Student Evaluations in Teaching - Emotion Classification using Neural Networks

Jaishree Ranganathan, University of North Carolina at Charlotte, Charlotte, USA
Angelina A. Tzacheva, University of North Carolina at Charlotte, Charlotte, USA

The Europian Conference on Education 2020
Official Conference Proceedings

Abstract
Student evaluation of teaching effectiveness plays an important role in Higher Education. Evaluations serve as Formative (identify areas of improvement in the process) and Summative (assess the end goal) measurements of teaching. Educational institutions collect these evaluations in both qualitative and quantitative forms. Qualitative evaluations serve as a bridge for students to express their feelings about the teaching methodology used, instructor efficiency, classroom environment, learning resources, and others. Identifying student emotions help instructors to have good intellectual insight about the actual impact of teaching. Teaching models include traditional models, modern flipped classroom models, and active learning approaches. The light-weight team is an active learning approach, in which team members have a little direct impact on each other's final grades, with significant long-term socialization. We propose and extend the previous method for assessing the effectiveness of the Light-weight team teaching model, through automatic detection of emotions in student feedback in computer science courses by using the Neural Network model. Neural Networks have been widely used and shown high performance in a variety of tasks including but not limited to Text Classification and Image Classification. It is highly deemed to work great with a huge volume of data. In this study, we discuss how sequential model can be used with smaller data sets and it performs well, compared to the baseline models such as Support Vector Machines and Naive Bayes.

Keywords: Classification, Educational Data Mining, Neural Networks, Student Evaluations, Teaching Methods

1. Introduction

Quality of education is one of the primary factors which requires constant attention and improvement. Student evaluations of teaching serve as both formative and summative measures in the process of quality education. Literature dates to 1920's (Wachtel, 1998) with the works of Remmers to assess the student evaluation agreements with alumni and peers (Remmers, 1928), (Remmers, 1930). Educational institutions collect student evaluations in both quantitative and qualitative forms. The quantitative feedbacks include a Likert-type scale in which responses are scored along with a range, to capture the level of agreement and disagreement. Qualitative feedbacks serve as a bridge for students to express their feelings about the teaching method used, instructor efficiency, classroom environment, learning resources, and others.

One of the emerging approaches in the field of teaching is the Active Learning approach. Light-Weight teams (Latulipe et. al., 2015) is an Active Learning approach, in which team members have no direct impact on each other's final grades, yet there is a significant component of peer teaching, peer learning, and long-term socialization. This innovative pedagogical approach has been studied in Computer Science undergraduate courses and has been reported to have high levels of student engagement (Latulipe et. al., 2015), (Macneil et. al., 2016).

Emotion Mining is the process of detecting and analyzing human feelings about events, issues, and or services. Qualitative feedbacks aids in the process of identifying student emotions. Authors Tzacheva et al. (Tzacheva & Ranganathan, 2018), (Tzacheva et. al., 2019), study the effectiveness of teaching model and their impact on student learning styles and experience in the classroom and identify factors that help in performance and positive attitude of students towards Computer science course. They propose a novel method for assessing pedagogical innovation through the detection of emotions in text, produced by student participants, in computer science courses. The results show that the implementation of Active Learning methods increase positive emotions among students and improve their learning experience.

Educational Data Mining is a new field that involves identifying patterns of student behaviors and learning by use of Machine Learning and Data Mining technologies. Neural Networks in Data Mining is a mathematical model which has its roots in biological neural network. Neural networks have achieved impressive results in several classification tasks (Aono & Himeno, 2018), (Kim et. al., 2018), (Lai et. al., 2015), (Severyn & Moschitti, 2015). It is widely perceived that Neural Networks performs well with a huge volume of data. Since student evaluations of teaching have limited data availability considering the number of students registering for a course, very limited works have used Neural networks in the education data mining field. Researchers use the classical machine learning models like Naive Bayes, Support Vector Machine for sentiment classification of student evaluations data. In this work, we use a sequential learning model on the student feedbacks for emotion classification and compare with the traditional models.

The remainder of the paper is organized as follows, section II focusses on related work; followed by method, experiments, and results in sections III and IV.

2. Related Works

2.1 Classification – Traditional Machine Learning Models

Authors (Leong et. al., 2012) use short message service (SMS) for student evaluation. They perform Sentiment Analysis ('positive' and 'negative') on SMS texts. Conceptual words and text link analysis visualization are used to explain the positive and negative aspects of the lecture. Authors (Altrabsheh et. al., 2014), classify real-time student feedback into three sentiment class 'positive', 'neutral', and 'negative' by experts. Naive Bayes, Complement Naive Bayes, Maximum Entropy, and Support Vector Machine models were used for evaluation. Support Vector Machine and Complement Naive Bayes yields better results compared to other models. Similarly, authors (Dhanalakshmi et al., 2016) classify student's feedback into 'positive' or 'negative' and suggest that Naive Bayes performs better with good recall.

Authors (Jagtap & Dhotre, 2014) use a hybrid approach combining Hidden Markov Model (HMM) and Support Vector Machine (SVM) to classify student feedback with sentiment polarity ('positive' and 'negative'). According to the authors, the advance feature selection method and hybrid approach work well for complex data. But they did not show the results of the classification model for validation.

Authors (Rajput et. al., 2016) use tag clouds, and sentiment scores from student feedback data to identify insights about teacher performance. Multi-Perspective Question Answering (MPQA) (Stoyanov et. al., 2005) sentiment dictionary is used to find positive and negative polarity. Word frequency and word attitude are combined to obtain the overall sentiment score for each feedback. They have compared the sentiment score with the Likert scale-based teacher evaluation. Results show that the Sentiment score with word cloud provides better insights than Likert scale results.

2.2 Classification – Neural Networks

Neural Networks are widely used in several classification tasks and proven to achieve the best results. But it is still in the infancy stage with Educational Data. Most of the works in the literature focus on predicting student performance using Artificial Neural Networks. For instance, (Guo et. al., 2015) use multiple level representations with unsupervised learning and fine-tune neural network layers through backpropagation. They use High school data with different kinds of information including background and demographic data, past study data, school assessment data, study data, and personal data. Compared to the traditional methods like Support Vector Machines and Naive Bayes, their model achieves better performance.

Authors (Musso et. al., 2013), also use student background information along with cognitive and non-cognitive measures to predict student academic performance using Artificial Neural Networks achieve greater accuracy compared to the discriminant analysis method.

While the above methods use non-text data for classification, the following researchers use text data. An online discussion forum is a popular tool for student communication and collaboration in web-based courses. Authors (Wei et. al., 2017) use Stanford MOOC posts dataset (Agrawal et. al., 2015) to identify `confusion', or

`urgency' and sentiment of the posts. They propose a transfer learning framework based on convolutional neural networks and long short-term memory models. Student Evaluation of Teaching Effectiveness (SETE) serves as an important aspect in validating the teaching models, resources, and effectiveness of teaching and learning outcomes. Authors (Galbraith et. al., 2012) use Neural Networks to measure student learning outcomes from SETE's.

There is not much work in applying Neural networks for sentiment classification from student evaluation of teaching. In this work, we use a sequential model with 1D convolution and word embedding for automatic classification of emotions from student evaluations.

3. Methodology

We use a Web-Based course evaluation system to collect data for the study. This system is administered by a third-party Campus Labs in assistance with the Center for Teaching and Learning. The data is collected for the following span of 2013 to 2017 including Fall, Spring, and Summer sections of courses handled by the instructor. After the data collection from Campus Labs, jsoup (Hedley, 2009) a Java library is used to process the Html files and extract the comments. The data contains 1070 instances. Sample student comments are shown in Table. 1.

Table 1: *Sample Student Comments*

Comments
Easily available to communicate with if needed.
The course has a lot of valuable information.
There was no enthusiasm in the class. The instructor should make the class more lively and interactive.

3.1 Pre-Processing

We use python Natural Language Toolkit (NLTK) (Perkins, 2010) to process the Qualitative student feedbacks and make it suitable for emotion labeling and classification. The steps include removing certain special characters like punctuation, splitting the sentence into pieces of words called tokens, case-folding, stop-words removal. The pre-processed dataset contains close to 800 records in the dataset.

3.2 Emotion Labeling

Labeling the data is the most significant task for any supervised machine learning algorithm. In this work, we use the National Research Council - NRC Lexicon (Mohammad & Turney, 2013), (Mohammad & Turney, 2010) for this purpose. NRC Emotion lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, disgust, surprise, trust, joy, and sadness) and two sentiments (positive and negative). The Annotations in the lexicon are at the WORD-SENSE level. Each line has the format: <Term> <AffectCategory> <AssociationFlag>.

Student comments are processed and if a match to a word is found then the score is incremented accordingly based on the Flag value in the lexicon, here if a word is

present twice then automatically based on the frequency score for that particular emotion will be incremented. After the entire comment is processed the Emotion which has the highest score is assigned as the final Emotion with respect to that student comment.

3.3 Classification

Classification is the process of predicting the class labels of given data points, and it belongs to the category of Supervised Learning. The learning algorithms for classification are broadly divided into two types as lazy learning (memory-based learning system) and eager learning (optimized learning system) algorithms. Lazy learning algorithms store all the training data and defer the process until it receives a query or test set to process. Whereas the eager learning algorithm learns the classifier structure with the training data and uses the learning to predict the test instances. The former takes less time learning and more time classifying while the latter is the opposite.

Some examples of lazy learning algorithms include K-Nearest Neighbor, Case-Based reasoning; while Naive Bayes, Neural networks, Decision Tree are examples of Eager learning. In this paper, we use Keras (Gulli & Paul, 2017) a high-level neural network API in python for automatic classification of emotion from student evaluation data. The classification model is based on Keras sequential model, which is a linear stack of layers. We use the 1D convolutional kernel with a dense (fully connected) layer compiled with Adaptive Moment Estimation (Adam) optimizer and categorical cross-entropy as a loss function. Finally, the model is trained using Epochs = 5 and Batch size = 2.

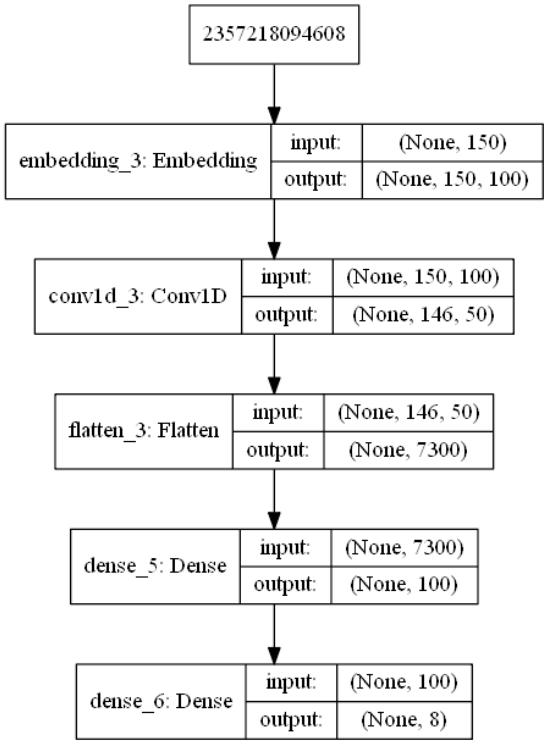

Figure 1: *Neural Network Model Summary*

4. Experiments and Results

In this section, we describe our experiments and results. The data for this study is collected from the Campus Labs website. The data extracted consists of 1070 records. The pre-processed dataset contains close to 820 records in the dataset. For labeling the data - student feedback comments with different types of Emotion, we use the National Research Council - NRC Lexicon (Mohammad & Turney, 2013), (Mohammad & Turney, 2010).

There are several classification algorithms that have been applied to text classification problems. In this work, we use traditional Naive Bayes and Support Vector Classification methods as a baseline to compare the neural network implementation.

4.1 Naïve Bayes Classifier and Support Vector Machine Classifier

One of the popular uses of text pre-processing in the traditional methods is the use of TF-IDF (Term Frequency - Inverse Document Frequency) which is a popular weighting scheme used in information retrieval and text mining applications. It is a statistical measure to evaluate the importance of words in the document or corpus. TF-IDF is mainly composed of two terms: Term Frequency (TF) and Inverse Document Frequency (IDF) as given below.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

The student evaluations dataset is processed with TF-IDF and given as input to the Naive Bayes and Support Vector classification. We achieve an accuracy of approximately 74.79% with Naive Bayes and 77.97% with Support Vector Machine.

4.2 Neural Networks Classifier

For the text input to be understood by the neural network algorithm, it is required to process the text before passing to the classifier model to be trained. For this purpose, words are replaced with unique numbers and combined with the embedding vector to make it semantically meaningful. We achieve an accuracy of approximately 76.7% which is very much in close approximation with the traditional models.

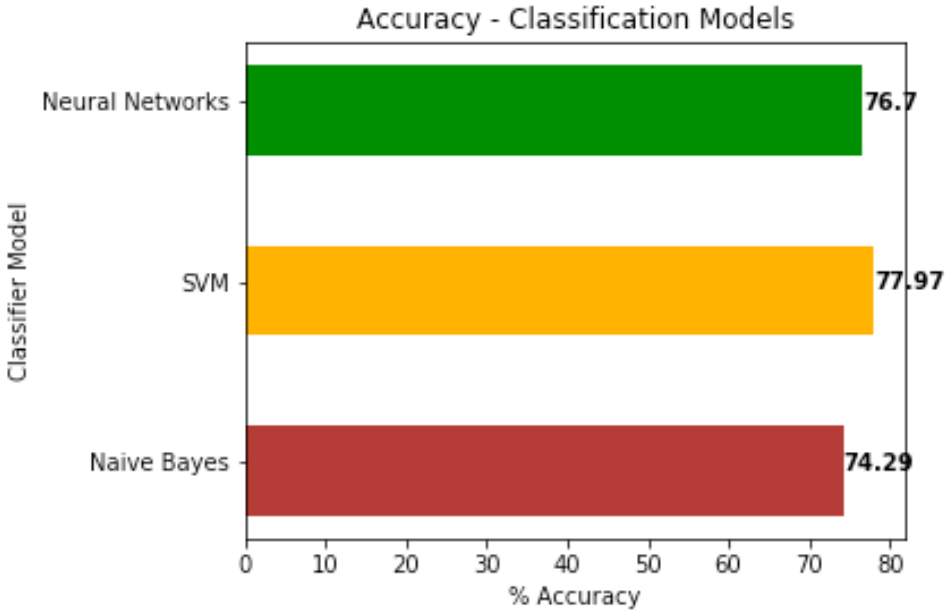

Figure 2: *Classifier Accuracy*

5. Conclusion

In this paper, we apply neural networks classifier for emotion detection in student evaluation of teaching. We use Keras Deep Learning API. Using an appropriate number of epochs for training on the source domain results in better performance. We also compare the neural networks model with the traditional text classification models like Naive Bayes and Support Vector Machine. We notice that neural networks yield (76.7%) similar performance to traditional text classification models like Naive Bayes (74.79%) and Support Vector Machine (77.97%), though the size of the dataset is not big. which is a drawback when using neural networks for classification. In future, we plan to extend this work by collecting student survey to identify actionable patterns that help improve the teaching model learning environment to a better state.

References

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips.

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014, September). Learning sentiment from students' feedback for real-time interventions in classrooms. In International Conference on Adaptive and Intelligent Systems (pp. 40-49). Springer, Cham.

Aono, M., & Himeno, S. (2018, June). Kde-affect at semeval-2018 task 1: Estimation of affects in tweet by using convolutional neural network for n-gram. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 156-161).

Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016, March). Opinion mining from student feedback data using supervised learning algorithms. In 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC) (pp. 1-5). IEEE.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. Research in Higher Education, 53(3), 353-374.

Gulli, A., & Pal, S. (2017). Deep Learning with Keras. Packt Publishing Ltd.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In 2015 International Symposium on Educational Technology (ISET) (pp. 125-128). IEEE.

Hedley, J. (2009). jsoup: Java html parser. 2009-11-29)[2015-06-12] http://jsoup. org.

Jagtap, B., & Dhotre, V. (2014). SVM and HMM based hybrid approach of sentiment analysis for teacher feedback assessment. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(3), 229-232.

Kim, Y., Lee, H., & Jung, K. (2018). AttnConvnet at SemEval-2018 Task 1: attention-based convolutional neural networks for multi-label emotion classification. arXiv preprint arXiv:1804.00831.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence.

Latulipe, C., Long, N. B., & Seminario, C. E. (2015, February). Structuring flipped classes with lightweight teams and gamification. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (pp. 392-397). ACM.

Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. Expert Systems with Applications, 39(3), 2584-2589.

MacNeil, S., Latulipe, C., Long, B., & Yadav, A. (2016, February). Exploring lightweight teams in a distributed learning environment. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 193-198). ACM.

Mohammad, S. M., & Turney, P. D. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 26-34). Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

Musso, M. F., Kyndt, E., Cascallar, E. C., & Dochy, F. (2013). Predicting General Academic Performance and Identifying the Differential Contribution of Participating Variables Using Artificial Neural Networks. Frontline Learning Research, 1(1), 42-71.

Perkins, J. (2010). Python text processing with NLTK 2.0 cookbook. Packt Publishing Ltd.

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-based sentiment analysis of teachers' evaluation. Applied Computational Intelligence and Soft Computing, 2016, 1.

Remmers, H. H. (1928). The relationship between students' marks and student attitude toward instructors. School & Society.

Remmers, H. H. (1930). To what extent do grades influence student ratings of instructors?. The Journal of Educational Research.

Severyn, A., & Moschitti, A. (2015, August). Twitter sentiment analysis with deep convolutional neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 959-962). ACM.

Stoyanov, V., Cardie, C., & Wiebe, J. (2005, October). Multi-perspective question answering using the OpQA corpus. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 923-930). Association for Computational Linguistics.

Tzacheva, A., Ranganathan, J., & Jadi, R. (2019, July). Multi-Label Emotion Mining From Student Comments. In Proceedings of the 2019 4th International Conference on Information and Education Innovations (pp. 120-124). ACM.

Tzacheva, A.A., & Jaishree, R. (2018). Emotion Mining from Student Comments a Lexicon Based Approach for Pedagogical Innovation Assessment.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. Assessment & Evaluation in Higher Education, 23(2), 191-212.

Wei, X., Lin, H., Yang, L., & Yu, Y. (2017). A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. Information, 8(3), 92.