Sentiment analysis: towards a tool for analysing real-time students feedback

Nabeela Altrabsheh
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
Email: nabeela.altrabsheh@port.ac.uk

Mihaela Cocea
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
Email: mihaela.cocea@port.ac.uk

Sanaz Fallahkhair
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
Email: sanaz.fallahkhair@port.ac.uk

***Abstract*—Students' real-time feedback has numerous advantages in education, however, analysing feedback while teaching is both stressful and time consuming. To address this problem, we propose to analyse feedback automatically using sentiment analysis. Sentiment analysis is domain dependent and although it has been applied to the educational domain before, it has not been previously used for real-time feedback. To find the best model for automatic analysis we look at four aspects: preprocessing, features, machine learning techniques and the use of the neutral class. We found that the highest result for the four aspects is Support Vector Machines (SVM) with the highest level of preprocessing, unigrams and no neutral class, which gave a 95 percent accuracy.**

***Keywords*-Sentiment Analysis, Educational Data Mining, Feature Selection, Real-time Feedback**

I. Introduction

Students' feedback can highlight different issues students may have with a lecture. An example of this is the students not understanding a specific example. Analysing feedback in real-time, however, is time consuming and stressful [1]. We propose to address this problem by creating a system that analyses students' feedback in real-time and then presents the results to the lecturer. To create such as system, sentiment analysis can be used.

Sentiment analysis is an application of natural language processing, computational linguistics and text analytics that identifies and retrieves sentiment polarity from the text by studying the opinion. Sentiment polarity is usually either positive or negative, although sometimes neutral is included.

Previous research has shown that sentiment analysis is more effective when applied to specific domains [2]. Sentiment analysis in the educational domain has mainly been focused on e-learning [3], [4], with little research done on classroom feedback [5]. Although e-learning and classroom education may seem similar, they differ in the types of interactions between the students and the lecturers and in the fact that the lecturer should respond to the students' feedback in real-time. Feedback from students in the classroom settings is different from distant learners due to the different situations and issues students may have. E-learning students can have issues such as lack of interaction. However, classroom feedback can be about class settings, e.g., class being hot. When training a model, words need to be related to the purpose of the application, and in our case, the model needs to be trained using classroom feedback.

To the best of our knowledge, sentiment analysis has not been applied for analysing students' classroom feedback before. Consequently, there is need of investigating different models and look at the best combination of preprocessing methods, features and machine learning techniques to create the best-suited model for our purpose. In this paper, we investigate the following aspects:

- what preprocessing techniques are the most effective?
- which feature(s) are most suitable for this data?
- which machine learning techniques give the highest performance?
- what is the effect of including the neutral class in the models?

Related research about data collection, preprocessing the data, feature selection and machine learning techniques is presented in section II. The data corpus that is used for this study is presented in section III. Our sentiment analysis models are presented in section IV, followed by results and discussion in section V. Conclusions and future work are outlined in section VI.

II. Related Research

There are four main steps in machine learning sentiment analysis: collecting the data, preprocessing it, selecting the features and applying the machine learning techniques. An overview of previous research related to these aspects is given in the following subsections.

A. Preprocessing

Sentiment analysis can be improved by preprocessing the data. Preprocessing is the process of cleaning the data from unwanted elements. It increases the accuracy of the results by reducing errors in the data.

Not using preprocessing, such as spelling corrections, may lead to the system ignoring important words. However, using preprocessing techniques wrongly sometimes may cause loss of important data – for instance the removal of punctuation, when it may add extra value to the sentiment.

There are many general preprocessing techniques, of which the most common are: remove stop words, remove

punctuation, remove number, covert text to lower or upper case, and removing repeated letters.

In the educational domain, most of the researchers have collected data from social media networks such as Facebook and Twitter. A common preprocessing technique is identifying emoticons, which is found in Troussas et al. [6]. Another preprocessing technique used widely with educational domains is the spelling check; this is found in Ortigosa et al. [4] and Martin et al. [7]. Most of the researchers have used the general preprocessing methods listed above as well.

B. Features

Features allow a more accurate analysis of the sentiments and detailed summarization of the results. One of the most commonly used feature is n-grams [2], [8], [9]. An n-gram is a sequence of n items from a text. Items can be letters, syllables, or words. N-grams are mostly based on words and the most common n-gram types are unigram (one word), bigram (two words), and trigram (three words).

There is no clear answer whether unigrams give better performance that bigrams and trigrams, or vice versa. Some researchers have had better performance using unigrams than bigrams and trigrams [8], [9]. For example, Go et al. [2] found that bigrams decreased performance. Similar results were found by several researchers. Oppositely, Pak and Paroubek [10] research about general Twitter analysis, found that bigrams give a higher accuracy than unigrams.

In the educational domain, some of the researchers used n-grams, such as Troussas et al. [6]. However, some of the researchers have used pos-tagging as a feature, such as in Ortigosa et al. [4] and Martin et al. [7].

The mixed picture about the contribution of different n-grams to model performance indicates the need to explore all the combinations between are unigrams, bigrams, and trigrams for our purpose.

C. Machine Learning Techniques

There are many techniques that have been used, of which the most common ones are Naive Bayes [11], [2], Maximum Entropy [2], and Support Vector Machines [2]. Naive Bayes does not work well with uneven class sizes, however Complement Naive Bayes can address this problem by estimating parameters from all the data in all sentiment classes except the class required.

In the education domain, Troussas et al. [6] found Naive Bayes to be the best technique, while Song et al. [12] found Support Vector Machines to be the best. Therefore, different machine learning techniques give different results even for the same domain, prompting a need for testing several techniques.

1) Neutral Class: Some researchers included a neutral class in their models, for example [2], [8], [9]. Others dismissed the neutral class due to the lack of neutral training instances in the data and the poor performances that it led to [10] and [9].

It has been argued that the neutral class is needed in real life applications [2], [9], including education [4], [7]. Students may have positive, negative or neutral opinions. Discarding the neutral class and focusing on only on positive and negative opinions does not show a complete picture of the class opinion and may distort the true proportions of the positive and negative opinions when the neutral proportion is not known. Many of the researchers have included the neutral class when analysing data from the educational domain [4], [7]. Consequently, the role of neutral class deserves further investigation.

III. Data Corpus

We used several methods for collecting the data for our experiments. We collected real-time feedback from lectures in the computing department at the University of Portsmouth. This included postgraduate and undergraduate students. The students were asked to submit in their own words, their feedback, opinions, and feelings about the lecture.

The second method used to collect feedback was end of unit student feedback from various institutes. The amount of data that was collected in total was 1036, one from each student. The data is distributed as 641 positive, 292 negative, and 103 neutral.

The data was labelled by three experts, of which two were experts in linguistics and one in sentiment analysis. 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 inter-rater reliability was calculated. The percent agreement was 80.6%, the Fleiss kappa was 0.625 and Krippendor's alpha was 0.626. The percent agreement is considered over-optimistic, while the other two measures are known to be more conservative.

IV. Sentiment Analysis for Students' Feedback

Our dataset is a relatively clean dataset as it is mostly collected from paper so there is a need for a relatively low level of preprocessing. To find the optimal level of preprocessing the data and to insure that the data was not over preprocessed, i.e. eliminating information that bears sentiment, we tested four preprocessing levels using different techniques which were chosen due to their popularity in previous studies:

1) Preprocessing P1: This is to test how well the model works without any preprocessing apart from converting the letters into lowercase. We included this to explore the role of different degrees of preprocessing, in which this one (P1) is the baseline;
2) Preprocessing P2: This is to remove numbers, punctuation, spaces and blanks, and special characters. In most cases these do not hold value or sentiment, therefore they are noise to the data;

Table I
EXPERIMENT RESULTS-HIGHEST FOR EACH MODEL

	Naive Bayes		CNB		ME		SVM-linear		SVM-radial basis		SVM-polynomial	
	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral
Best N-gram	UNI+BI	TRI	UNI	UNI+BI	TRI	TRI	UNI	UNI	BI+TRI	UNI+BI+TRI	UNI	UNI
Preprocessing	P1	P4	P4	P1	P4	P4	P4	P4	P3	P1	P4	P4
Accuracy	0.517	0.597	0.842	0.863	0.717	0.683	0.945	0.930	0.888	0.689	0.687	0.619
Precision	0.526	0.338	0.878	0.865	0.342	0.320	0.947	0.932	0.900	0.734	0.472	0.384
Recall	0.521	0.332	0.842	0.864	0.407	0.323	0.945	0.930	0.889	0.666	0.687	0.619
F-Score	0.499	0.335	0.848	0.859	0.372	0.308	0.944	0.929	0.882	0.568	0.560	0.474

Table II
PREPROCESSING DIFFERENCES BETWEEN THE FOUR LEVELS OF PREPROCESSING

	NB		CNB		ME		SVM-linear		SVM-radial basis		SVM-polynomial	
	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral	W/O	Neutral
Best N-gram	UNI+BI	TRI	UNI	UNI+BI	TRI	TRI	UNI	UNI	BI+TRI	UNI+BI+TRI	UNI	UNI
P2-P1	-0.020	-0.005	0.001	-0.450	-0.004	-0.003	-0.003	-0.008	-0.049	-0.023	0.000	0.000
P3-P2	-0.007	0.000	0.001	0.200	0.004	0.003	0.007	0.010	0.200	-0.046	0.000	0.000
P4-P3	0.020	0.060	0.002	-0.042	0.019	0.000	0.000	0.001	-0.200	0.000	0.000	0.000

3) Preprocessing P3: This includes all of preprocessing P2 with an additional two techniques which are replacing n't with not and removing repeated letters. These two are a fairly important step to preprocess the data as it increases the probability of matching words to their right sentiment; and
4) Preprocessing P4: This includes all of preprocessing P3 plus the removal of stop words. This step is a final step that removes all words that are irrelevant to the analysis.

Like most researchers, we focus on n-grams. The features that were experimented with are: Unigrams, Bigrams, Trigrams, Unigrams and Bigrams combined, Unigrams and Trigrams combined, Bigrams and Trigrams combined, and Unigrams, Bigrams, and Trigrams combined.

Machine learning techniques were selected next. The techniques used in our experiments are Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME) and Support Vector Machines (SVM). NB, ME and SVM were chosen due to their popularity and high performance in previous research in sentiment analysis, e.g., [13]. Complement Naive Bayes (CNB) is rarely used in sentiment analysis; however, it was used to test its potential in solving the uneven class problem for our application. We investigate the combination of preprocessing methods, features, machine learning techniques and the use of the neutral class. The results of these experiments are explained in the following section.

V. RESULTS AND DISCUSSION

All the models were tested using 10-fold cross-validation; accuracy, precision, recall, and F-score were calculated. The non-neutral model is abbreviated by W/O. The preprocessing levels P1 to P4 correspond to the descriptions in section IV. The features are abbreviated by: UNI: unigrams; BI: bigrams; and TRI: trigrams.

The highest results are presented in Table I. From the results, we observe the following:

- Nearly all models performed better when preprocessing was applied, which was expected. There are however some interesting exceptions which are discussed below;
- Unigrams gave a high performance in several models; unigrams combined with bigrams performed well for CNB, and trigrams performs relatively well with ME;
- All the methods except NB have relatively high accuracy, with the SVM linear kernel having the best performance at 95% and SVM radial basis kernel the second best at 88%;
- Precision, recall, and F-score are high in both SVM and CNB models but low in NB and ME models; and
- SVM and CNB have a good performance when the neutral class is considered.

A. Preprocessing results

In relation to preprocessing the text, most of the models showed higher accuracy when using the highest level of preprocessing. However, in NB and CNB where the highest performance was found in unigrams with bigrams, using no preprocessing gave the highest results. This could be due to the numbers and punctuation having sentiment – for instance the bigram '2 hours' was negative as many of the students disliked long lectures.

The difference between each preprocessing level was calculated for the highest results in each model. Table II shows these results.

There are some models where the difference is zero, indicating no improvement in performance when increased levels of preprocessing are used. Some results show negative

values meaning that increasing the level of preprocessing made the results worse. This is found in nearly all the models when calculating the difference between P1 and P2.

For the CNB model when the neutral class is considered, P1 performs the best, which could mean that preprocessing removes information that is valuable for this model.

When calculating the difference between P2 and P3, there is an increment in the results by up to 20 percent in some models. This means that removing repeated letters and replacing 'n't' with 'not' is valuable in these models; this could be because students write informally and sometimes in chat language to give feedback. Repeated letters and negation exists in nearly 10 percent of the raw data; consequently, preprocessing these aspects improves prediction.

Although P3 and P4 have just a slight difference, which is removing stop words, one of the models showed an increment of up to 6 percent. This model used trigrams, in which stop words are common and, therefore, valuable for prediction.

For the best preprocessing combination, we found that using all the preprocessing techniques gives a higher accuracy in most of the models; however, in CNB using all preprocessing decreased the performance. Therefore, we will continue to experiment with both P1 and P4 for our future experiments, and look at the contribution of numbers, punctuation and negation in carrying sentiment.

B. Features results

Unigrams performed well in most of the best performing models. Some examples of frequent unigrams that are common in our dataset are 'technology', 'blackboard', 'examples', 'helpful', 'learning', 'questions', 'knowledge', 'topics', 'notes', 'lecturer' and 'subject'. Similarly, previous research also found unigrams to be the best feature [8], [9].

Trigrams performed well with the ME model. This indicates that certain 3-word combinations have a high informative value for our particular application. Some examples of three word combinations that appeared often are "how to use", "I have learnt", "learnt lots of", "understand the content", "use the technology", "learn to use" and "maths is fun". In opposition, research done by Wang and Wu [9] about general Twitter sentiment analysis showed that trigrams performed poorly.

Unigrams combined with bigrams gave good performance in the CNB model. This is similar to the research of Go et al. [2] about general Twitter sentiment analysis.

Bigrams on their own and unigrams combined with trigrams showed some relatively good results in the CNB and SVM-linear models.

Similarly, bigrams combined with trigrams, and unigrams combined with bigrams and trigrams in the CNB, SVM-linear and SVM radial basis models had a good performance.

In future research, we will identify the unigrams, bigrams and trigrams that are most informative for our purposes.

C. Machine learning results

The best method was SVM linear with a 95% accuracy when using unigrams. SVM with the radial basis kernel gave high results at 88% accuracy using bigrams and trigrams with a 90% precision and 88% recall, when the neutral class was not used. When the neutral class was used, combining all three n-grams gave the best performance at a 68% accuracy and a 73% precision.

SVM polynomial is known to work well with natural language processing models; however, in our experiments we found it gave the lowest results out of all three kernels. This is similar to research done by Jain and Nayak [14] about movie reviews from IMDB, where they found that the polynomial kernel performed the worse; at the same time they found the linear kernel to work best.

Our results showed that SVM gave extremely high performance. SVM has been found to perform well in several domains, including customer feedback and movie reviews [13], while in the educational domain, Ortigosa [4] showed that SVM gave the lowest results. This could have been influenced by the fact that they used part-of-speech (POS) tagging as a feature with data from Facebook messages. For our data, SVM also shows the best performance when using the neutral class with only a 0.15 loss in accuracy compared with the non-neutral model.

For Maximum Entropy, the best n-gram for both neutral and non-neutral models was trigrams. The non-neutral model results were much higher for trigrams than the other feature combinations, although in the models with the neutral class the results of the trigrams were only slightly higher than the any of the other n-grams. This is similar to the research about movie reviews done by Pang and Lee [13], where the use of trigrams resulted in approximately 80% accuracy.

The Naive Bayes Classifier gave the lowest performance. We noticed that our results were extremely high with the positive class and lower in the neutral and negative class. These results are linked to the number of instances in each class: the size of the negative and neutral classes in our data was significantly lower than the size of the positive class. For the model without the neutral class, the results were higher using unigrams combined with bigrams; this is similar to previous research, e.g., [2], [13]. In the NB model with the neutral class, using trigrams led to the best performance.

Complement Naive Bayes addresses the problem of uneven training sets in Naive Bayes. Our experiments, similar to Gokulakrishnan [15] research about analysing tweets, showed that Complement Naive Bayes performed better than Naive Bayes alone. Therefore, this method is useful in addressing the problem of a small training set for the neutral class, a problem which appeared in previous research [2], [9]. Complement Naive Bayes with the neutral class showed the highest performance when combining the unigrams with the bigrams, while the non-neutral model showed a better

performance when using unigrams alone.

To the best of our knowledge, Complement Naive Bayes has not been previously used for sentiment analysis in the educational domain. Our results indicate that it works well for educational data. Moreover, the results imply that CNB can be a good solution when having uneven data classes, hence this could be a solution for the neutral class problem where training data is usually hard to obtain. It may also be beneficial for sentiment analysis applications in other domains where these problems are encountered.

In conclusion, the best performing machine learning techniques are Support Vector Machines and Complement Naive Bayes. Consequently, we will further investigate the use of these models for our real-time feedback analysis.

D. Neutral class results

The CNB model with the neutral class performed better than without the neutral class. As for the other models the neutral class led to a fairly good performance in all the models except for Naive Bayes, where both neutral and non-neutral performed poorly.

SVM showed the best performance when using the neutral class with an accuracy loss of only 1.5% compared with the non-neutral model. We will continue to explore the use of the neutral model, as it is valuable for the educational domain, especially in relation to providing a complete picture of the proportions of different opinions.

VI. Conclusions and Future Work

In this paper, we investigated different combinations of machine learning techniques, features, preprocessing levels and the use of neutral class for analysing real-time students' feedback.

We found that preprocessing the data increased the accuracy of some models by up to 20 percent, which was expected. However, interestingly we found that in the NB and CNB models with the neutral class, using no preprocessing gave the highest performance. Numbers and punctuation for our application may hold some value, therefore eliminating these through preprocessing may lead to loss of valuable information, which, in turn, leads to a decreased performance of the models.

We experimented with the use of different n-gram combination and found that the best features were unigrams, unigrams combined with bigrams, and in some cases trigrams.

We found that the best models were SVM and CNB; therefore, they could be used for feedback analysis. Our experiments indicate that CNB can be a good solution for uneven training classes and that this can be beneficial when there is not enough data in the neutral class.

Future work includes testing pos tagging as a feature, including different preprocessing techniques such as negation, and keeping the numbers and punctuation. We also plan to widen our research area into detecting specific emotions related to learning.

References

[1] J. Novak and M. Cowling, "The implementation of social networking as a tool for improving student participation in the classroom," *ISANA International Academy Association Conference Proceedings*, vol. 22, pp. 1–10, 2011.

[2] A. Go, R. Bhayani, and L. Huang. (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford. [Online]. Available: http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf

[3] F. Tian, Q. Zheng, R. Zhao, T. Chen, and X. Jia, "Can e-learner's emotion be recognized from interactive Chinese texts?" *International Conference on Computer Supported Cooperative Work in Design*, vol. 13, pp. 546–551, 2009.

[4] A. Ortigosa, J. M. Martin, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, pp. 527 – 541, 2014.

[5] M. Munezero, C. S. Montero, M. Mozgovoy, and E. Sutinen, "Exploiting sentiment analysis to track emotions in students' learning diaries," *Koli Calling International Conference on computing Education Research*, vol. 13, pp. 145–152, 2013.

[6] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of facebook statuses using naive bayes classifier for language learning," *Information, Intelligence, Systems and Applications*, vol. 4, pp. 1–6, 2013.

[7] J. Martin, A. Ortigosa, and R. Carro, "Sentbuk: Sentiment analysis for e-learning environments," *International Symposium on Computers in Education (SIIE)*, vol. 12, pp. 212–217, Oct 2012.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, Stroudsburg, PA, USA, 2011, pp. 30–38.

[9] W. Wang and J. Wu, "Emotion recognition based on cso&svm in e-learning," *International Conference on Natural Computation (ICNC)*, vol. 7, pp. 566–570, 2011.

[10] A. Pak and P. Paroubek, "Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives," *International Workshop on Semantic Evaluation*, vol. 5, pp. 436–439, 2010.

[11] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *Annual Meeting on Association for Computational Linguistics*, vol. 42, pp. 271–278, 2004.

[12] D. Song, H. Lin, and Z. Yang, "Opinion mining in e-learning system," *International Conference on Network and Parallel Computing Workshops*, vol. 6, pp. 788–792, Sept 2007.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," *ACL-02 Conference on Empirical Methods in Natural Language Processing*, vol. 10, pp. 79–86, 2002.

[14] S. Jain and S. Nayak. (2013) Sentiment analysis of movie reviews: A study of features and classifiers. CS221 Project Report, Stanford. [Online]. Available: http://web.stanford.edu/~nayaks/reportsFolder/cs221_report.pdf

[15] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," *Advances in ICT for Emerging Regions (ICTer)*, vol. 13, pp. 182–188, Dec 2012.