Detecting sarcasm from students' feedback in Twitter

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. Sarcasm is a sophisticated form of act where one says or writes the opposite of what they mean. Sarcasm is a common issue in sentiment analysis and detecting it is a challenge. While models for sarcasm detection have been proposed for general purposes (e.g. Twitter data, Amazon reviews), there is no research addressing this issue in an educational context, despite the increased use of social media in education. In this paper we experiment with several machine learning techniques, features and preprocessing levels to identify sarcasm from students' feedback collected via Twitter.

Keywords: sentiment analysis; sarcasm detection; students feedback

1 Introduction

Students' feedback can be a valuable data source for detecting students' emotions. Detecting students' emotions is important because previous research highlights that positive emotions increase students' interest in learning, increase engagement in the classroom and motivate students. Moreover, students who are happy are generally more motivated to accomplish their learning goals.

Students' feedback can be analysed using sentiment analysis, which is also known as opinion mining. Sentiment analysis can help the lecturer by analysing large amounts of feedback and summarising it, thus giving a complete picture of what students' opinions are. Unfortunately, students sometimes use sarcasm in their feedback leading to inaccurate results when analysing the data.

Sarcasm is a form of irony conveying criticism in an indirect way. It could also cover up embarrassment in a way that is dramatic or humorous [4]. Unlike lying, sarcasm is used to highlight reality rather than hide it, and it is perceived as being more polite and less aggressive than direct confrontation [4].

In education, sarcasm has been studied from the perspective of the teacher/ lecturer. More specifically, studies looked at the use of sarcasm when communicating with or giving feedback to students. It was found to be negative when sarcasm was used to belittle students, but also found to be positive when sarcasm was used with humour and not targeted at particular students. To the best of our knowledge, there is no research looking at sarcasm in students' feedback.

Detection of sarcasm for text is a difficult problem and there are only a few approaches looking at this problem [2, 3, 6]. None of these approaches, however, look at sarcasm in an educational context.

2 Related work

Detection of sarcasm received increased attention from the research community, with several works reported for the detection of sarcasm from text only [2, 3, 6]. Unlike previous research, our experiments focus on data from an educational context, and more specifically, from students' feedback. We experiment with several classifiers to investigate the reliability of the prediction. In the following we give an overview of the several processes involved in the development of prediction models: preprocessing, feature selection and machine learning techniques.

Preprocessing involves cleaning the data from unwanted elements which may negatively affect the performance of the machine learning techniques. The most common techniques are: tokenization, removing punctuation, removing numbers, and converting text to lower or upper case. Most research about sarcasm in sentiment analysis has been on tweets. Some examples of preprocessing Twitted data are removal of Twitter special characters [6] and removal the HTML links [6].

Feature selection is the process of selecting relevant features for the particular prediction problem. The most commonly feature is unigrams, i.e. individual words. In sarcasm, n-grams are commonly used as well [2].

A variety of *machine learning techniques* were used for sentiment analysis in general and sarcasm detection in particular. We experimented with classifiers that have previously performed well for polarity, emotion prediction, and sarcasm detection [5]: Naive Bayes (NB), Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Maximum Entropy (ME), Support Vector Machines (SVM), Sequential Minimal Optimization (SMO), and Random Forest (RF).

3 Predicting sarcasm from students' feedback

Twitter was used to collect students feedback about the lecture. For each tweet, students were asked to choose an emotion from a set of 19 learning-related emotions provided, e.g. amused, bored, confused, frustrated, and to provide a polarity label (positive/ negative/ neutral). A list of all the emotions can be found in [1]. 1522 tweets were collected with their corresponding emotion and polarity labels.

We identified sarcasm in the tweets by: (a) detecting the word or hashtag "sarcasm" (or other variants such as "sarcastic"), e.g. "Nothing great than doing some vectors on sunday morning #vectors #lecture #sarcasm #justuniversity-things", and (b) detecting if the tweets were labeled an emotion that was opposite to the tweet's meaning, e.g. "in need of some inspiration #literature #lecture #boring #feelingdown" with label amused. Using these two methods, we found 194 sarcastic tweets; all the other tweets were labeled as non-sarcastic.

We tested three preprocessing levels using different techniques which were chosen due to their popularity in previous studies: (a) P1, i.e. the baseline, includes converting text to lower case; (b) P2, includes P1 plus removal of numbers, punctuation, spaces and blanks, and special characters; (c) P3, which includes P2 plus the removal of hashtags, emoticons and Twitter special characters.

Like most researchers, we focus on n-grams, and in particular on unigrams. Unigrams have been widely used in previous sentiment analysis research, as well

as for detecting sarcasm [2]. We also explored other features in combination with unigrams (U): (a) emotion label (E) (from students); (b) polarity label (P) (from students); (c) number of all punctuation characters (PC) (e.g. commas, semicolons, brackets, apostrophes, etc.); (d) number of question and exclamation marks (QE); (e) number of emoticons (NE); (f) number of hashtags (H); (g) time of the tweet during the lecture (T), i.e. beginning, middle, end. All combinations of these additional features with unigrams were experimented with.

4 Results and Discussion

All models were tested using 10-fold cross-validation; accuracy, precision, recall, F-score, and error rate are used to evaluate the models. Precision, recall, F-score and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) for the sarcasm class are also reported. The best models for sarcasm detection are displayed in Table 1.

Table 1. Best models for sarcasm prediction

	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	All	U+E+P	E+P	U	All	E+P	All	U+H
Accuracy	0.82	0.78	0.72	0.87	0.80	0.85	0.86	0.84
Precision	0.84	0.84	0.84	0.76	0.82	0.84	0.84	0.81
Recall	0.82	0.78	0.72	0.87	0.80	0.85	0.86	0.84
F-score	0.83	0.80	0.76	0.81	0.81	0.85	0.84	0.82
Error rate	0.18	0.22	0.28	0.13	0.20	0.15	0.14	0.16
Precision (Sarcasm)	0.34	0.30	0.26	0.00	0.28	0.42	0.43	0.30
Recall (Sarcasm)	0.47	0.51	**0.63**	0.00	0.35	0.33	0.25	0.22
F-score (Sarcasm)	0.39	0.37	0.36	0.00	0.31	0.36	0.31	0.25
AUC	0.76	0.71	0.68	0.50	0.61	0.75	0.60	0.68

Regarding *preprocessing*, we first experimented with unigrams only and found that most of the models showed higher performance in terms of sarcasm detection when using the lowest level of preprocessing (P1), which could be due to hashtags, numbers, and punctuation holding value to the prediction of sarcasm. In previous work, punctuation was used by Tsur et al. [6], where they showed that on its own it is not a good predictor, but that in combination with other features it can increase the performance. Consequently, for the unigrams combinations with the other features, we only experimented with level P1 of preprocessing.

Regarding the *features*, we found that adding the emotion and polarity labels to the unigrams led to an increase of 1 to 6% in sarcasm detection, i.e. recall, in five out of the eight classifiers (NB, MNB, CNB, ME and SMO). We found that overall, the punctuation features (PC and QE), the emoticons count (NE) and the time of the lecture (T) made little difference to the performance, and in some cases led to a small decrease of 1% to 2%. Using all the features led to an increase for the recall of the sarcasm class for the NB, LIN and SMO classifiers by 7%, 4% and 9%, respectively.

In terms of *machine learning techniques*, we found that the best classifiers to detect sarcasm are Complement Naive Bayes, Multinomial Naive Bayes and

Naive Bayes, mainly due to their high recall rate. Naive Bayes has not been used before to detect sarcasm, however, previous research shows that it is performs well in sentiment analysis [1]. Multimonial Naive Bayes has been used in sarcasm detection by Justo et al. [3], obtaining a recall of 77%. We notice that interestingly, CNB performs the lowest in terms of accuracy, despite previous research showing that it performs well with uneven classes [1], which is applicable to our dataset as the sarcasm class represents only 13% of the data; however, CNB led to the highest sarcasm recall of 63%.

We found that the worst classifiers to detect sarcasm are RB, SMO and RF. However, these algorithms are prone to overfitting. Despite the high accuracies of these classifiers, the recall and other measures for the sarcasm class were low, indicating that the good performance is due to the recognition of the majority class, i.e. non-sarcasm. Similarly, research by González-Ibánez et al. [2] to identify sarcasm from tweets showed that SMO led to a high accuracy of 71% using unigrams as features. They did not, however, report the measures for the sarcasm class; consequently, we do not know how much of the accuracy performance is due to the recognition of sarcasm.

Although the results we reported in this paper are relatively low for what is considered a good performance for a classifier, when looking at the previous results on sarcasm detection, we can see that our results are comparable, and on some metrics even better than previous research. In future work we will investigate the use of lexicons and other features such as POS tagging.

References

1. Altrabsheh, N., Cocea, M., Fallahkhair, S.: Predicting students' emotions using machine learning techniques. In: The 17th International Conference on Artificial Intelligence in Education (AIED) (2015), (Forthcoming)
2. González-Ibánez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 581–586 (2011)
3. Justo, R., Corcoran, T., Lukin, S.M., Walker, M., Torres, M.I.: Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems 69, 124 – 133 (2014)
4. Shany-Ur, T., Poorzand, P., Grossman, S.N., Growdon, M.E., Jang, J.Y., Ketelle, R.S., Miller, B.L., Rankin, K.P.: Comprehension of insincere communication in neurodegenerative disease: Lies, sarcasm, and theory of mind. Cortex 48(10), 1329 – 1341 (2012)
5. Tian, F., Gao, P., Li, L., Zhang, W., Liang, H., Qian, Y., Zhao, R.: Recognizing and regulating e-learners emotions based on interactive chinese texts in e-learning systems. Knowledge-Based Systems 55, 148–164 (2014)
6. Tsur, O., Davidov, D., Rappoport, A.: ICWSM-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: The 4th International AAAI Conference on Weblogs and Social Media. pp. 162–169 (2010)