Mining Feature-opinion in Educational Data for Course Improvement

Alaa El-Halees
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
alhalees@iugaza.edu.ps

ABSTRACT

In academic institutions, student comments about courses can be considered as a significant informative resource to improve teaching effectiveness. This paper proposes a model that extracts knowledge from students' opinions to improve and to measure the performance of courses. Our task is to use user-generated contents of students to study the performance of a certain course and to compare the performance of some courses with each others. To do that, we propose a model that consists of two main components: Feature extraction to extract features, such as teacher, exams and resources, from the user-generated content for a specific course. And classifier to give a sentiment to each feature. Then we group and visualize the features of the courses graphically. In this way, we can also compare the performance of one or more courses.

KEYWORDS

mining student opinions, opinion feature extraction, opinion mining, student evaluation, opinion classification.

1 INTRODUCTION

Opinion mining is a research subtopic of data mining aiming to automatically obtain useful knowledge in subjective texts [3]. It has been widely used in real-world applications such as e-commerce, business-intelligence, information monitoring and public polls [4].

In this paper we propose a model to extract knowledge from students' opinions to improve teaching effectiveness in academic institutes. One of the major academic goals for any university is to improve teaching quality. That is because many people believe that the university is a business and that the responsibility of any business is to satisfy their customers' needs. In this case university customers are the students. Therefore, it is important to reflect on students' attitudes to improve teaching quality. Students post comments of courses using Internet forums, discussion groups, and blogs which are collectively called *user-generated content*. Our task is to use user-generated contents of students' comments to study the performance of a certain course and to compare the performance of some selected courses.

To do that, we used the following tasks which are usually used in opinion mining: First, extract courses' features that are commented by students in the user-generated contains. Course feature is an attribute of a specific course such as contain, teacher, ..etc. In this step a feature extraction method will be proposed. Second, determine the attitude of the student toward the feature (i.e. if the student like or dislike the teacher of the course). In this step sentiment analysis classification will be used. Third, visualizing and summarizing all features for each course. Finally, comparing the features of a course with the features of another course.

To test our work we collected data from students who expressed their views in discussion forums dedicated for this purpose. The language of the discussion forums is Arabic. As a result, some techniques are used especially for Arabic language.

The rest of the paper is structured as follows: section two discusses related work, section three about the proposed method, section four describes the conducted experiments, section five gives the results of experiments and section six concludes the paper.

2 RELATED WORKS

Hu and Liu in [3] used frequent features generation to summarize products customer review. They summarized only specific features of the product that customers have opinion on and also whether the opinions are positive or negative. Also, Somprasertsri and Lalitrojwong in [4] proposed a method to summarize product customer review. But they used a different method; they applied dependency relations and ontological knowledge with probabilistic based model.

Using opinion mining in education, we found three works that mentioned the idea of using opinion mining in education. First, Lin et al. in [5] discussed the idea of Affective Computing which they defined as a "Branch of study and development of Artificial Intelligence that deals with the design of systems and devices that can recognize, interpret, and process human emotions". In there work, the authors only discussed the opportunities and challenges of using opinion mining in E-learning as an application of Affective Computing. Second, Song et. al. in [6] proposed a method that uses user's opinion to develop and evaluate E-learning systems. The authors used automatic text analysis to extract the opinions from the Web pages on which users are discussing and evaluating the services. Then, they used automatic sentiment analysis to identify the sentiment of opinions. They showed that opinions extraction is helpful to evaluate and develop E-learning system. Third work of Thomas and Galambos in [7] investigated how students' characteristics and experiences affect their satisfaction. They used regression and decision tree analysis with the CHAID algorithm to analyze student opinion data. They concentrated on student satisfactions such as faculty preparedness, social integration, campus services and campus facilities.

3 THE PROPOSED METHOD

Opinion mining discovers opinioned knowledge at different levels such as at clause, feature, sentence or document levels [8]. This paper discusses how to extract opinions in feature level. Features of a product are attributes, components and other aspects of the product. For course improvement feature may be course content, teacher, resources …etc.

We can formulate the problem of extracting features for each course as follows: Given user-generated contents about courses, for each course *C* the mining result is a set of pairs. Each pair is denoted by (*f*, *SO*), where *f* is a feature of the course and *SO* is the semantic orientation of the opinion expressed on feature *f*.

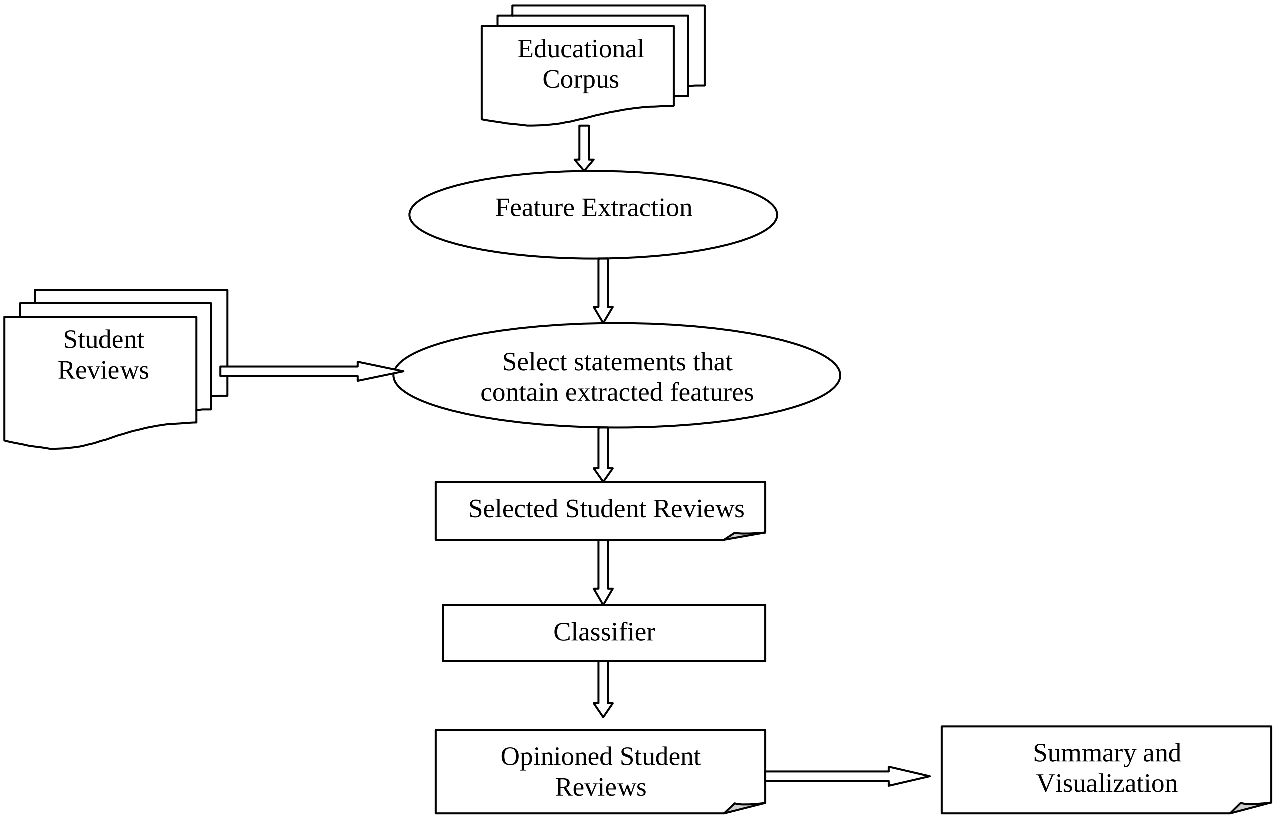
Educational Corpus
Feature Extraction
Student Reviews
Select statements that contain extracted features
Selected Student Reviews
Classifier
Opinioned Student Reviews
Summary and Visualization

Figure 3.1 Steps of the proposed method

We proposed a model, as in figure 3.1, has the following steps:

1) Arabic Corpus, which contains opinion expressions in higher education domain, was collected from the Internet.
2) After preprocessing the corpus and tagged each word in the dataset using Part of Speech (POS), we used modified version of WhatMatter System proposed by Siqueira and Barros in [9] to select and extract features. It is a system for feature extraction from opinions on services. The system's process receives as input a text containing an opinion, and returns the extracted features list. It includes the following tasks:

a) Frequent nouns identification: To collect all frequent nouns in the corpus. A noun is frequent if it is within certain threshold (i.e. we used 4% as the best value).

b) Relevant nouns identification: It also collects all nouns adjacent to adjectives.

c) Unrelated nouns removal: It filters irrelevant nouns using the PMI-IR measure from [10]. This measure is given by the formula bellow:

$$PMI(t_1, t_2) = \log_2 \left(\frac{Hits(t_1 \wedge t_2)}{Hits(t_1) * Hits(t_2)} \right)$$

where Hits(t_1) is the number of pages containing t_1, Hits(t_2) is the number of pages containing t_2 and Hits(t_1 ^ t_2) is the number of pages with both terms. Using quires in Google, t_1 is the tested noun and t_2 is education domain (i.e. course).

3) Given the student review for a course, this step simply filters all reviews which do not contain one of the features in the feature list extracted in the previous step.

4) Determining whether the opinion on the feature is positive or negative, a binary classifier is used (i.e. *Naïve Bays).*

5) Aggregates all opinions in a course by counting how many positive and negative opinions were given for each feature. In this case we need to look at a synonym of the feature's name. That because many features may have a different name for the same entity (i.e. professor and teacher).

6) Visualize the feature for a course, two courses or more can be graphed together to obtain more clear comparison.

4 EXPERIMENTS

To evaluate our method, a set of experiments was designed and conducted. In this section we describe the experiments design including the corpus, the preprocessing stage, the used data mining method and evaluation metrics.

4.1 Corpus

Initially we collected data for our experiments using 4,957 discussion posts which contain 22 MB of data from three discussion forums dedicated to discuss courses. We focused on the content of five courses including all threads and posts about these courses. Table 1 gives some details about the extracted data. Details of data for each selected course are given in table 2.

Table 1: A summary of the used corpus

Total Number of posts	167
Total Number of Statements	5017
Average number of statements in a post	30
Total Number of Words	27456
Average number of words in a post	164

Table 2: Details about data collected for each of the five courses.

Course To Review	Number of Posts	Number of Sentences	Number of Words
Course_1	69	1920	13228
Course_2	34	1321	7280
Course_3	23	617	3587
Course_4	21	524	3183
Course_5	20	635	3407

The rest of the collected data is used to extract features and as training set for the classifiers. For that, we manually annotated each post to positive and negative.

4.2 preprocessing

After we collected the data associated with the chosen five courses, we striped out the HTML tags and non-textual contents. Then, we separated the documents into posts and converted each post into a single file. For Arabic scripts, some alphabets have been normalized (e.g. the letters which have more than one form) and some repeated letters have been cancelled (that happens in discussion when the student wants to insist on some words). After that, the sentences are tokenized, stop words removed and Arabic light stemmer applied. Also, each sentence is analyzed by a Part-of-Speech (POS) Tagger. Then, we obtained vector representations for the terms from their textual representations by performing TFIDF weight (term frequency–inverse document frequency) which is a well known weight presentation of terms often used in text mining [11]. We also removed some terms with a low frequency of occurrence.

4.3 Classifier's Methods

In our experiments to classify posts, we applied a machine learning method which is Naïve Bayes. Naïve Bayes classifiers are widely used because of their simplicity and computational efficiency. It uses training methods consisting of relative-frequency estimation of words in a document as words probabilities and uses these probabilities to assign a category to the document. To estimate the term *P(d | c)* where *d* is the document and *c* is the class, Naïve Bayes decomposes it by assuming the features are conditionally independent [12].

4.4 Evaluation Metrics

There are various methods to determine effectiveness; however, precision and recall are the most common in this field. Precision is the percentage of predicted reviews class that is correctly classified. Recall is the percentage of the total reviews for the given class that are correctly classified. We also computed the F-measure, a combined metric that takes both precision and recall into consideration [13].

$$F - measure = \frac{2 * precision * recall}{precision + recall}$$

5 EXPERIMENTAL RESULTS

We have conducted experiments on students' comments on five selected courses. In the preprocessing step we used Stanford tagger [14] for Arabic POS tagging. Also, we used the text transformation operator in Rapidminer from [15] to do the other preprocessing steps (i.e. light stemming, tokenization and vector representations). Evaluation of opinion classification relies on a comparison of results on the same corpus annotated by humans [16]. We evaluated our main steps in our approach which are: feature extraction and opinion classification as follows: In feature extraction, first we manually assigned a feature for each student subjective comments. Then we used our method, described in section 3, to extract features from student reviews. Table3 gives results of the precision, recall and f-measure to evaluate features extraction.

Table 3: Extraction performance of the proposed method

Extraction Precision	83.5%
Extraction Recall	81.3%
Extraction F-measure	82.4%

Then, we manually assigned a sentiment label for each student subjective comments. Also,, we used Rapidminer from [15] as data mining tool to classify and evaluate the results of students' posts. Table 4 gives results of the precision, recall and f-measure for the courses using Naïve bays. Table 4 gives the performance of the evaluation.

Table 4: Polarity of the system

Polarity Precision	77.58
Polarity Recall	79.22
Polarity F-measure	77.83

After that, we grouped the features. Figure 1 gives an example of features extraction for course_1. Figure 2 visualizes the opinion extraction summary as graph.

Course_1:
- *Contain:*
  - *Positive: 15*
  - *Negative: 26*
- *Teacher:*
  - *Positive: 24*
  - *Negative: 20*
- *Exams:*
  - *Positive: 19*
  - *Negative: 20*
- *Marks:*
  - *Positive: 20*
  - *Negative: 7*
- *Books:*
  - *Positive: 12*
  - *Negative: 21*

Figure 1: Feature_ based opinion extraction for course_1

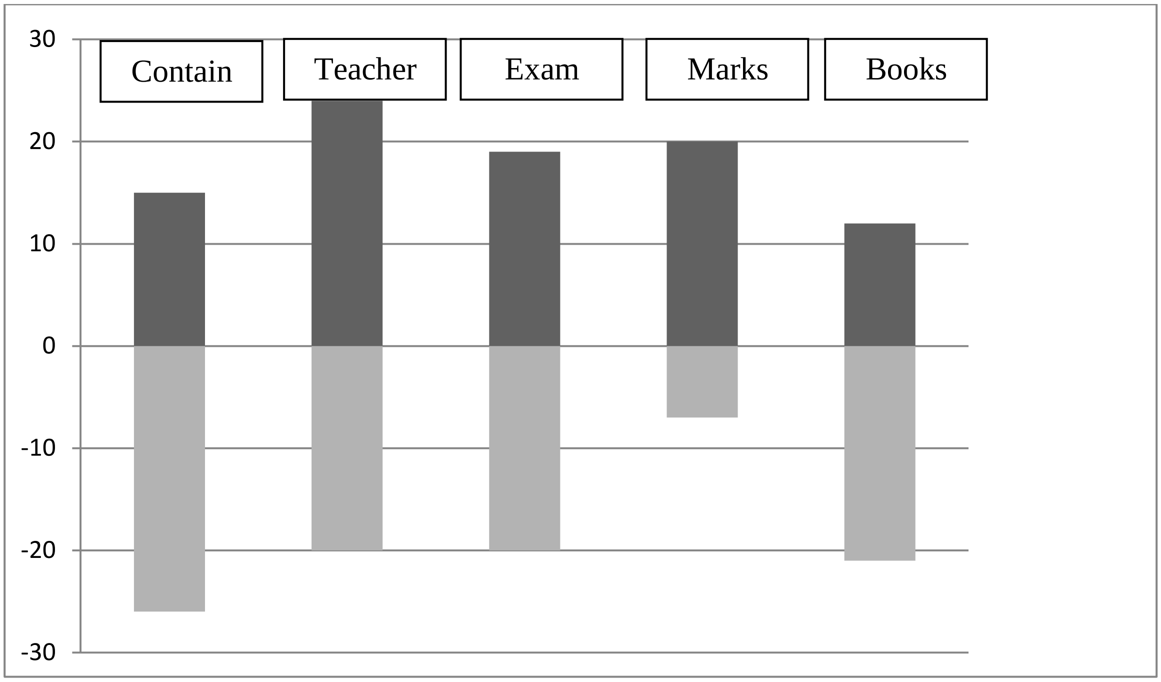

Figure 2: Graph of feature_ based opinion extraction for course_1

In figure 2, it is easy to envisage the positive and negative opinions for each feature. For example, we can figure out that *Books category* has negative attitude while *marks category* has positive attitude from the point of view of the students.

Also, using the visualization we can compare the performance of two courses, for example figure 3 compares the performance of course 1 and course 2.