CIRGIRDISCO at RepLab2014 Reputation Dimension Task: Using Wikipedia Graph Structure for Classifying the Reputation Dimension of a Tweet

Muhammad Atif Qureshi[1,2], Arjumand Younus[1,2], Colm O'Riordan[1], and Gabriella Pasi[2]

[1] Computational Intelligence Research Group, National University of Ireland Galway, Ireland
[2] Information Retrieval Lab,Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
muhammad.qureshi@nuigalway.ie,arjumand.younus@nuigalway.ie, colm.oriordan@nuigalway.ie, pasi@disco.unimib.it

Abstract. Social media repositories serve as a significant source of evidence when extracting information related to the reputation of a particular entity (e.g., a particular politician, singer or company). Reputation management experts manually mine the social media repositories (in particular Twitter) for monitoring the reputation of a particular entity. Recently, the online reputation management evaluation campaign known as RepLab at CLEF has turned attention to devising computational methods for facilitating reputation management experts. A quite significant research challenge related to the above issue is to classify the reputation dimension of tweets with respect to entity names. More specifically, finding various aspects of a brand's reputation is an important task which can help companies in monitoring areas of their strengths and weaknesses in an effective manner. To address this issue in this paper we use dominant Wikipedia categories related to a reputation dimension in a random forest classifier. Additionally we also use tweet-specific features, language-specific features and similarity-based features. The experimental evaluations show a significant improvement over the baseline accuracy.

1 Introduction

Over the past few years social media has emerged as an effective marketing platform with brands using it for broadening their reach and enhancing their marketing. At the same time social media users excessively voice out their opinions about various entities (e.g. musicians, movies, companies) [5]. This has given birth to a new area within the marketing domain known as "online reputation management" whereby automated methods for monitoring reputation of entities are essential requiring novel computational algorithms to facilitate the work of reputation management experts [1, 2, 4]. This paper describes our experience in devising an algorithm for dealing with the "reputation dimension classification" challenge in the context of RepLab2014 where we are given a set of entities within two domains (i.e., automotives and banking) and for each entity a set of tweets, which contain labels along eight different dimensions.

We utilize the Wikipedia category graph structure of an entity to observe the amount of discussion related to a reputation dimension within the tweet. The experimental results show an improved accuracy over the baseline and our system stands at position 5 among the overall submissions.

2 Task Overview and Dataset

The main aim of the RepLab activity within CLEF is to focus on online reputation of companies on Twitter, and the RepLab activity for 2014 comprises of reputation dimension classification which is a fairly new task and differs from past two year's tasks (i.e., RepLab 2012 and RepLab 2013) [3]. The task comprises classification of tweets according to the dimension of the reputation thereby implying identification of various aspects significant to a company's reputation and following are the standard categories used[3]:

- Products & Services
- Innovation
- Workplace
- Citizenship
- Governance
- Leadership
- Performance
- Undefined

As an example, a tweet containing information about employees' resignations within a company would fall under the dimension "Workplace" whereas a tweet containing information about net profits earned by the company in a financial period is likely to fall under the dimension "Performance."

The corpus is a multilingual collection of tweets referring to a set of 31 entities spread into two domains: automotive and banking[4]. Table 1 shows a summary of the tweets within each dimension from within the automotive and banking domain; we show the numbers from within the tweets we were able to crawl and these numbers may differ from the ones available at time of evaluation. For each entity, at least 2,200 tweets have been collected. The 700 first tweets have been taken to compose the training set, and the other ones are for the test set.

[3] Note that these are the standard categories provided by the Reputation Institute.

[4] Note that the reputation dimension classification task within RepLab uses the same dataset as RepLab2013; however, it utilizes tweets within two domains out of the four domains.

Table 1. No. of Tweets within Each Reputation Dimension

Domain	Products/Services	Innovation	Workplace	Citizenship	Governance	Leadership	Performance	Undefined
Automotive	6757	1691	192	1069	44	85	409	1503
Banking	1141	47	276	1140	1259	212	534	725

3 Methodology

Our algorithm uses the following sets of features:

- Wikipedia-based features
- Statistical features which we further categorize into tweet-specific features, language-specific features and word-occurrence features.

In this section we first present a background on Wikipedia category-article structure followed by a description of the feature set we have utilized.

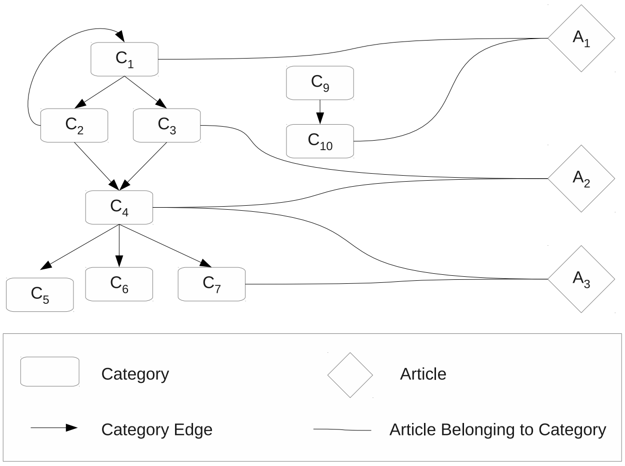

Fig. 1. Wikipedia Category Graph Structure along with Wikipedia Articles

3.1 Background

Our algorithm makes use of the encyclopedic structure in Wikipedia; more specifically the knowledge encoded in Wikipedia's graph structure is utilized for the classification of reputation dimension within tweets. Wikipedia is organized into categories in a taxonomical structure (see Figure 1). Each Wikipedia category can have an arbitrary number of subcategories as well as being mentioned inside an arbitrary number of supercategories (e.g., category C_4 in Figure 1 is a subcategory of C_2 and C_3, and a supercategory of C_5, C_6 and C_7.) Furthermore, in

Wikipedia each article can belong to an arbitrary number of categories, where each category is a kind of semantic tag for that article [8]. As an example, in Figure 1, article A_1 belongs to categories C_1 and C_{10}, article A_2 belongs to categories C_3 and C_4, while article A_3 belongs to categories C_4 and C_7.

3.2 Wikipedia-Based Feature Set: Relatedness Score Based on Wikipedia Category-Article Structure

In this section we present a description of the Wikipedia-based feature set we utilize. The extracted phrases (i.e., n-grams) from a tweet are matched with Wikipedia categories and a voting mechanism is used to score the frequently matched Wikipedia categories. Note that we use the technique similar to the one proposed in [7] with the only difference being that we directly utilize Wikipedia categories corresponding to a reputation dimension. We maintain a voting count corresponding to each Wikipedia category through which the probability of a Wikipedia category belonging to a particular reputation dimension is calculated, the final phase involves a manual analysis for fetching the categories most representative of a particular dimension. For the manual analysis, we plot the obtained categories using Gephi whereby probabilities are plotted to select the Wikipedia categories most closely related to a given reputation dimension. As an example, Figure 2 illustrates the graph of Wikipedia categories corresponding to the reputation dimension of "Innovation" for the automotive domain. The red-colored nodes in Figure 2 represent the Wikipedia categories that occur in a particular dimension with a probability of 1.0, the white-colored nodes represent a probability of 0.0, and the various green-colored nodes represent probabilities around 0.5. Using these graphs, we select the Wikipedia categories that represent high probabilities for a dimension and these categories are then used in our relatedness scoring framework.

After the selection of Wikipedia categories we extract possible n-grams from a tweet and then we score relatedness of n-grams of a tweet with a given dimension. The extracted phrases from a tweet which are contained in the selected Wikipedia categories are used to calculate the relatedness score. The following summarizes important factors which contributes in calculating our relatedness score for a tweet

- $Depth_{significance}$ denotes the significance of category depth at which a matched phrase occurs; the deeper the match occurs in the taxonomy the less its significance to the dimension under consideration. This implies that the matched phrases in the parent category of the dimension under investigation are more likely to be relevant to the dimension than those that are at a deeper depth.
- $Cat_{significance}$ denotes the significance of a matched phrase's categories corresponding to the dimension under investigation.

The relatedness scores constitute our Wikipedia-based feature set for the reputation dimension classification task.

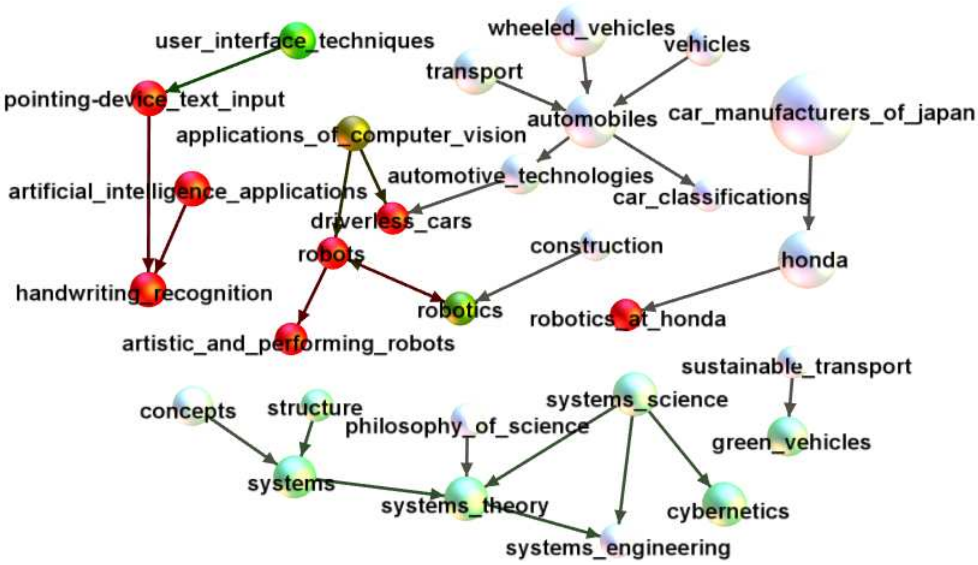

Fig. 2. Wikipedia Category

3.3 Additional Set of Statistical Features

We also use three additional set of statistical features which similar to the ones proposed by Kothari et al.[6] fall under the following categorization

- Tweet-specific features: We used four tweet-specific features that relate to how a tweet is written. They are: (1) presence of hashtag (#tag); (2) presence of user mention (some_user); (3) presence of url in a tweet; (4) language of the tweet (i.e., English or Spanish).
- Language-specific features: We used three language-specific features that relate to various aspects of reputation dimension for a brand. They are: (1) occurrence of a percentage symbol in a tweet; (2) occurrence of currency symbol in a tweet; (3) proportion of common-noun POS tags, proper-noun POS tags, adjective POS tags and verb POS tags in a tweet.
- Word-occurence features: We used two word-occurrence features of which the first checks for the presence of other entity names of same domain; note that products and services dimension contains a lot of tweets whereby other entities are mentioned in the tweet. The second feature first counts the number of times a word occurs in a given dimension for different entities (i.e., checks for word occurrence in 20 entities of automotive domain, and 11 entities of banking domain) and if the number of occurrences is above an empirically-set threshold we add that particular word to our dictionary of dimension terms. The number of dimension terms present are then used as features.

3.4 Machine Learning and Experimental Runs

Using the feature sets described in Section 3.2 and 3.3, we train a random forest classifier over the training data and then use it to predict labels for the test data. We perform three machine learning runs as follows:

1. For the first run, we use only Wikipedia-based features of Section 3.2 whilst training random forest classifier per domain i.e. combining all tweets related to a a particular domain into one training and one test set
2. For the second run, we use only the additional features of Section 3.3 whilst training a random forest classifier per domain i.e. combining all tweets related to a a particular domain into one training and one test set
3. For the third run, we use all features i.e. both Wikipedia-based features and additional features of Section 3.2 and 3.3 whilst training a random forest classifier per domain i.e. combining all tweets related to a a particular domain into one training and one test set

Table 2. Results of Filtering Task of RepLab 2013

Team	Accuracy	F-measure
uogTr_RD_4	0.7318	0.4735
DAE_RD_1	0.7231	0.3906
Lys_RD_1	0.7167	0.4774
SIBTEX_RD_1	0.7073	0.4057
CIRGIRDISCO_RD_3	0.7071	0.3195

4 Experimental Results

4.1 Dataset

We performed our experiments by using the data set provided by the organizers of RepLab 2014 [2]. In this data set 31 entities were provided, and for each entity at least 2200 tweets were collected: the first 700 constituted the training set, and the rest served as the test set. The measures used for the evaluation purposes are Accuracy, Precision and Recall.

4.2 Results

Table 2 presents a snapshot of the official results for the filtering task of RepLab 2014, where CIRGIRDISCO is the name of our team. As can be seen from Table 2, out of a total of 8 participating teams in RepLab2014 reputation dimension classification task 4 teams outperform our best run. Our system shows good results for the evaluation measure of accuracy; however, the evaluation measures of precision and recall show an average performance and we believe these errors come from some noisy Wikipedia categories. As future work, our aim is to remove the noisy Wikipedia categories along with automation of the category selection process of Section 3.2.

References

1. E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. WePS3 Evaluation Campaign: Overview of the On-line Reputation Management Task. In *2nd Web People Search Evaluation Workshop (WePS 2010), CLEF 2010 Conference, Padova Italy*, 2010.
2. E. Amigo, J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martin, E. Meij, M. de Rijke, and D. Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Fourth International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain. Proceedings*, Springer LNCS, 2013.
3. E. Amigó, J. Carrillo-de-Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, and D. Spina. Overview of RepLab 2014: author profiling and reputation dimensions for Online Reputation Management. In *Proceedings of the Fifth International Conference of the CLEF initiative*, Sept. 2014.
4. E. Amigó, A. Corujo, J. Gonzalo, E. Meij, and M. d. Rijke. Overview of replab 2012: Evaluating online reputation management systems. In *CLEF 2012 Labs and Workshop Notebook Papers*, 2012.
5. C. Dellarocas, N. F. Awad, and X. M. Zhang. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. In *MANAGEMENT SCIENCE*, pages 1407–1424, 2003.
6. A. Kothari, W. Magdy, A. K. Darwish, and A. Taei. Detecting comments on news articles in microblogs. *ICWSM 2013*, 2013.
7. M. A. Qureshi, C. ORiordan, and G. Pasi. Exploiting wikipedia for entity name disambiguation in tweets. In *15th International Conference on Applications of Natural Language to Information Systems, NLDB 2014*, 2014.
8. T. Zesch and I. Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.