Doctoral dissertation, University of Minnesota.).

[30] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). ACL.

[31] M.F. Porter. An algorithm for suffix stripping, Program, 14 (3):130-137, 1980

[32] Scott, G., Grebennikov, L., & Shah, M. (2008). Using qualitative data to prove and improve quality in Australian higher education. *Evidence based Decision Making: Scholarship and Practice*, 97.

[33] K. P. Mohanan, the place of student feedback in teaching evaluation. http://www.cdtl.nus.edu.sg/publications/studfeedback/StudFeedback_TeachQuality.pdf