View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Coordinación de Bibliotecas de la Universidad Politécnica de Madrid

DAEDALUS at RepLab 2014: Detecting RepTrak Reputation Dimensions on Tweets

César de Pablo-Sánchez[1], Janine García-Morera[1],
Julio Villena-Román[1,2], José Carlos González-Cristóbal[3,1]

[1] DAEDALUS - Data, Decisions and Language, S.A.
[2] Universidad Carlos III de Madrid
[3] Universidad Politécnica de Madrid
cdepablo@daedalus.es, jgarcia@daedalus.es,
jvillena@daedalus.es, josecarlos.gonzalez@upm.es

Abstract. This paper describes our participation at the RepLab 2014 reputation dimensions scenario. Our idea was to evaluate the best combination strategy of a machine learning classifier with a rule-based algorithm based on logical expressions of terms. Results show that our baseline experiment using just Naive Bayes Multinomial with a term vector model representation of the tweet text is ranked second among runs from all participants in terms of accuracy.

Keywords: RepLab, CLEF, reputation analysis, reputation dimensions, machine learning classifier, Naive Bayes Multinomial, rule-based approach, hybrid approach, combination.

1 Introduction

RepLab [1] is a competitive evaluation exercise for reputation analysis, launched in 2012 edition of CLEF [2] campaign, which started focusing on the problem of monitoring the reputation of entities (mainly companies) in Twitter, dealing with the tasks of entity name disambiguation, reputation polarity, topic detection and topic ranking. However, RepLab 2014 introduced two new tasks, categorization of messages with respect to standard reputation dimensions and the characterization of Twitter profiles (author profiling) with respect to a certain activity domain.

Specifically, the reputation dimensions scenario consists on a classification task that must return the implicit reputational dimension in a given tweet, to be chosen among the standard categorization provided by the Reputation Institute[1]: (1) Products/Services, (2) Innovation, (3) Workplace, (4) Citizenship, (5) Governance, (6) Leadership, (7) Performance, and (8) Undefined. Participants are provided with a training corpus containing collection of tweets in Spanish and English referring to a

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

selected set of entities in the automotive or banking domain. Each tweet is categorized into one of the aforementioned reputation dimensions.

This paper describes our participation at the RepLab 2014 reputation dimensions scenario. We are a research group led by DAEDALUS[2], a leading provider of language-based solutions in Spain, and research groups of Universidad Politécnica and Universidad Carlos III of Madrid. We are long-time participants in CLEF [2], in many different tracks and tasks since 2003, also in both previous years of RepLab [3].

The idea behind our participation was to evaluate the best combination strategy of a machine learning classifier with a rule-based algorithm based on logical expressions of terms. Our experiments and results achieved are presented and discussed in the following sections.

2 Experiments

The dataset for the Reputation Dimension task is composed of two languages, English and Spanish, in two different domains, automotive and banking. Our system uses a different pipeline for each of the two languages as we were interested in the comparison between rule based classifiers developed for the Spanish language and statistical machine-learning classifiers. We submitted five runs that combine the statistical and rule-based classifiers.

We invested a certain effort to the process of tokenization of the tweet text and URL as some preliminary experiments using cross-validation on the training corpus. These experiments showed that this tokenization process was much more important than the selection of an algorithm. Our runs use information from the *text* and *extended_url* fields in the tweet.

Our baseline run (Run #1) is based on a supervised classifier for each language. Multinomial Naive Bayes (NBM) classifier on a simple bag of words representation was selected with cross-validation among a collection of different algorithms.

We used Weka 3.7 implementation of NBM [7] and the provided WordTokenizer that allows to define split characters that are removed from the term vector space representation of the text. Besides the usual split symbols, spaces and some punctuation, we use tweet specific delimiters such as hashtags (#), usernames (@) and emoticons, and also URL specific delimiters such as slashes, ampersands, question marks and hyphens that are used to separate words in SEO optimized URLs. Finally, as a high number of terms were low frequency numerals we decided to add numbers as well to help in normalization.

During the development process, we tested different parameters configuration and algorithms to reach the conclusion that NBM was robust enough and other representations (bigrams, feature selection) were not adding additional value.

Regarding the language, each of the two classifiers has different performance as the amount of training data for each language was quite different. English training data is composed of 11 869 tweets but Spanish data is about one third in size (3 692

[2] http://www.daedalus.es/

tweets). In our preliminary experiments using cross-validation, the Spanish classifier performed about 10% in accuracy lower than the English classifier and that was particularly meaningful for categories with few labelled instances (Innovation, Leadership or Workplace).

Table 1. Category distribution in training corpora

Dimension	Spanish	English
Citizenship	645	1 564
Governance	429	874
Innovation	38	178
Leadership	41	256
Performance	228	715
Products and services	1 477	6 421
Undefined	782	1 446
Workplace	52	415
TOTAL	3 692	11 869

The rest of the runs make use of different combinations of this NBM classifier with a rule-based classifier for business reputation developed prior to our participation in the task. This rule-based classifier is an adaptation for tweets of a previous model developed for longer texts like news and blogs. This classifier was only available in Spanish, so English just uses the initial baseline NBM classifier.

The combination of methods in the different runs is described in next table.

Table 2. Description of runs.

Run	Description
Run #1	NBM classifier for each language
Run #2	NBM classifier for English, rule-based classifier for Spanish
Run #3	NBM classifier for English, rule-based classifier for Spanish with improvements according to this specific domain
Run #4	English: NBM classifier Spanish: stacked combination of the statistical classifier with the rule base classifier: first the rule-based classifier is used, but if the result is "Undefined", NBM is used
Run #5	English: NBM classifier Spanish: voted combination of the two classifiers priming the rule-based classifier. When the two classifiers disagree on a classification, the rule-based one is used.

The rule-based classifier is build using Textalytics Text Classification API [4], which, despite its name, itself is based on a hybrid algorithm [5] [6] that combines statistical classification, which provides a base model that is relatively easy to train, with rule-based filtering, which is used to post-process and improve the results provided by the previous classifier by filtering false positives and dealing with false negatives and allows to obtain a high degree of precision for different environments.

The machine-based classifier uses an implementation based on kNN and we also have a simple rule language that allows to express lists of positive, negative and relevant (multiword) terms appearing in the text.

The classifier uses a slightly modified RepTrak ontology that contains more detailed classes, for instance, "Products and services" include "Satisfaction of necessities", "Reclamations", "Customer relationship management", "Value for money", "Quality of products and services" and "Warranty". Moreover, it is a multilabel classifier and can assign several labels to a single message.

3 Results

The reputation dimensions task has been evaluated as a classification problem, so accuracy and precision/recall measures over each class are reported, using accuracy as the main measure.

Results achieved by our runs are shown in Table 3. The columns in the table are accuracy and the ratio of classified tweets, i.e., the ratio from the set of tweets that were available at the time of evaluation. The organizers state that a baseline that classifies every tweet with the most frequent class would get 56% accuracy.

Table 3. Results for our runs.

Run	Accuracy	Ratio of classified tweets
Run #1	0,72	0,96
Run #4	0,70	0,98
Run #3	0,66	0,91
Run #2	0,59	0,82
Run #5	0,59	0,82

Next table shows the final ranking for the dimension task in terms of accuracy for the top 5 runs. Our baseline run achieved the second best result among all.

Table 4. Results of best runs.

Run	Accuracy	Ratio of classified tweets
uogTr_RD_4	0,73	0,99
Run #1	0,72	0,96
LyS_RD_1	0,72	0,91
SIBtex_RD_1	0,71	0,95
CIRGIRDISCO_RD_3	0,71	0,95

The following table and figure represents the distribution of classes in the gold standard and in the output of our runs. Our runs, as most runs from participants, are clearly biased to the most frequent class ("Products and services"), as can be seen comparing with the gold standard.

Table 5. Distribution of classes.

Run	Innovation	Citizenship	Leadership	Workplace
GOLD	306	5 027	744	1 124
Run #1	9	3 760	34	147
Run #4	79	3 226	138	319
Run #3	36	2 225	163	303
Run #2	79	2 235	138	317
Run #2	79	2 235	138	317

Run	Governance	Undefined	Performance	ProductsServices
GOLD	3 395	4 349	1 598	15 903
Run #1	2 649	1 678	982	22 645
Run #4	2 067	939	1 173	23 963
Run #3	1 498	2 986	1 036	23 657
Run #2	1 574	6 151	1 126	20 284
Run #5	1 574	6 151	1 126	20 284

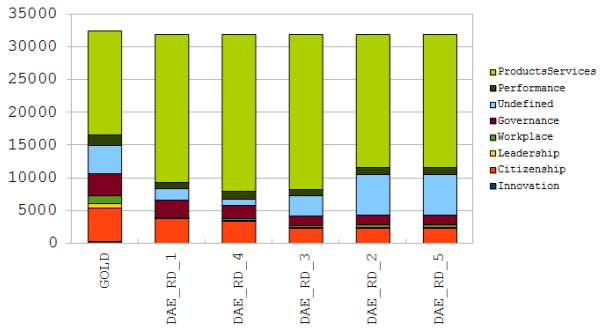

Figure 1. Distribution of classes

The following table represents the precision and recall of our runs, and the best ranked experiment in terms of accuracy. Apparently, our problem is on recall rather than precision of results.

Table 6. Recall/precision of each run

Run	Recall	Precision
Run #1	0,029411765	1,000
Run #2	0,045751634	0,175
Run #3	0,039215686	0,333
Run #4	0,045751634	0,175
Run #5	0,045751634	0,175
uogTr_RD_4	0,212418301	0,286

4 Conclusions and Future work

Results show that our baseline experiment using Naive Bayes Multinomial with a term vector model representation of the tweet text is ranked second among runs from all participants in terms of accuracy. No definite conclusion can be drawn from this fact, whether the Naive Bayes algorithm achieves better or worse accuracy for prediction reputation dimensions than our rule-based model, as approaches are mixed in both languages. If we had had the rule-based model migrated to English in time, the comparison among runs would be easier. Moreover, again due to lack of time and resources, we have not been able yet to carry out an individual analysis by language so we do not understand yet the contribution of each approach to the final result.

However, accuracy values show that, despite of the difficulty of the task, results are quite acceptable and somewhat validate the fact that this technology may be already included into an automated workflow process for the first step towards social media mining and online reputation analysis.

Moreover, a manual inspection of the training data reveals certain miss classifications and lack of criteria in the assignment of categories, with some points of ambiguity and disagreement regarding the consideration of whether a tweet must be assigned or not to a given reputation dimension, specifically for the case of product and services and citizenship. We would thank the clear description of guidelines with the annotation criteria in function of the context.

Acknowledgements

This work has been supported by several Spanish R&D projects: *Ciudad2020: Towards a New Model of a Sustainable Smart City* (INNPRONTA IPT-20111006), *MA2VICMR: Improving the Access, Analysis and Visibility of Multilingual and Multimedia Information in Web* (S2009/TIC-1542) and *MULTIMEDICA: Multilingual Information Extraction in Health Domain and Application to Scientific and Informative Documents* (TIN2010-20644-C03-01).

References

1. Enrique Amigó, Jorge Carrillo-de-Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, Damiano Spina. 2014. *Overview of RepLab 2014: author profiling and reputation dimensions for Online Reputation Management*. Proceedings of the Fifth International Conference of the CLEF Initiative, September 2014, Sheffield, United Kingdom.
2. CLEF. 2014. CLEF Initiative (Conference and Labs of the Evaluation Forum). http://www.clef-initiative.eu/
3. Julio Villena-Román, Sara Lana-Serrano, Cristina Moreno-García, Janine García-Morera, José Carlos González-Cristóbal. 2012. *DAEDALUS at RepLab*

2012: Polarity Classification and Filtering on Twitter Data. CLEF 2012 Labs and Workshop Notebook Papers. Rome, Italy, September 2012.

4. Textalytics Text Classification API. 2014. Text Classification v1.1. http://textalytics.com/core/class-info
5. Julio Villena-Román, Sonia Collada-Pérez, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2011. *Método híbrido para categorización de texto basado en aprendizaje y reglas*. Procesamiento del Lenguaje Natural, Vol. 46, 2011, pp. 35-42.
6. Julio Villena-Román, Sonia Collada-Pérez, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2011. *Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization*. Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press 2011.
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.