University of Glasgow Terrier Team / Project Abacá at RepLab 2014: Reputation Dimensions Task

Graham McDonald, Romain Deveaud, Richard McCreadie, Timothy Gollins, Craig Macdonald and Iadh Ounis

School of Computing Science
University of Glasgow, G12 8QQ, Glasgow, UK
firstname.lastname@glasgow.ac.uk

Abstract. This paper describes our participation in the RepLab 2014 Reputation Dimensions task. The task is a multi-class classification task where tweets relating to an entity of interest are to be classified by their reputation dimension. For our participation we investigate two approaches; Firstly, we use a term's gini-index score to quantify the term's representativeness of a specific class and construct class profiles for tweet classification, and secondly, we perform tweet enrichment using a web scale corpus to derive terms representative of a tweet's class, before training a classifier with the enriched tweets. Our tweet enrichment approach performed exceedingly well, showing that this approach is effective for classifying tweets by their reputation dimensions and a promising direction for future work.

1 INTRODUCTION

This notebook paper describes our participation in the Reputation Dimensions task of RepLab 2014 [1]. RepLab is a competitive exercise for Online Reputation Management (ORM) systems, organized as an activity of the Cross Language Evaluation Forum (CLEF)[1]. ORM is concerned with the tracking and monitoring of media to identify what is being said about an entity [2]. With the increased popularity of social media and communication platforms such as Twitter[2] that allow users to reach a global audience and share their experiences in real time, it is especially important for companies to be able to monitor their public perception, assisted by ORM tools, and react in an appropriate and timely manor.

For a company to be able to respond to changing public opinion in online communications, relating to the company, there are three components of the communication that must be understood. Firstly, the company must be aware of the aspects (*Dimensions*) of its business the communication is about, for example Products & Services. Secondly, the company must understand the type of author, for example is the author a journalist, and thirdly how influential the author is.

For our participation in RepLab 2014, we participated in the Reputation Dimensions task which addresses the first of these three components. The remainder of this paper is

[1] http://clef2014.clef-initiative.eu/
[2] http://twitter.com

structured as follows. Section 2 gives an overview of the Reputation Dimensions task before Section 3 outlines our classification approaches. Section 4 presents the results of our submitted runs and, finally, in Section 5 we present our conclusions.

2 THE REPUTATION DIMENSIONS CORPUS AND TASK

In this section we give an overview of the Reputation Dimensions corpus and task. The corpus consisted of English and Spanish tweets crawled during the period 1st June 2012 to 31st December 2012, with just over 75% English tweets. Each tweet related to at least one of 31 entities of interest from the Banking and Automotive industries. For each entity, there were at least 2,200 tweets, with at least 700 and 1,500 tweets for the training and test sets respectively. The most recent tweets were used for the test set. Participants were provided the tweet ids and had to download the tweet text directly from Twitter. To retrieve the tweet text, we used the Java tool provided by the RepLab organisers.

The Reputations Dimensions task is a multi-class classification task. Participating systems were to classify tweets as one of the seven reputation dimensions (Innovation, Citizenship, Leadership, Workplace, Governance, Performance and Products & Services) defined in the RepTrak Framework by the Reputation Institute[3]. The data set also included the "Undefined" class for tweets that were not classified as one of these seven dimensions. Undefined tweets are not included in the evaluation.

3 CLASSIFICATION APPROACHES

In this section we give an overview of the approaches we deployed for our participation in the Reputation Dimensions task. The research questions we address in our participation are twofold: (1) Can we use the gini-index of a term as a measure of the terms belonging to a reputation dimension to construct dimension profiles for tweet classification? and 2) Can we identify a tweet's reputation dimension with greater accuracy by enriching the tweet using a web scale corpus?

The remainder of this section is structured as follows: Section 3.1 outlines our Reputation Dimension Profiling approach, then Section 3.2 gives an overview of our approach for tweet enrichment.

3.1 REPUTATION DIMENSION PROFILING

For our dimension profiling approach we convert the tweets to lower case, remove non-alphanumeric characters, new lines and URLs, before using the Terrier Information Retrieval Platform [3] to remove stop-words and index the tweets for each class (as defined in the gold standard). We discard terms with a term frequency of < 3 before calculating the conditional probability of a term belonging to each class, normalised by the class distribution over the collection. Using this probability, we calculate a term's gini-index [4] score to quantify the terms discriminatory power between classes. Tweets

[3] http://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework

and profiles are then represented as term frequency vectors and we classify tweets to their closest dimension profile using cosine similarity. For developing this approach, we performed a 5-fold cross validation on the training data creating profiles from the training split of each fold.

The "Undefined" class is included in our gini-index calculations, resulting in scores in the range of 0.125 (least discriminative terms) to 1 (most discriminative terms). For terms with a suitably high gini-index score, we use the term's class conditional probability to determine the class that the term is most representative of and add the term to the class profile. Appropriate gini-index and class conditional probability thresholds were determined by parameter analysis, resulting in profiles constructed from terms with a gini-index score $>= 0.3$ and a class conditional probability > 0.1.

We submitted three runs employing this gini-index technique: Firstly, uogTr_RD_1 classifies tweets using profiles constructed by the process. Secondly, uogTr_RD_2 constructs profiles using this process before enriching the profiles with expansion terms derived from Wikipedia[4]. Finlay, uogTr_RD_2 constructs profiles using this process before enriching the profiles with class specific query expansion terms.

3.2 TWEET ENRICHMENT

For our tweet enrichment approach we pre-process tweets using the same approach as in Section 3.1 (converting to lowercase, removing non-alphanumeric characters, new lines and URLs), before enriching the tweets.

To obtain enrichment terms, we use Terrier to submit a raw tweet as a query to a large contemporaneous web corpus. The top 10 retrieved documents then form a pseudo-relevant document set. We calculate the entropy of each term within the set of retrieved documents, and we select the top 20 terms with the highest entropy as the most informative terms related to the tweet. Then, we enrich the pre-processed tweets by appending these informative terms. Stop-words are removed from the enriched tweets, which are further converted into term frequency feature vectors that are used to train several types of classifiers in Weka [5], using 10-fold cross validation.

We submitted two runs employing this technique: Firstly, for uogTr_RD_4 we train an SVM model using Weka and LibSVM [6]. Secondly, for uogTr_RD_5 we train the Weka implementation of the Random Forests [7] classification algorithm.

4 RESULTS

A total of 30 systems were submitted for the Reputation Dimensions task. The task organisers also reported on a baseline text classification approach that used tweet words as feature vectors to train an SVM classifier. Results were reported ranked by the system's overall Accuracy.

Table 1 presents the accuracy scores of our runs plus the baseline approach and the mean over all 31 submissions.

[4] http://en.wikipedia.org/wiki/Main_Page

Table 1. Submitted Runs: Overall Accuracy Score

System	Accuracy
uogTr_RD_1	0.4960
uogTr_RD_2	0.6205
uogTr_RD_3	0.6086
uogTr_RD_4	**0.7318**
uogTr_RD_5	0.6871
mean	0.6424
baseline	0.6221

We see that our Tweet Enrichment approach with SVM model performed excellently being ranked first with an accuracy score of 0.7318. The Tweet Enrichment approach with a Random Forests model also performs well, achieving an accuracy score of 0.6871 markedly above the baseline score of 0.6221.

Our Dimension profiling approach performed less well due to the fact that increasing gini-index and class conditional probability thresholds results in the selection of fewer discriminative terms for a class profile, therefore profiles become smaller as they become more class specific. This makes the task of classifying previously unseen tweets increasingly difficult due to the sparse nature of tweets. Enriching the dimension profiles counteracted this somewhat, as we see increased performance using the enriched profiles from 0.4960 for uogTr_RD_1 to 0.6086 and 0.6205 for uogTr_RD_3 and uogTr_RD_2 respectively.

Table 2 shows the relative frequency of classes for classified tweets calculated as $\#class\,predictions/total\,tweets\,classified*100$ and Table 3 shows the precision and recall for each of the classes. We see that most of the runs are slightly biased towards the largest class "Products and Services" but the runs that performed best achieved notably higher precision on smaller classes such as "Innovation" and "Leadership". We would expect to be able to further improve our results by achieving higher precision scores for "Workplace" and "Performance".

Table 2. Relative Frequency of Classes for Classified Tweets.

	Innovation	Citizenship	Leadership	Workplace	Governance	Performance	Products and Services
uogTr_RD_1	9.24	18.09	4.95	14.90	**13.03**	4.97	34.79
uogTr_RD_2	3.41	14.91	2.60	9.58	10.36	2.59	**56.52**
uogTr_RD_3	4.52	14.20	**2.61**	7.78	8.99	2.83	59.04
uogTr_RD_4	**0.77**	**17.87**	1.34	**2.51**	9.46	**5.40**	64.69
uogTr_RD_5	0.03	9.78	0.16	0.18	4.34	1.69	83.79
baseline	0.10	12.26	1.32	0.90	8.01	2.51	74.88
gold standard	1.08	17.89	2.64	4.00	12.08	5.68	56.60

Table 3. Precision (p) and Recall (r) Values per Dimension.

	Innovation		Citizenship		Leadership		Workplace		Governance		Performance		Products and Services	
	p	r	p	r	p	r	p	r	p	r	p	r	p	r
uogTr_RD_1	0.0537	**0.5294**	0.5775	0.6677	0.1626	**0.3534**	0.0886	**0.3798**	0.4203	0.5231	0.2328	0.2352	**0.7236**	0.5162
uogTr_RD_2	0.0828	0.2973	0.6485	0.6166	0.2600	0.2943	0.1238	0.3398	0.5013	0.4963	0.3384	0.1783	0.6614	0.7668
uogTr_RD_3	0.0622	0.2973	0.6643	0.6021	0.2538	0.2876	0.1399	0.3122	0.4603	0.3952	0.2960	0.1708	0.6380	0.7718
uogTr_RD_4	0.2863	0.2124	0.7358	**0.7388**	0.3878	0.2231	0.4034	0.2882	0.5882	**0.5263**	0.3710	**0.4123**	0.6768	0.8861
uogTr_RD_5	**0.9000**	0.0294	**0.8722**	0.5390	**0.8490**	0.0604	0.5517	0.0284	**0.6117**	0.2506	0.3857	0.1351	0.5718	**0.9695**
mean	0.2749	0.0883	0.7180	0.5529	0.4379	0.1307	0.4343	0.1751	0.5667	0.3552	0.3598	0.2388	0.6134	0.8468
baseline	0.1791	0.0392	0.8453	0.5514	0.4192	0.1989	**0.6419**	0.1387	0.5473	0.3469	**0.4464**	0.1983	0.6528	0.8316

5 CONCLUSIONS

In this paper, we described our participation in RepLab 2014 Reputation Dimensions task. We investigated two distinct approaches; firstly we use a term's gini-index score to identify terms representative of specific classes to build class profiles for classifying tweets, and secondly, we take a tweet enrichment approach using a large contemporaneous web corpus to derive terms representative of the tweet's class before training a classifier on the enriched tweets.

We found that our tweet enrichment approach performed very well. In particular, we note that when training a SVM classifier our tweet enrichment approach achieved the best overall Accuracy results of the task. This approach also performed markedly above average when training a Random Forests classifier. These results are very encouraging and we intend to explore this methodology further as future work.

ACKNOWLEDGMENTS The authors would like to thank Information Technology as a Utility Network (ITaaU) and the EC co-funded project SMART (FP7-287583).

References

1. Amigó, E., Carrillo-de-Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2014: author profiling and reputation dimensions for Online Reputation Management. In: Proceedings of the Fifth International Conference of the CLEF Initiative. (2014)
2. Madden, M., Smith, A.: Reputation management and social media. Washington, DC: Pew Internet & American Life Project. Retrieved May **26** (2010)
3. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proc. OSIR. (2006)
4. Aggarwal, C., Zhai, C.: A survey of text classification algorithms. In Aggarwal, C.C., Zhai, C., eds.: Mining Text Data. Springer US (2012) 163–222
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1) (2009) 10–18
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3) (2011) 27:1–27:27
7. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32