Proceedings of the Ninth International AAAI Conference on Web and Social Media

Tweet Enrichment for Effective Dimensions Classification in Online Reputation Management

Graham McDonald, Romain Deveaud, Richard McCreadie, Craig Macdonald and Iadh Ounis
University of Glasgow
School of Computing Science
{firstname.lastname}@glasgow.ac.uk

Abstract

Online Reputation Management (ORM) is concerned with the monitoring of public opinions on social media for entities such as commercial organisations. In particular, we investigate the task of reputation dimension classification, which aims to classify tweets that mention a business entity into different dimensions (e.g. "financial performance" or "products and services"). However, producing a general reputation dimension classification system that can be used across businesses of different types is challenging, due to the brief nature of tweets and the lack of terms in tweets that relate to specific reputation dimensions. To tackle these issues, we propose a robust and effective tweet enrichment approach that expands tweets with additional discriminative terms from a contemporary Web corpus. Using the RepLab 2014 test collection, we show that our tweet enrichment approach outperforms effective baselines including the top performing submission to RepLab 2014. Moreover, we show that the achieved accuracy scores are very close to the upper bound that our approach could achieve on this collection.

Introduction

Online Reputation Management (ORM) is concerned with the tracking and monitoring of media to identify what is being said about an entity, such as a business. Real-time social media and communication platforms, e.g. Twitter, allow users to instantly inform a global audience on their experiences and opinions. The propagation of these personal messages can result in a swell of public opinion. Indeed, this forming of consensus in public opinion can have a serious impact on the reputation of commercial organisations.

Therefore, it is important for commercial organisations to be able to monitor and understand what is being said about them online, so as to react efficiently and respond appropriately. A crucial step in the process of reputation management is understanding how the conversation relates to specific aspects or *dimensions* of the business, e.g. "Governance", "Financial Performance" or "Products & Services", since different dimensions can require different types or levels of responses.

Automatically classifying tweets with respect to particular dimensions is challenging, since tweets are short and often do not contain terms that are specific to a reputation dimension or that state the potential impact to the company's reputation. For example, the tweet "Bank of America Corp. has agreed to pay $2.43 billion to settle a legal dispute with its European counterpart" is ambiguous as it could refer to a settlement agreed directly with another company, e.g. over a "products and services" patent, and might not likely generate very much public interest. However, this tweet is about a lawsuit that had a significant negative impact on the company's reputation. Therefore, to infer this extra context, we need more information than can be extracted from the tweet text itself. Given this example tweet, our approach expands the tweet with terms such as "lawsuit", which are related to the "Governance" reputation dimension. This expansion enables the classifier to classify the tweet correctly.

In this paper, we propose an enhanced text classification approach to automatically classify tweets into different reputation dimensions. The process consists of three stages: firstly, we use the tweet as a query to retrieve a ranking of contemporary Web documents related to the tweet. Secondly, we identify the most informative terms from the returned documents to enrich the tweet representation and finally, we apply a classifier using the enriched tweet collection.

The contributions of this paper are two-fold. First, we present an effective approach to reputation dimensions classification that achieves a state-of-the-art performance on the RepLab 2014 dataset by enriching the tweet with highly informative terms. Second, by varying the number of Web documents used for selecting expansion terms and the number of expansion terms selected to be added to a tweet, we show that, surprisingly, unlike in classical query or document expansion scenarios, expanding the tweet with a single expansion term is most effective.

ORM and Reputation Dimensions Classification

ORM has received increased interest within the Information Retrieval (IR) community in recent years. Much of this interest has been generated through the RepLab evaluation campaigns (Amigó et al. 2014) that focus on the develop-

Products & Services	Products and services offered by the company or reflecting the consumers' satisfaction.
Innovation	Innovativeness shown by the company, nurturing novel ideas and incorporating them into products.
Workplace	Employees' satisfaction or the company's ability to attract, form and keep talented and highly qualified people.
Citizenship	Company acknowledgment of community and environmental responsibility, including ethic aspects of the business: integrity, transparency and accountability.
Governance	The relationship between the company and the public authorities.
Leadership	The leading position of the company.
Performance	The company's long term business success and financial soundness.

Table 1: Reputation Dimension Definitions.

ment of reputation management systems for managing the reputation of companies.

The task of Reputation Dimensions Classification was first introduced within RepLab 2014 (Amigó et al. 2014) and aims to classify tweets relating to a business entity into one of the seven reputation dimensions shown in Table 1. The task is defined as a multi-class classification task where, given a tweet about an entity of interest E and a set of reputation dimensions $D = \{d_1, d_2, ...\}$, the goal is to automatically classify the tweet to the single reputation dimension that the tweet relates to.

The approach that we present and expand upon here achieved the best accuracy results in the RepLab 2014 task (McDonald et al. 2014). In particular, we deploy a query expansion (QE) technique to identify terms that relate to the reputation dimension of the tweet. It has been shown in the TREC microblog track (Ounis et al. 2011) that using QE could enhance tweet retrieval effectiveness. However, such approaches select query expansion terms from the set of retrieved tweets in the collection. Differently, we use an external, contemporaneous Web corpus to obtain expansion terms for enriching the collection (Kwok and Chan 1998) before classifying the tweets. While addressing the different task of identify concepts in tweets, Meije *et al.* (2012) represents a related work, in that they enrich the representation of tweets by identifying related articles from Wikipedia based on word n-grams. On the other hand, we use the expansion terms obtained from a Web corpus to perform an enhanced text classification for identifying reputation dimensions.

Tweet Enrichment for Dimension Classification

In this section we present our proposed approach to Reputation Dimensions Classification. Our approach enhances text classification by enriching the tweets with externally sourced expansion terms that are related to the tweet's reputation dimension.

As illustrated in Figure 1, our reputation dimension classification approach is a sequence of three main steps: (1) the retrieval of Web documents that are topically related to the tweet of interest, (2) the extraction of the most informative terms from these documents, and (3) the classification of the tweet after being further enriched with informative terms.

Firstly, we submit the entire tweet as a query to a Web corpus in order to retrieve a ranked list of documents that are topically related to this tweet (step 1). We then use the top n documents to extract a set of terms that will be used to

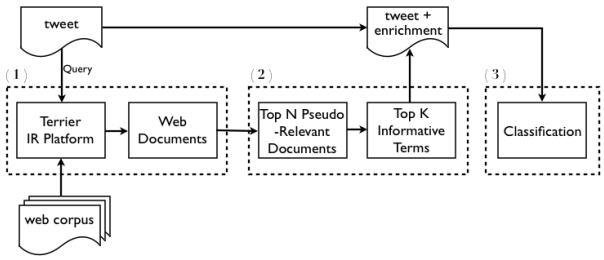

Figure 1: Tweet enrichment and classification process.

enrich the tweet. To do so, we first clean the retrieved Web documents from non-informative content by removing all HTML (including CSS and JavaScript) markup, stopwords, and non-word characters. Then, we build a language model θ_n from the cleaned and concatenated textual content of the top n documents, before computing the entropy $H(w)$ of each term $w \in \theta_n$ (step 2):

$$H(w) = -P(w|\theta_n) \log P(w|\theta_n) \quad (1)$$

We chose entropy to score terms since it has been shown to be effective in identifying highly informative terms and multi-word terms within documents (Villada Moirn and Tiedemann 2006). Here, the probability of word w is computed using a maximum likelihood estimate:

$$P(w|\theta_n) = \frac{\text{tf}_{w,\theta_n}}{\sum_{w' \in \theta_n} \text{tf}_{w',\theta_n}}. \quad (2)$$

Finally, we select the top k terms, ranked by entropy, and concatenate them to the tweet text to form the enhanced tweet representation before converting the enhanced tweet to a term frequency (tf) vector to train a classifier. Our tweet enrichment approach is a general solution and any classifier can be applied to the enriched collection. However, we found the combination of tf features and an SVM classifier (Chang and Lin 2011) to be the most effective[1] and so due to space constraints we report this configuration.

Experimental Setup

In this section we first give an overview of the Reputation Dimensions Classification corpus, before detailing the baselines and evaluation measures we use. Finally, we give an overview of our classification setup.

Corpus: The RepLab 2014 Reputation Dimensions corpus is a collection of tweets, in which each tweet relates to one of 31 entities from the automotive or banking domains (20 automotive, 11 banking). The collection contains English and Spanish tweets crawled between 1st June 2012 and 31st December 2012. The ratio of English to Spanish tweets is approx. 3:1. The task organisers supplied the tweet ids and corresponding ground truth labels for the seven reputation dimensions.

There are 15,294 training tweets and 32,447 test tweets. The number of tweets and the time period the tweets were

[1]Binary, tf and tf-idf feature representations, and SVM, Naive Bayes and Random Forests classifiers were tested.

Approach	Training	Precision	Recall	F_1	Accuracy
RepLab Baseline	Per-Entity	0.5332	0.3293	0.4072	0.6222
RepLab Average	N/A	0.4865	0.3412	0.3942	0.6425
RepLab Best	Across-Entities	0.4928	**0.4697**	0.4810	0.7319
SVM No-Enrichment	Across-Entities	0.4904	0.1568	0.2361	0.4854
Enriched-SVM $(n = 10, k = 1)$	Across-Entities	**0.7502**$^{\triangle}$	0.3861$^{\triangle}$	**0.5016**$^{\triangle}$	**0.7431**$^{\triangle}$
Enriched-SVM $(n = 30, k = 1)$	Oracle	0.7522$^{\triangle}$	0.3862$^{\triangle}$	0.5022$^{\triangle}$	0.7451$^{\triangle}$

Table 2: Reputation dimension classification performance of each baseline and our proposed approach. The best performance is highlighted in bold. Statistical significant improvements (averaged across the entities) over the SVM Baseline (two-sided pairwise t-test $p < 0.01$) are denoted $^{\triangle}$.

collected varies between entities in the training and test splits of the corpus. For each entity, at least 700 tweets from the start of the crawl period were selected as training tweets, leaving at least 1,500 test tweets for each entity.

Baselines: We compare our approach with two different baselines from RepLab 2014. The first one, *RepLab Baseline*, is composed of 31 different SVM classifiers, each trained on tweets relating to a single entity (Amigó et al. 2014). The second baseline, *RepLab Average*, is the average of all the systems that were submitted to RepLab 2014.

Evaluation measures: In our experiments, we report Precision, Recall, F_1, and Accuracy. Precision, Recall and F_1 are calculated by macro averaging across reputation dimensions and micro averaging across entities.

Classification: As shown in step 1 of Figure 1, to retrieve the set of top n documents, we use the Terrier IR platform (Ounis et al. 2006) with the BM25 retrieval model, using the default Terrier settings. Documents are retrieved from an index of ClueWeb12B[2], a collection of 52 million Web pages that were crawled in the first half of 2012. No stemming or stopword removal was applied to the ClueWeb12 corpus or the tweet queries.

Next, having retrieved the top n documents, we select the top k terms ranked by entropy as enrichment terms and add the terms to the tweet to construct the enriched tweet, as shown in step 2 of Figure 1.

After adding the enrichment terms to the tweet, we remove English stopwords and convert @ people mentions and # hashtags to simple terms by removing the non-alphanumeric characters. The enriched tweet is then converted to tf feature vectors for input to the classification stage of our approach, shown in step 3 of Figure 1.

To perform the text classification, we use the WEKA (Hall et al. 2009) machine learning platform with the LIBSVM (Chang and Lin 2011) extension and a linear kernel. As the task is a multi-class classification task where we want to classify a tweet into one of the seven reputation dimensions, we perform a 1-against-1 classification using pairwise coupling. Moreover, we are interested in developing a general Reputation Dimensions Classification solution that is not dependent on prior knowledge of the entity. Therefore, different to the RepLab 2014 baseline detailed above, we train a single global model for all 31 business entities. Finally, we learn the effective values of n documents and k terms by optimising for Accuracy on the training split of the collection ($n = 10$,$k = 1$).

[2]http://www.lemurproject.org/clueweb12.php

Reputation Dimensions Classification Results

In this section, we report and analyse the results of our proposed tweet enrichment approach. We seek to answer two research questions. Firstly, "Is our tweet enrichment approach more effective for classifying tweets to reputation dimensions than a non-enriched approach?" and, secondly, "Which setting of n documents and k terms achieves the best reputation dimensions classification Accuracy?".

Table 2 shows the results of our tweet enrichment approach, denoted as *Enriched-SVM* $(n = 10, k = 1)$, and the baselines *RepLab Baseline* and *RepLab Average*. Table 2 also shows the results this approach achieved in RepLab 2014 (McDonald et al. 2014), denoted as *RepLab Best*, and the results achieved by text classification when the tweet enrichment approach is not applied (i.e. setting the n and k parameter values to 0), denoted as *SVM No-Enrichment*. Finally, in order to assess the best achievable performance that our approach could attain, we also report an oracle where the values of n and k maximise the Accuracy on the test set.

On analysing Table 2, we first see that our tweet enrichment approach achieves the best performance in terms of Precision, F_1 measure, and Accuracy when compared to the baselines. Moreover, it shows substantial and significant improvements over the SVM classification when tweet enrichment is not applied. The deployed configuration of $n = 10$ documents and $k = 1$ terms also outperforms *RepLab Best* by greatly improving Precision (+52.2%) while slightly degrading Recall (-17.8%).

Finally, we see that *Enriched-SVM* $(n = 10, k = 1)$ achieves performances that are very close to the upper-bound Accuracy that our tweet enrichment approach could possibly achieve on the RepLab 2014 dataset, i.e. when we set $n = 30$ documents and $k = 1$ terms.

In answer to our first research question, we conclude that enriching tweets with terms extracted from a Web corpus achieves a significant improvement in classification results over a non-enriched approach, while achieving state-of-the-art results on the RepLab 2014 dataset.

Next, using Table 3 we examine the influence of the number of documents n and the number of enrichment terms k over the Accuracy of our approach.

Firstly, on analysing Table 3, we note that our approach outperforms the 0.7319 Accuracy achieved by *RepLab Best* for each value of n when k is set to 1. Secondly, we observe

k, # of enrichment terms	n, # of documents						
	1	5	10	20	30	40	50
1	0.7440	0.7443	**0.7431**	0.7427	0.7451	0.7437	0.7435
5	0.7361	0.7411	0.7395	0.7352	0.7361	0.7380	0.7414
10	0.7316	0.7396	0.7363	0.7332	0.7390	0.7354	0.7389
20	0.7254	0.7335	0.7319	0.7390	0.7377	0.7374	0.7365
30	0.7191	0.7310	0.7327	0.7332	0.7330	0.7347	0.7335
40	0.7161	0.7292	0.7319	0.7334	0.7323	0.7353	0.7334
50	0.7131	0.7298	0.7268	0.7269	0.7340	0.7298	0.7315
75	0.7033	0.7251	0.7313	0.7322	0.7281	0.7300	0.7310
100	0.6983	0.7223	0.7305	0.7310	0.7302	0.7275	0.7286

Table 3: Accuracy scores for our tweet enrichment approach as we vary the values of the n and k parameters. The deployed configuration, $n = 10$ $k = 1$, is highlighted in bold.

that our approach yields similar performances when using any number of documents from 5 to 50, although Accuracy tends to decrease when enriching the tweets with a large number of terms. This demonstrates the robustness of our tweet enrichment approach.

Thirdly, and surprisingly, we see that for each n (number of documents) retrieved, our approach achieves its best Accuracy when performing a single-term enrichment (i.e. k is set to 1). This differs from similar experiments in microblog retrieval where using 10 to 20 enrichment terms has been shown to be effective (Li et al. 2011; Amati et al. 2011). Moreover, in ad-hoc document retrieval, it has been shown that expanding a query with 10 to 40 terms tends to achieve the best performance (Cui et al. 2002).

Indeed, although the performance slightly decreases when more terms are added to the tweet, Accuracy stays well above the non-enriched SVM classification (0.4854 Accuracy), showing again the effectiveness of our approach for this task.

In answer to our second research question, we conclude that using one single enrichment term always allows to achieve the best results, while we can use any number of documents, between 5 and 50, without markedly altering the Accuracy. An initial inspection of the single terms selected as expansion terms suggests that our tweet enrichment approach selects terms that are specific to a reputation dimension for the entity in the tweet. Future work will investigate further what makes a good expansion term for this task.

Conclusions

In this paper, we have shown that enriching a tweet with terms from a contemporary external Web corpus helps to classify the tweet by the reputation dimension that the tweet relates to. However, surprisingly, we found that using a single expansion term achieves the best performance on the RepLab 2014 test collection. Moreover, we showed that our tweet enrichment approach to reputation dimensions classification achieves an Accuracy score that is very close to the upper-bound that we could achieve on this collection. As future work, we intend to conduct a thorough investigation into how our tweet enrichment approach performs for each entity and on each reputation dimension.

References

Amati, G.; Amodeo, G.; Bianchi, M.; Marcone, G.; Bordoni, F. U.; Gaibisso, C.; Gambosi, G.; Celi, A.; Di Nicola, C.; and Flammini, M. FUB, IASI-CNR, Univaq at TREC 2011 Microblog track. In *Proc. of TREC'11*.

Amigó, E.; Carrillo-de-Albornoz, J.; Chugur, I.; Corujo, A.; Gonzalo, J.; Meij, E.; de Rijke, M.; and Spina, D. Overview of RepLab 2014: Author profiling and reputation dimensions for Online Reputation Management. In *Proc. of CLEF'14*.

Chang, C.-C., and Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2(3):27:1–27:27, 2011.

Cui, H.; Wen, J.-R.; Nie, J.-Y.; and Ma, W.-Y. Probabilistic Query Expansion Using Query Logs. In *Proc. of WWW'02*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18, 2009.

Kwok, K. L., and Chan, M. Improving two-stage ad-hoc retrieval for short queries. In *Proc. of SIGIR'98*.

Li, Y.; Zhang, Z.; Lv, W.; Xie, Q.; Lin, Y.; Xu, R.; Xu, W.; Chen, G.; and Guo, J. PRIS at TREC 2011 Microblog track. In *Proc. of TREC'11*.

McDonald, G.; Deveaud, R.; McCreadie, R.; Gollins, T.; Macdonald, C.; and Ounis, I. University of Glasgow Terrier Team / Project Abac at RepLab 2014: Reputation Dimensions Task. In *Proc. of CLEF'14*.

Meij, E.; Weerkamp, W.; and de Rijke, M. 2012. Adding semantics to microblog posts. In *Proc. of WSDM'12*.

Ounis, I.; Amati, G.; Plachouras, V.; He, B.; Macdonald, C.; and Lioma, C. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR'06*.

Ounis, I.; Macdonald, C.; Lin, J.; and Soboroff, I. Overview of the TREC-2011 Microblog track. In *Proc. of TREC'11*.

Shen, D.; Pan, R.; Sun, J.-T.; Pan, J. J.; Wu, K.; Yin, J.; and Yang, Q. Query enrichment for web-query classification. *ACM Trans. on Information Systems* 24(3):320–352, 2006.

Villada Moirn, B., and Tiedemann, J. Identifying idiomatic expressions using automatic word alignment. In *Proc. of EACL'06*.