Predicting learning-related emotions from students' textual classroom feedback via Twitter

Nabeela Altrabsheh
School of Computing
Lion Terrace
University of Portsmouth
nabeela.altrabsheh@port.ac.uk

Mihaela Cocea
School of Computing
Lion Terrace
University of Portsmouth
mihaela.cocea@port.ac.uk

Sanaz Fallahkhair
School of Computing
Lion Terrace
University of Portsmouth
sanaz.fallahkhair@port.ac.uk

ABSTRACT

Teachers/lecturers typically adapt their teaching to respond to students' emotions, e.g. provide more examples when they think the students are confused. While getting a feel of the students' emotions is easier in small settings, it is much more difficult in larger groups. In these larger settings textual feedback from students could provide information about learning-related emotions that students experience. Prediction of emotions from text, however, is known to be a difficult problem due to language ambiguity. While prediction of general emotions from text has been reported in the literature, very little attention has been given to prediction of learning-related emotions. In this paper we report several experiments for predicting emotions related to learning using machine learning techniques and n-grams as features, and discuss their performance. The results indicate that some emotions can be distinguished more easily then others.

Keywords

Emotion prediction from text, Machine learning, Learning-related emotions

1. INTRODUCTION

Detecting emotions is important in the learning process [4]. Positive emotions may increase students' interest in learning, increase engagement in the classroom and motivate students [4]. Additionaly, students who are happy generally are more motivated to accomplish their learning goals.

Sentiment analysis research has grown considerably in the last decade, mainly due to the availability of rich text resources such as social networking sites, blogs and microblogs, and product reviews. Despite the name of this area, sentiment analysis is mostly focused on detection of polarity (negative or positive sentiment) rather than specific emotions. Thus, there is relatively little research on the prediction of specific emotions from text [2, 3], with even fewer reports of such research in education [9]. Moreover, from these studies (both within the educational field and outside of it), an even smaller number use machine learning to predict emotion from text, e.g. [2, 3, 9].

In this paper we focus on the prediction of emotions relevant for learning from students' textual feedback via Twitter in a classroom context using machine learning techniques. To investigate the prediction of the identified emotions from text, we experiment with several preprocessing methods, n-gram features, and machine learning techniques.

2. RELATED RESEARCH

There are four main steps to create predictive models from text with machine learning: preprocessing the data, selecting the features, applying the machine learning techniques and evaluating the results.

Preprocessing the data involves preparing the data and cleaning it from unwanted elements which may negatively affect the performance of the machine learning techniques. Some of the general preprocessing techniques used with basic text are: tokenization, convert text to lower or upper case, remove punctuation, remove numbers and, remove stop words [8].

Preprocessing Twitter data requires additional techniques due to the presence of emoticons, hashtags and chat language. Some of the Twitter-specific data preprocessing techniques from previous research [8, 11] are: removing hashtags, removing URLs, removing retweets, identifying emoticons, removing user mentions in tweets, removing Twitter special characters, and slang/chat language handling.

In relation to specific emotions detection, both general preprocessing techniques and Twitter-related preprocessing techniques have been used, e.g. removal of stop words and stemming [3], removing URLs [5], and tokenization [5].

Feature selection refers to the process of selecting relevant features for the particular prediction problem, while eliminating the features that are redundant or irrelevant. In prediction problems where the data is in the form of text, the most common features are n-grams [7]. The most commonly used n-gram for emotion detection is unigrams (one word) [7]. In contrast, there are very few studies investi-

gating the use of bigrams (two words) and trigrams (three words) in emotion prediction. However bigrams and trigrams has been used in sentiment analysis of tweets [7]. In this paper, we investigate the influence of these different n-grams and their combination on emotion detection.

Various *machine learning techniques* have been used for polarity and emotions prediction from text. In our experiments we used classifiers previously shown to work well [9]: Naive Bayes (NB), Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Support Vector Machines (SVM), Maximum Entropy (ME), Sequential Minimal Optimization (SMO), and Random Forest (RF).

Previous research on emotions related to learning indicates a variety of emotions experienced by learners [6]. In previous research [1], we identified from the literature a number of common emotions that are associated with learning: amused, anxiety, appreciation, awkward, bored, confusion, disappointed, embarrassed, engagement, enthusiasm, excitement, frustration, happy, motivated, proud, relief, satisfaction, shame and uninterested.

3. DATA CORPUS

The data was collected from lectures taught in English in Jordanian universities on different topics: calculus, English communication skills, database, engineering, molecular biology, chemistry, physics, science, contemporary history of the world and architecture.

Twitter was used to collect students feedback, opinions, and feelings about the lecture. For each tweet, they were asked to choose one emotion from a set of emotions provided, i.e. the 19 emotions listed in the previous section. Although tweets were used the language was formal and did not include chat language or slang, however, they did include emoticons and hashtags.

A total number of 1522 tweets were collected with their corresponding emotion label. There was one label per feedback. Some of the emotions appeared more frequently than others. The most frequent emotions that were used in our research were: Bored (336), Amused (216), Frustration (213), Excitement (178), Enthusiasm (176), Anxiety (130), Confusion (73), and Engagement (67). The least frequent ones were discarded due to insufficient data for training and testing machine learning algorithms: Happy (32), Satisfaction (31), Appreciation (26), Embarrased (18), Dissapointed (12), Uninterested (4), Proud (3), Relief (3), Shame (2), Awkward (1), and Motivated (1).

4. PREDICTION OF EMOTIONS FROM STUDENTS' FEEDBACK

Two different preprocessing levels were experimented with: (a) high preprocessing, which includes: tokenization, convert text to lower case, remove punctuation, remove numbers, remove stop words, remove hashtags, remove URLs, remove retweets, remove user mentions in tweets, and remove Twitter special characters; (b) low processing, which includes: tokenization, convert text to lower case, and remove stop words.

The high preprocessing was only used for one of the models which contained all the emotions combined, due to the low results that it led to in comparison with the low level of preprocessing for this model. Consequently, for the other models only the low preprocessing was experimented with.

The negative influence of preprocessing on the performance of the models indicates that information that is typically discarded for polarity prediction has value for the identification of specific emotions, as for example in the case of punctuation [11].

We experimented with different n-grams, i.e. unigrams, bigrams, and trigrams, and all combinations between them to find which n-gram or combination of n-grams leads to the best performance for the different models. The features that were experimented with are: Unigrams (UNI); Bigrams (BI); Trigrams (TRI); Unigrams and Bigrams combined; Unigrams and Trigrams combined; Bigrams and Trigrams combined; and Unigrams, Bigrams, and Trigrams combined.

We used the classifiers mentioned previously in section 2 due to their common use in previous research. Additionally, we used two common kernels for SVM: radial basis (RB) and linear (LIN) kernel.

We experimented with all the emotions combined and then subtracted, in turn, the emotion with the lowest number of instances. The total number of models experimented with was 16 models, which are: 7 emotions (All except engagement) + other (8 classes); 6 emotions (7 emotions except confused) + other (7 classes); 5 emotions (6 emotions except anxiety) + other (6 classes); 4 emotions (5 emotions except enthusiasm) + other (5 classes); 3 emotions (4 emotions except excitement) + other (4 classes); 2 Emotions (Amused, Bored) + other (3 classes); and each emotion + other (2 classes).

All the models were tested using 10-fold cross-validation; the accuracy and the error rate were used to assess the overall performance of the classifiers, while the precision, recall, and F-score were used to assess the ability of the classifiers to correctly identify the specific emotion(s).

The results indicate that the models with a single emotion perform better than the multi-emotion models in terms of accuracy, although one has to bare in mind that the baseline for multi-class models is lower than the baseline for 2-class models.

The results show that two classifiers performed best in terms of accuracy: the Support Vector Machine with Radial Basis kernel (RB), mainly for the 2-class models, and Sequential Minimal Optimization (SMO), mainly for the multi-class models. In term of features, unigrams and trigrams were found to lead to the best performance for the 2-class models, while unigrams combined with bigrams and trigrams led to the best performance for the multi-class models.

Despite the fact that accuracy can be useful in predicting the models performance, it does not indicate how well a classifier can predict specific emotions. As the recall indicates the percentage of correctly identified instances for a class of in-

Table 1: Highest recall for each model

Model	Tecbnique	N-gram	Accuracy	Error rate	Precision	Recall	F-score
ALL Preprocessed	ME	UNI+BI+TRI	0.32	0.68	0.34	0.33	0.33
ALL W/O Preprocessing	ME	UNI+BI	0.32	0.68	0.33	0.32	0.32
7 Emotions+ other	NB	BI+TRI	0.26	0.74	0.24	0.25	0.25
6 Emotions+ other	MNB	UNI	0.27	0.73	0.27	0.26	0.27
5 Emotions+ other	MNB	UNI+TRI	0.25	0.75	0.32	0.32	0.32
4 Emotions+ other	MNB	BI	0.26	0.74	0.29	0.38	0.33
3 Emotions + other	ME	UNI+BI+TRI	0.51	0.49	0.43	0.36	0.39
2 Emotions+ other	ME	UNI+BI+TRI	0.57	0.43	0.40	0.51	0.45
Amused	CNB	TRI	0.49	0.51	0.19	0.70	0.30
Anxiety	CNB	TRI	0.45	0.55	0.12	0.77	0.21
Bored	CNB	TRI	0.44	0.56	0.28	0.85	0.42
Confused	CNB	TRI	0.28	0.72	0.06	0.81	0.11
Engagement	CNB	TRI	0.24	0.76	0.04	0.68	0.08
Enthuisiasm	CNB	TRI	0.36	0.64	0.14	0.76	0.24
Excitement	CNB	TRI	0.37	0.63	0.15	0.86	0.26
Frustration	CNB	TRI	0.40	0.60	0.19	0.84	0.31

Table 2: Best overall models for identification of specific emotions

Model	Tecbnique	N-gram	Accuracy	Error rate	Precision	Recall	F-score
Amused	CNB	Bi+Tri	0.64	0.36	0.24	0.62	0.35
Bored	CNB	UNI+BI+TRI	0.71	0.29	0.43	0.63	0.51
Excitement	CNB	UNI+TRI	0.64	0.36	0.21	0.64	0.32

terest, it can be used to assess the ability of the classifiers to predict emotions; in addition, precision can indicate where the identification problems occur.

For most of the models with the highest accuracy, the recall is extremely low or even 0% in some cases. In addition, precision is also low for most of the models (with a few exceptions). For instance in the "engagement + other" model where the accuracy is 95% and the precision, recall, and F-score are (0-0.05)% for the emotion class. This indicates that the high accuracy is due to the correct identification of the "other" class rather than the correct identification of emotion(s).

Table 1 displays the best experimental results when focusing on the recall, i.e. the correct identification of the emotion(s). In terms of machine learning techniques, Complement Naive Bayes (CNB) performs best for half of the models, which could be explain by the ability of this technique to compensate for uneven class sizes. In terms of features, trigrams led to the best performance in the 2-class models, while unigrams combined with bigrams and trigrams led to the best performance in the multi-class models.

The fact that the models with high recall rates have low accuracy and low precision values indicates that many instances of the "other" class are wrongly classified as indicating particular emotions. In other words, although the classifiers have a higher sensitivity for the emotion classes, they are not precise in distinguishing the "other" class from the emotion class(es).

When looking at the overall picture and the balance of the evaluation metrics considered (i.e. accuracy, error rate, precision and recall), some of the models stand out – these are presented in Table 2. We found that the best classifier is Complement Naive Bayes (CNB). When looking at the features, one can notice that different combinations of n-grams led to the best performance for different classifiers. This indicates that a combination of various n-grams instead of a single n-gram is useful for the prediction of specific emotions and should be investigated further.

It is not surprising that the best performing models are for the emotions for which we had larger number of instances (see section 3), i.e. bored, amused and excitement. Interestingly, the models for excitement performed better that the ones for frustration, although there were more instances for frustration than for excitement.

From previous research studies focusing on the prediction of emotions using machine learning techniques, only one study was conducted in an educational context [9]. This research used part-of-speech (POS) tags as features, and more specifically, they experimented with the combination of the following part-of-speech tags: verb, adverb, adjective and noun. They evaluated their models using precision, recall, and F-score and found that Random Forest performed better than the other classifiers with a weighted average F-score at 0.638. Similar to our research they found that the recall score was higher than the precision. From the emotions that we identified as relevant for learning from previous literature, they only looked at anxiety, for which they obtained a precision value of 0.6 using a LogitBoot classifier. However, this re-

search was conducted on Chinese text, which has different characteristics and structures compared with English text. Moreover, the research was based on text from online chats and discussion groups. Furthermore, they used in their approach an affective words base (i.e. lexicon), where each affective word had a number associated with its degree of reflection of a particular emotion.

Outside the educational domain, there are very few studies that looked at the prediction of specific emotions from text only, which are described below.

One study, which used unigrams and a experimented with a multi-class model with 5 emotions [3], found that the Naive Bayes and Support Vector Machine classifiers performed well, leading to an accuracy of 67%. This data, however, is not representative for other types of text expressing emotions, as indicated by the low accuracy, i.e. less than 35%, of these models on test sets with other data. Similarly to the research described above, they also experimented with lexicons for specific emotions.

Another study which used unigrams as a feature and machine learning looked at predicting the presence of emotion versus the lack of emotion [2]; they obtained a maximum accuracy of 74%. However, they did not discuss the performance in terms of identifying the presence of emotion (i.e recall for the emotion). They have also used lexicons with emotion-related words.

However, very few studies investigated the use of other n-grams. Youn and Purver [10] investigated the prediction of emotions from the Chinese microblog service Sina Weibo; in their experiments they found that the models with bigrams and trigrams outperformed the models using unigrams. Similarly, our results showed that using all of the n-grams (i.e. unigrams, bigrams, and trigrams) combined led to the best identification of emotions for the multi-emotion models. Additionally, we found that trigrams led to the best identification of emotions for the 2-class models.

While it is difficult to compare the performance of our models with previous work given the variations in different experimental set-ups (e.g. data origin, language, choice of emotions, choice of features and the use of lexicons), one aspect that seems to be prevalent in previous research is the used of lexicons. Consequently, in out future work, we will investigate the use of such an affective word base for education and its effect on the prediction models.

5. CONCLUSIONS AND FUTURE WORK

In this paper we conducted several experiments with the purpose to investigate the prediction of specific emotions related to learning from students' textual classroom feedback. We focused on several learning emotions which were found to be relevant from previous literature: Amused, Anxiety, Bored, Confusion, Engagement, Enthusiasm, Excitement, and Frustration. We experimented with several preprocessing and machine learning techniques, and also with different combinations of n-gram features.

The models were evaluated using 10-fold cross-validation and using the following evaluation metrics: accuracy, error rate, precision, recall, and F-score. The best performing models were obtained for three particular emotions using 2-class models: amused, bored and excitement. The best classifier was Complement Naive Bayes (CNB). A combination in n-grams led to the best performance in most models.

In future work we will investigate the influence on prediction of a learning-related emotion lexicon; we will also investigate the relation between learning emotions and polarity.

6. REFERENCES

[1] N. Altrabsheh, M. Cocea, and S. Fallahkhair. Predicting students' emotions using machine learning techniques. In *The 17th International Conference on Artificial Intelligence in Education*, 2015. forthcoming.

[2] S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[3] T. Danisman and A. Alpkocak. Feeler: Emotion classification of text using vector space model. In *Convention Communication, Interaction and Social Intelligence*, volume 1, pages 53–59, 2008.

[4] S. D'Mello, T. Jackson, S. Craig, et al. Autotutor detects and responds to learners affective and cognitive states. In *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems*, 2008.

[5] F. Keshtkar and D. Inkpen. A corpus-based method for extracting paraphrases of emotion terms. In *Proceedings of the NAACL HLT Workshop on Computational approaches to Analysis and Generation of emotion in Text*, pages 35–44, 2010.

[6] B. Kort, R. Reilly, and R. W. Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies*, pages 43–436. IEEE Computer Society, 2001.

[7] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, volume 10, pages 1320–1326, 2010.

[8] A. Pak and P. Paroubek. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, volume 5, pages 436–439, 2010.

[9] F. Tian, P. Gao, L. Li, W. Zhang, H. Liang, Y. Qian, and R. Zhao. Recognizing and regulating e-learners' emotions based on interactive chinese texts in e-learning systems. *Knowledge-Based Systems*, 55:148–164, 2014.

[10] Z. Yuan and M. Purver. Predicting emotion labels for chinese microblog texts. In *The 1st International Workshop on Sentiment Discovery from Affective Data*, volume 40, 2012.

[11] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan. Analyzing twitter for social TV: Sentiment extraction for sports. In *Proceedings of the 2nd International Workshop on Future of Television*, volume 2, pages 11–18, 2011.