Mining Opinions in User-Generated Contents to Improve Course Evaluation

Alaa El-Halees

Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
`alhalees@iugaza.edu.ps`

Abstract. The purpose of this paper is to show how opinion mining may offer an alternative way to improve course evaluation using students' attitudes posted on Internet forums, discussion groups and/or blogs, which are collectively called *user-generated content.* We propose a model to mine knowledge from students' opinions to improve teaching effectiveness in academic institutes. Opinion mining is used to evaluate course quality in two steps: opinion classification and opinion extraction. In opinion classification, machine learning methods have been applied to classify an opinion as positive or negative for each student's posts. Then, we used opinion extraction to extract features, such as teacher, exams and resources, from the user-generated content for a specific course. Then we grouped and assigned orientations for each feature.

Keywords: mining student opinions, E-learning evaluation, opinion mining, student evaluation, opinion classification, opinion extraction.

1 Introduction

The increased use of the Internet has changed people's behavior in the way they express their views and opinions. Nowadays, the quality of products and services are often discussed by customers on the Web. Customers can now post reviews of products and services using Internet forums, discussion groups, and blogs which are collectively called *user-generated content* [1] [2]. In recent years, many researchers used opinion mining to extract knowledge from these user-generated contents. Opinion mining is a research subtopic of data mining aiming to automatically obtain useful knowledge in subjective texts [3]. This technique 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. One way to improve teaching quality is to use traditional student evaluations. The

most common criticism of the traditional student evaluations is summarized in three issues; first, it may be biased because students tend to give higher ratings when they expect higher grades in the course [5]. Second, evaluation mainly takes place at the end of a semester. In this case, it is hard to improve the evaluated course and any suggestions will be applied in the subsequent semesters. And, third, regardless of the size of a university, student evaluations generate an enormous quantity of data making the analysis time consuming [6].

To overcome these limitations, we propose to use opinion mining to evaluate course quality. The method can give additional support to traditional student evaluation. There are certain advantages to this method. First, it takes away from traditional classes where most of the time a student takes into account the grade when he/she expresses his/her evaluation. Second, it happens during the semester not at the end, so any recommendations may be taken into account in the same semester not the coming ones. Third, because of the data mining nature which is built for huge data, it is easy to work with an enormous quantity of data generated by students' opinions.

We use opinion mining to evaluate course quality in two steps: opinion classification and opinion extraction. In opinion classification, machine learning methods have been used to classify an opinion as positive or negative for all posts in all courses. Then, in the second step we mined opinions for specific course. In this step, we extracted features for specific courses. Examples of course features contain the teacher and exams. Subsequently, we assigned opinion orientation of the feature (positive or negative). Finally, we grouped the features for each 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 contains opinion classification, section four is about opinion extraction, section five describes the conducted experiments, section six gives the results of experiments and section seven concludes the paper.

2   Related Work

In publications, we found three works that mentioned the idea of using opinion mining in education. First, Lin et al. in [7] 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 [8] 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 [9] 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 Opinion Classification

Opinion mining concerned with enabling system to determine opinions from text written in natural language by human [10]. Opinion classification is a subtopic of opinion mining that classifies an opinionated document as expressing positive or negative. It is also commonly known as sentiment classification or document-level sentiment classification. It aims to find the general sentiment of the author in an opinionated text [4]. For example, in educational data, a student may express his/her opinion about a course using a discussion forum or a blog. Opinion classification determines whether the student attitude is positive or negative about that course.

Formally, given a set of user-generated content review R by students containing opinions about a course, opinion classification aims to classify each document $r \in R$ to determine whether the review is positive, negative or neutral.

We can consider opinion classification as document-level polarity classification which is a special case of text categorization with sentiment positive or negative rather than topic-based categories. With the exception that in traditional document classification, topic words are important. However, with opinion classification, sentiments of the words are more important [11]. Therefore, in opinion classification, we may be able to improve polarity classification by removing objective sentences [10].

In our work, we used modified version of AMOD approach proposed in [12] as follows:

1- A seed set representative of the two adjective categories positive and negative has been provided. Since the work is tested in Arabic language, the seeds are Arabic Adjectives.

2- Synonymy from Online dictionary is used to find new terms that will also be considered representative of the two categories (positive and negative). The new terms, once added to the original ones, yield two new sets of terms.

3- Arabic Datasets, which contained opinion expressions in education, were collected from the Internet.

4- To classify each collected document, we calculate its positive or negative orientation by computing the difference between the number of positive and negative adjectives, from both the previous lists, encountered in the studied document. We count the number of positive adjectives, then the number of negative adjectives, and we simply compute the difference. If the result is positive (greater to given threshold), the document will be classified in the positive class. The same process is done for negative. Otherwise, the document is neutral which is eliminated.

5- We used a binary classifier using the previous documents as training set and user-generated contents as testing set to assign a polarity (e.g. positive or negative) to each student review.

4   Opinion Extraction

Opinion mining discovers opinioned knowledge at different levels such as at clause, feature, sentence or document levels [13]. In the previous section we discussed a way to classify student opinion at document level. This section 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.

For our work, we used Hu, M., Liu approach in [14]. The goal of the approach is to identify, extract and group the features for each course as follows:

1) Identifying and extracting the course features that have been commented by the students. Traditional Information extraction is used where we search as specific course features in the text.
2) Determining whether the opinion on the feature is positive or negative. In this case we used opinion classification discussed in the previous section.
3) Grouping the features orientations and produce a summary for each course and for each feature in that course.

5   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 methods and evaluation metrics.

5.1   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. Then, 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

5.2   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. 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 [15]. We also removed some terms with a low frequency of occurrence.

5.3   Methods

In our experiments to classify posts, we applied three machine learning methods, which are Naïve Bays, k-nearest and Support Vector Machine.

Naïve Bays 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 $P(d \mid c)$ where d is the document and c is the class, Naïve Bayes decomposes it by assuming the features are conditionally independent [16].

k-Nearest Neighbor is a method to classify documents. In the training phase, documents have to be indexed and converted to vector representation. To classify new document d; the similarly of its document vector to each document vector in the training set has to be computed. Then its k nearest neighbor is determined by measuring similarity which may be measured by, for example, the Euclidean distance [17].

Support Vector Machine is a learning algorithm proposed by [18]. In its simplest linear form, it is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. Test documents are classified according to their positions with respect to the hyperplanes.

5.4 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 [19].

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (1)$$

6 Experimental Results

We have conducted experiments on students' comments on five selected courses using two steps: Opinion Classification and Opinion extraction.

First, we evaluated opinion classification. Evaluation of opinion classification relies on a comparison of results on the same corpus annotated by humans [20]. Therefore, to evaluate our approach, first we manually assigned a label for each student subjective comment. Then, we used Rapidminer from [21] as data mining tool to classify and evaluate the results of students' posts. Table 3 gives results of the precision, recall and f-measure for each course using three data mining methods Naïve bays, k-nearest and Support Vector Machine. The last row gives the average results.

Table 3. Precision, recall and F-measure of the five courses using three data mining methods

Course	K-nearest			Naïve Bays			Support Vector Machine		
	Pr	Re	F-m	Pr	Re	F-m	Pr	Re	F-m
Course_1	69.23	75	72	76.7	66.7	71.35	70	72.16	71.06
Course_2	79.09	59.09	72.36	84.29	81.82	83.04	74.74	81.21	77.84
Course_3	58.24	72.23	64.48	67.67	93.94	78.69	61.25	88.15	72.28
Course_4	70	53.85	60.87	72.5	76.92	74.65	69.09	75.52	72.02
Course_5	87.5	82.35	84.85	86.74	76.74	81.43	80	94.12	86.49
Average	70.81	68.50	70.91	77.58	79.22	77.83	71.02	82.23	76.22

From the table, with precision of 77.58 %, we can conclude that Naïve Bays method has better performance than the other two methods. However, with recall of 82.23% Support Vector machine has better performance. Overall, Naïve Bays has the best f-measure with 77.83%.

In the second step, we selected a set of features for course evaluation. From the data, we found that the most frequent features are Teacher, Content, Exams, Marks and Books. We used Gate Information Extraction tool from [22] for feature extractions .Then, we used the system assignment proposed in opinion classification to assign the orientation of the posts. After that, we grouped the features. Figure 1 gives an

Course_1:
    Contain:
            Positive: 18
            Negative: 25
    Teacher:
            Positive: 28
            Negative: 19
    Exams:
            Positive: 18
            Negative: 21
    Marks:
            Positive: 20
            Negative: 6
    Books:
            Positive: 11
            Negative: 21

Fig. 1. Feature_ based opinion extraction for course_1
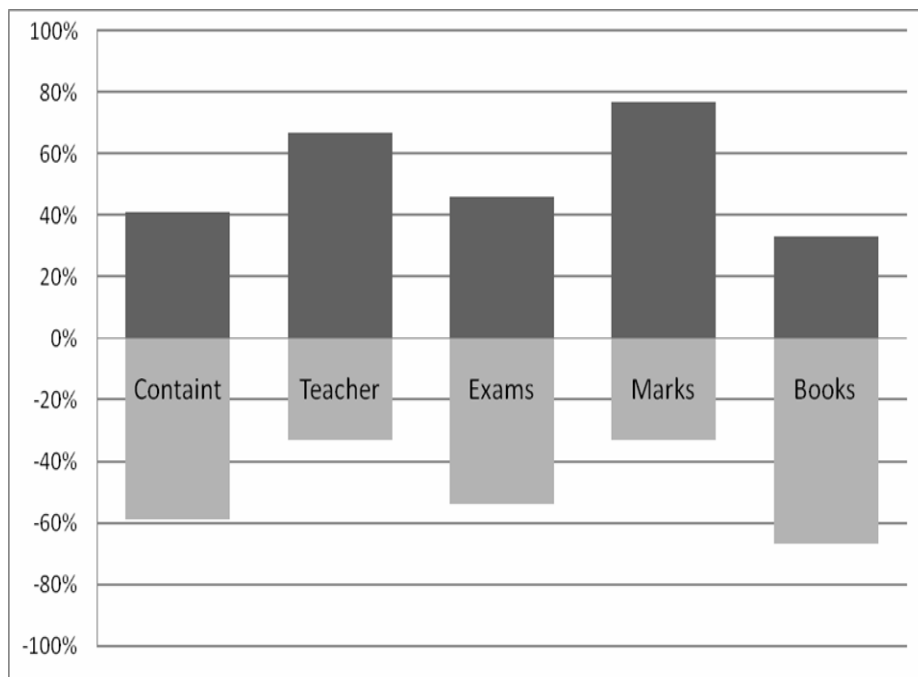

Fig. 2. Graph of feature_ based opinion extraction for course_1

example of features extraction for course_1. Figure 2 visualizes the opinion extraction summary as graph.

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.

7   Conclusion

The aim of this work is to present the usefulness of discovering knowledge from user-*generated content* to improve course performance. Our goal is to supplement student evaluation of courses, not to replace the traditional way of course evaluation. We used opinion mining in two steps: first to classify student posts for courses where we used three machine learning methods. Then, to extract and group features for   each course we used opinion extraction method.

We think this is a promising way of improving course quality. However, two drawbacks should be taken into consideration when using opinion mining methods in this case. First, if the student knew that his posts will be used for evaluation, then he/she will behave in the same way of filling traditional student evaluation forms and no additional knowledge can be found. Second, some students, or even teachers may put spam comment to bias the evaluation. However, for latter problem methods of spam detection, such as work of [23], can be used in future work. Also, comparing courses for each lecturer or semester could be useful for course evaluations.

References

[1] Liu, B.: Searching Opinions in User-Generated Contents. In: Invited talk at the Sixth Annual Emerging Information Technology Conference (EITC 2006), Dallas, Texas, August 10-12 (2006)

[2] Leung, C.W.K., Chan, S.C.F.: Sentiment Analysis of Product Reviews. In: Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining Information Science Reference, 2nd edn., pp. 1794–1799 (August 2008)

[3] Song, H., Yao, T.: Active Learning Based Corpus Annotation. In: IPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, pp. 28–29 (August 2010)

[4] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Information Retrieval 2, 121–135 (2008)

[5] Huemer, M.: Student Evaluations: a Critical Review, `http://home.sprynet.com/~owl/.sef.htm` (accessed in January 2011)

[6] Kozub, R.M.: Student Evaluations Of Faculty: Concerns And Possible Solutions. Journal of College Teaching & Learning 5(11) (November 2008)

[7] Lin, H., Pan, F., Wang, Y., Lv, S., Sun, S.: Affective Computing in E-learning. E-learning, Marina Buzzi, InTech, Publishing (February 2010)

[8] Song, D., Lin, H., Yang, Z.: Opinion Mining in e-Learning. In: IFIP International Conference on Network and Parallel Computing Workshops (2007)

[9] Thomas, E.H., Galambos, N.: What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis. Research in Higher Education 45(3), 251–269 (2004)

[10] Xia, L., Gentile, A.L., Munro, J., Iria, J.: Improving Patient Opinion Mining through Multi-step Classification. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 70–76. Springer, Heidelberg (2009)

[11] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)

[12] Harb, A., Plantié, M., Dray, G.: Web opinion mining: how to extract opinions from blogs? In: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology. ACM, New York (2008)

[13] Balahur, A., Montoyo, A.: A Feature Dependent Method For Opinion Mining and Classification. In: International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE, pp. 1–7 (2008)

[14] Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA (2004)

[15] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)

[16] Du, R., Safavi-Naini, R., Susilo, W.: Web filtering using text classification (2003), http://ro.uow.edu.au/infopapers/166

[17] Dasarathy, B.: Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos (1991)

[18] Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20 (1995)

[19] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA (February 1999)

[20] Osman, D., Yearwood, J.: Opinion search in web logs. In: Proceedings of the Eighteenth Conference on Australasian Database, Ballarat, Victoria, Australia, vol. 63 (2007)

[21] http://www.Rapidi.com

[22] http://gate.ac.uk/

[23] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H.: Detecting Product Review Spammers using Rating Behaviors. In: The 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, October 26 - 30 (2010)