Learning sentiment from students' feedback for real-time interventions in classrooms

Nabeela Altrabsheh, Mihaela Cocea, and Sanaz Fallahkhair

School of Computing, University of Portsmouth
Lion Terrace, Portsmouth, United Kingdom
{nabeela.altrabsheh,mihaela.cocea,sanaz.fallahkhair}@port.ac.uk

Abstract. Knowledge about users sentiments can be used for a variety of adaptation purposes. In the case of teaching, knowledge about students sentiments can be used to address problems like confusion and boredom which affect students engagement. For this purpose, we looked at several methods that could be used for learning sentiment from students feedback. Thus, Naive Bayes, Complement Naive Bayes (CNB), Maximum Entropy and Support Vector Machine (SVM) were trained using real students' feedback. Two classifiers stand out as better at learning sentiment, with SVM resulting in the highest accuracy at 94%, followed by CNB at 84%. We also experimented with the use of the neutral class and the results indicated that, generally, classifiers perform better when the neutral class is excluded.

Keywords: Sentiment Analysis; Educational Data Mining; Students Feedback

1 Introduction

Students feedback can help the lecturers understand their students learning behaviour [5] and improve teaching [19]. Taking feedback can highlight different issues that the student may have with the lecture. One example of this is when the student does not understand part of the lecture or a specific example. Another example is when the lecturers' teaching pace is too fast or too slow. Feedback is usually collected at the end of the unit, but it is more beneficial taken in real-time.

Collecting feedback in real-time has numerous benefits for the lecturer and their students, such as improvement in teaching [19] and understanding students' learning behaviour [5]. Moreover, students' feedback improves communication between the lecturer and the students [5], allowing the lecturer to have an overall summary of the students opinion.

One way of collecting feedback in real-time is using Student Response Systems (SRS) which is a term used for devices that collect real-time data from students. Clickers, mobile phones and social media are types of SRS that have been used in the past to collect feedback in real time. Despite their usefulness in collecting real-time feedback, SRS systems can not be used to their full advantage without support for the analysis of the collected data. For example, in

a study using Twitter to collect feedback, the lecturer had to read through all the students' tweets sequentially [16]; therefore, the lecturer had to read from the beginning to understand the tweets, causing time loss. Furthermore, other research showed concern that using this tool will put such an additional workload onto the lecturers that they would require additional training to effectively use the tweets as feedback [26].

To address this problem we propose the creation of a system that will automatically analyse students' feedback in real-time and present them to the lecturer. The system will be trained offline, to insure there will be no delay in presenting the results to the lecturer. The system will visualise the students' feedback in a meaningful way giving the lecturer the most important information from the feedback. To analyse the students' feedback we propose the use of sentiment analysis.

Sentiment analysis, an application of natural language processing, computational linguistics and text analytics, identifies and retrieves information from the text. Sentiment analysis can be applied to general data, although it is more effective when applied to specific domains [23] because word meanings and sentiment may differ across domains. An example for this is the word 'early' which may reflect negative subjectivity in education as in the instance "The lecture is too early!". Then again, when describing a parcel service such as "The parcel arrived early", this is most likely a positive sentiment.

One scenario that shows the benefits of the system is described in the following. Rob, a lecturer, has just finished presenting an example, and he wanted to know whether to move on to the next part of the lecture. He looks at the visualisation provided by our system, illustrating different proportions of positive, negative and neutral sentiment. He cans also see frequent words with their polarity. He found words such as 'example', 'confused','complicated' and 'lost' show on the screen with the negative polarity. He then looks at the percentage of negative feedback, which is 60 percent of the class, the neutral is 30 percent and the positive is 10 percent. He then decided to explain the example in a different way.

To insure that the system delivers optimal results, there is need of studying and designing sentiment analysis models that are trained with real students feedback. In this paper we focus on assessing the ability of several machine learning techniques to learn sentiment from students' textual feedback. Consequently we trained four models, i.e. Naive Bayes, Complement Naive Bayes, Maximum Entropy and Support Vector Machine, and compared their performance.

The rest of the paper is organised as follows. Related research is presented in section 2. The data corpus used for this study is presented in Section 3. The sentiment analysis models are presented in Section 4, followed by results and discussion in Section 5. To finish, conclusions and future work are outlined in Section 6.

2 Related Work

Sentiment analysis looks at the polarity of sentiment. In most cases, researchers are interested in the positive and negative sentiments, although some researchers advocate the use of a neutral category as well.

Agarwal et al. [1] investigated the contribution of the neutral class to the performance of classifiers by comparing 2-class (Positive/Negative) models with 3-class (Positive/Negative/Neutral) models. They found that the 2-class models have higher accuracy; however, other researchers obtained a good performance for the 3-class models [3].

We believe that a neutral class is needed in a real life applications for education, while acknowledging that enough training data labelled as neutral is necessary to get good results. For this paper, we used the method proposed by Agarwal et al. [1], comparing the model with and without the neutral class, to investigate the effect of the neutral class on the performance of classifiers in the educational domain.

There have been some studies about sentiment analysis for education, however they have been focused mainly on e-learning [17, 24], with some exceptions looking at classroom learning. For example, in Munezero [15], sentiment analysis was applied in-class to detect students' emotions from students' learning diaries. This work, however, did no look at real-time interventions based on the analysed feedback.

Machine learning sentiment analysis approaches have four main steps: collecting the data, preprocessing it, selecting the features and applying the machine learning techniques. These are reviewed in the following subsections.

2.1 Preprocessing

Preprocessing is the process of cleaning the data from noise such as removing special characters. It increases the accuracy of the results by reducing errors in the data [2]. There are different types of preprocessing used according to where the data is collected from; for example, data collected from Twitter needs extra preprocessing such as removing hashtags, retweets and links. Some of the most common general preprocessing techniques [20] that will be used in our experiments are:

- Remove stop words: Removal of the stop words will help reduce index space, improve response time, and improve effectiveness. There is not one set of stop words that can be removed. Stop words can be words such as 'a', 'the' and 'there'. More examples can be found in [20];
- Remove punctuation: Removal of punctuation, such as question and exclamation marks, has been applied by Prasad [20]. However, exclamation marks can indicate the presence of emotion such as in the sentence 'I passed!!!'; here the exclamation may mean strong positive sentiment or strong joy emotion. Moreover, the question mark may represent confusion. Most of the researchers removed numbers and punctuation, therefore for this paper we decided to eliminate it; however, it will be investigated in the future;

- Remove numbers: Numbers in chat language can represent words; for example, 'to' or 'too' can be written as '2', 'for' can be written as '4' and the word 'great' can be written in chat language as 'gr8'. However, most of the time numbers do not have meaning by themselves and are irrelevant in sentiment analysis. In most sentiment analysis research numbers are removed; therefore, we also remove them for our experiments;
- Convert text to lower or upper case: converting the letters into upper or lower case is used to match occurrences in the training data. Words in capitals sometimes suggests strong emotion [20]. However, this will not be investigated in the experiments presented in this paper;
- Spelling check/Removing repeated letters: Spelling can be corrected by removing extra letters such as in Prasad [20] and Ortigosa et al. [17]. Go et al. [9] replaced the letters with two letters. However, Agarwal et al. [1] replaced the letters with three letters. In our research, we replaced repeated letters with two letters, as most words in English have a maximum of two repeated letters. However, this affects words which should be single letters, such as 'looooove', which will become 'loove' and, therefore, it will not be matched to other occurrences of 'love'. On the other hand, it covers most common situations, rather than exceptions.

2.2 Features

Feature selection allows a more accurate analysis of the sentiments and detailed summarization of the results. One of the most common feature is n-grams [1, 9]. An n-gram is a sequence of n items from a text. It can be letters, syllables, or words. The most common n-gram is unigram which is selecting single words, as found in many research works [1, 8, 27]. Consequently, for the purpose of this paper, unigrams alone will be experimented with.

2.3 Machine Learning Techniques

In the educational domain, Tan et al. [23] and Troussas et al. [25] found Naive Bayes to be the best technique, while Song et al. [22] used Support Vector Machines. This research indicates that different machine learning techniques give different results even for the same domain, prompting a need for testing several techniques. The techniques used in our experiments are Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME), and Support Vector Machines (SVM), due to their popularity and high results in previous research.

3 Data Corpus

We used two methods for data collection: real-time collection of feedback in lectures and end of unit feedback. The first method we used is real-time feedback

from computing lectures at the University of Portsmouth. The lectures included postgraduate and undergraduate level students.

Due to the difficult circumstances in collecting real-time students' feedback, we collected end of unit student feedback from various institutes. The total amount of data is 1036 instances, as shown in Table 1.

Table 1. Data Sources

Data Source	Number of instances
End of Unit Other Institutes	768
End of Unit University of Portsmouth	117
Real-time feedback University of Portsmouth	190

The data was labelled by three experts, one with background in data mining (including sentiment analysis) and two with background in linguistics. The labels were assigned using a majority rule. When there was no majority, the neutral label was assigned. To verify the reliability of the labels provided by the three experts, we looked at inter-rater reliability. The percent agreement was 80.6%, the Fleiss kappa [7] was 0.625 and Krippendorff's alpha [12] was 0.626. The percent agreement is considered over-optimistic, while the other two measures are known to be more conservative [14].

Table 2. Distribution of sentiment labels in our corpus

	Positive	Negative	Neutral
Frequency	641	292	103

4 Learning sentiment from students' feedback

Using the 1036 labelled instances, we investigated the learning performance of different machine learning techniques: Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME) and Support Vector Machines (SVM). These methods are briefly described in the following subsections.

4.1 Naive Bayes and Complement Naive Bayes

Naive Bayes is a classifier that uses a probabilistic model; its origin is from Bayes theorem, which assumes independence between features. It has been found to perform well for sentiment analysis, e.g., [18,20]. Some advantages of Naive Bayes are that it only needs a small amount of training data to estimate parameters, it is fast and incremental, and can deal with discrete and continuous attributes.

Naive Bayes does not work well with uneven class sets. Complement Naive Bayes addresses this problem and has been proven to give higher results than Naive Bayes when the classes are uneven [10]. Complement Naive Bayes estimates the probability of a class using parameters of all the classes excluding the

class itself. The NB algorithm was implemented in R[1], while for CNB Weka [29] was used.

4.2 Maximum Entropy

The Maximum Entropy classifier is similar to the Naive Bayes classifier, except that instead of the features acting independently, the model finds weights for the features that maximize the likelihood of the training data using search-based optimization. One advantage of maximum entropy is that it does not make assumptions about the relationships between features; consequently, in contrast to Naive Bayes and SVM, it could potentially perform better when conditional independence assumptions are not met.

One drawback is that it is not very realistic in many practical problems, as real datasets contain random errors or noise which create a less clean dataset [11]. For our experiments with Maximum Entropy, we use the R maxent package [4].

4.3 Support Vector Machines (SVM)

A Support Vector Machine (SVM) is a non-probabilistic binary linear classifier. It finds hyperplanes that separate the classes. SVM is highly effective at traditional text categorization and generally outperforms Naive Bayes. SVM is effective in high dimensional spaces and when the number of dimensions is greater than the number of samples. Moreover, it is memory efficient.

The effectiveness of the SVM can be affected by the kernel [6]. There are different types of kernels, of which the most common kernel methods are linear, polynomial, and radial basis functions. The linear kernel results in a simple classifier. It can work best with larger amounts of data and is graphed as a straight line. Non-linear kernels are more flexible and often give better performance [6]. From non-linear kernels, most common are polynomial kernel (SVM Poly) and radial basis kernel (SVM RB). The polynomial kernel works well with natural language processing [6] and is usually presented as a curved line. The radial basis kernel is popular [6] and flexible, and is graphed as a curved path. LibSVM in Weka [29] was used for our experiments.

5 Results and Discussion

To test the learning performance of the four models, 10-fold cross-validation was used. The results are displayed in Table 3, which includes the performance of the models without the neutral class, i.e., positive and negative, and with the use of the neutral class, i.e., positive, negative and neutral.

From the results we observe the following:

1. Two methods have a very good performance in terms of accuracy, precision and recall: Support Vector Machine with radial basis kernel and Complement Naive Bayes models;

[1] `http://www.r-project.org/`

Table 3. Experiment results - without (W/O) and with (Neu) the neutral class.

	Naive Bayes		CNB		ME		SVM Linear		SVM Poly		SVM RB	
	W/O	Neu	W/O	Neu	W/O	Neu	W/O	Neu	W/O	Neu	W/O	Neu
Accuracy	0.50	0.55	0.84	0.80	0.57	0.63	0.69	0.62	0.68	0.61	0.94	0.93
Precision	0.49	0.32	0.87	0.84	0.33	0.33	0.74	0.66	0.47	0.35	0.94	0.93
Recall	0.49	0.31	0.84	0.80	0.30	0.33	0.69	0.62	0.68	0.61	0.94	0.93
F-Score	0.48	0.28	0.84	0.81	0.31	0.32	0.57	0.48	0.56	0.47	0.94	0.92

2. Precision and recall are high in both Support Vector Machine and Complement Naive Bayes models, but low in Naive Bayes and Maximum Entropy models.
3. Naive Bayes has a relatively poor performance despite being considered a good learning method for sentiment analysis;
4. When the neutral class is considered, performance decreases for most metrics and classifiers.

Our results show that SVM gave the highest accuracy, as opposed to the research of Ortigosa et al. [17] in the educational domain, and more specifically, e-learning. This could be due to the use of unigrams as opposed to Ortigosa et al. [17] who used pos (part of speech)-tagging as a feature.

Although our data is relatively clean, the Naive Bayes classifier had the lowest performance. This may be due to the uneven class sets, which could explain why the Complement Naive Bayes classifier had a high performance.

The recall values show that SVM RB is the most sensitive of the four models, i.e., it correctly identifies instances of all classes, while the Maximum Entropy is the least sensitive. Precision is highest for SVM RB and lowest for Naive Bayes with the neutral class. The best balance between precision and recall is achieved by SVM RB, making it the best classifier. This balance is also present for CNB as well, which is the second best performing method.

To investigate if the classifiers results are significantly different when the neutral class is not used compared with when the neutral class is used, we used two statistical tests: the paired t-test and the binomial test. The t-test is widely used for testing statistical differences on data mining methods [29]; however, some authors argue that it is not the best test for comparing the performance of different algorithms on the same data set and propose the use of the binomial test [21]. Consequently, we report the results for both of these tests, which are displayed in Table 4, where the t-test is represented as A and the binomial test as B. The significant values are marked in bold.

The significance tests show that the classifiers perform significantly better in terms of accuracy when the neutral class is not used for CNB and SVM with polynomial and radial basis kernels. Precision is significantly better when the neutral class is excluded for NB, CNB (just for the t-test; the binomial test indicated that the difference in not significant) and SVM with polynomial kernel. Recall is significantly better without the neutral class for NB, CNB and all versions of SVM; however, for SVM with linear kernel the t-test results indicate

Table 4. Level of significance (p-values) for differences between classifiers with the neutral class and without it.

	Naive Bayes		CNB		ME		SVM linear		SVM Poly		SVM RB	
	A	B	A	B	A	B	A	B	A	B	A	B
Accuracy	**0.01**	**0.00**	**0.00**	**0.02**	**0.00**	**0.00**	0.05	0.10	**0.00**	**0.00**	**0.00**	**0.00**
Precision	**0.00**	**0.00**	**0.01**	0.10	0.91	1.00	0.05	0.34	**0.00**	**0.00**	0.106	0.109
Recall	**0.00**	**0.00**	**0.00**	**0.00**	0.07	0.10	**0.01**	0.10	**0.00**	**0.00**	**0.00**	**0.00**
F-score	**0.00**	**0.00**	**0.00**	0.10	0.26	0.75	0.06	0.34	**0.00**	**0.00**	**0.00**	**0.00**

that the difference is significant, while the binomial test indicates that it is not. Finally, the F-scores are significantly better when the neutral class is excluded for NB, CNB (t-test only), SVM Poly and SVM RB.

The ME classifier performs significantly better when the neutral class is used in terms of accuracy, but with no significant difference in terms of precision, recall and F-score. The NB classifier has a significantly better accuracy when the neutral class is used, but significantly lower precision, recall and F-score.

Consequently, for most classifiers the evaluation metrics, i.e. accuracy, precision, recall and F-score, improve when the neutral class is not used. This may be due to the low number of training instances for the neutral class, i.e. 103 out of 1036, an aspect that has been pointed out in previous research, e.g., [27].

Given the results of our classifiers, arguments can be found for both using and disregarding the neutral class. On one hand, ignoring the neutral class seems to be consistent with people's tendency to give their opinions when they feel stronger about them, i.e., positive or negative, rather than when they do not have a particular view on the subject, i.e., neutral; consequently opinion mining often does not consider the neutral class as it is viewed as absence of opinion, e.g., [28].

Using the neutral class, on the other hand, may prevent problems such as overfitting [13]. It also provides a more complete picture of the data, where lack of sentiment is still important to be considered [13] as much as the positive and negative classes are.

Consequently, for our purposes, we will continue to investigate the use of neutral class for the educational domain, not just in terms of performance of classifiers, but also from the point of view of the users, i.e. lecturers, with regard to the usefulness of knowing how many students have a neutral view with regards to their teaching.

When looking at the t-test and the binomial test results, there is a disagreement between these tests only in 3 instances out of 48, with the t-test indicating a significant difference, while for the binomial test the difference in not significant.

6 Conclusions and Future work

In this paper we investigated the learning capabilities of four machine learning methods for learning sentiment from students' textual feedback: Naive Bayes,

Complement Naive Bayes, Maximum Entropy and Support Vector Machines (with three types of kernel).

A dataset of 1036 instances of teaching-related feedback was used, which was labelled by 3 experts. We experimented with the use of unigrams as features and a range of standard preprocessing techniques. Our experiments indicate that two methods in particular, i.e. SVM with radial basis kernel and CNB, give very good results; therefore, they could be used for real-time feedback analysis.

We also explored the use of the neutral class in the models and found that, in most cases, performance is better when the neutral class in not used. There are, however, arguments for using a neutral class from practical point of view, as it provides a more complete picture of a situation. Moreover, for the best performing method, i.e. SVM with radial basis kernel, the difference between using the neutral class and not using it, is 0.01 for accuracy, precision and recall. Consequently, one can argue that such a small loss is acceptable for having a more complete picture.

Future work includes an analysis of more preprocessing techniques and their impact on model performance, as well as experimentation with other features, such as bigrams, trigrams and pos(part of speech)-tagging. In addition, we will test the models using more real-time collected data.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media. pp. 30–38. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
2. Altrabsheh, N., Gaber, M., Cocea, M.: SA-E: Sentiment Analysis for Education. International Conference on Intelligent Decision Technologies 255, 353–362 (2013)
3. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. International Conference on Computational Linguistics 23, 36–44 (2010)
4. Bhargavi, P., Jyothi, S.: Applying naive bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security 9(8), 117–122 (2009)
5. Calders, T., Pechenizkiy, M.: Introduction to the special section on educational data mining. SIGKDD Explorations 13(2), 3–6 (May 2012)
6. Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J.: Training and testing low-degree polynomial data mappings via linear svm. Journal of Machine Learning Research 11, 1471–1490 (Aug 2010)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin 76(5), 378 (1971)
8. Gamon, M.: Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. International Conference on Computational Linguistics 20, 841–847 (2004)
9. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. CS224N Project Report, Stanford (2009), `http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf`
10. Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A.: Opinion mining and sentiment analysis on a twitter data stream. Advances in ICT for Emerging Regions (ICTer) 13, 182–188 (Dec 2012)

11. de Groot, R.: Data mining for tweet sentiment classification. Master's thesis, Utrecht Univerity (2012)
12. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 1(1), 77–89 (2007)
13. Koppel, M., Schler, J.: The importance of neutral examples for learning sentiment. Computational Intelligence 22(2), 100–109 (2006)
14. Lombard, M., Snyder-Duch, J., Bracken, C.C.: Practical resources for assessing and reporting intercoder reliability in content analysis research projects (2004), http://astro.temple.edu/~lombard/reliability/
15. Munezero, M., Montero, C.S., Mozgovoy, M., Sutinen, E.: Exploiting sentiment analysis to track emotions in students' learning diaries. Koli Calling International Conference on computing Education Research 13, 145–152 (2013)
16. Novak, J., Cowling, M.: The implementation of social networking as a tool for improving student participation in the classroom. ISANA International Academy Association Conference Proceedings 22, 1–10 (2011)
17. Ortigosa, A., Martin, J.M., Carro, R.M.: Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior 31, 527 – 541 (2014)
18. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Annual Meeting on Association for Computational Linguistics 42, 271–278 (2004)
19. Poulos, A., Mahony, M.J.: Effectiveness of feedback: the students perspective. Assessment & Evaluation in Higher Education 33(2), 143–154 (2008)
20. Prasad, S.: Micro-blogging sentiment analysis using bayesian classification methods. CS224N Project Report, Stanford (2010), http://nlp.stanford.edu/courses/cs224n/2010/reports/suhaasp.pdf
21. Salzberg, S.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1(3), 317–328 (1997)
22. Song, D., Lin, H., Yang, Z.: Opinion mining in e-learning system. International Conference on Network and Parallel Computing Workshops 6, 788–792 (Sept 2007)
23. Tan, S., Cheng, X., Wang, Y., Xu, H.: Adapting naive bayes to domain adaptation for sentiment analysis. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 5478, pp. 337–349. Springer Berlin Heidelberg (2009)
24. Tian, F., Zheng, Q., Zhao, R., Chen, T., Jia, X.: Can e-learner's emotion be recognized from interactive Chinese texts? International Conference on Computer Supported Cooperative Work in Design 13, 546–551 (April 2009)
25. Troussas, C., Virvou, M., Junshean Espinosa, K., Llaguno, K., Caro, J.: Sentiment analysis of facebook statuses using naive bayes classifier for language learning. Information, Intelligence, Systems and Applications (IISA) 4, 1–6 (July 2013)
26. Vohra, M.S., Teraiya, J.: Applications and challenges for sentiment analysis: A survey. International Journal of Engineering 2(2), 1–5 (2013)
27. Wang, W., Wu, J.: Emotion recognition based on cso&svm in e-learning. International Conference on Natural Computation (ICNC) 7, 566–570 (2011)
28. Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? finding strong and weak opinion clauses. In: Proceedings of the 19th National Conference on Artifical Intelligence. pp. 761–767. AAAI Press (2004)
29. Witten, I.H., Frank, E., Mark, A.H.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2011)