# LyS at CLEF RepLab 2014: Creating the State of the Art in Author Influence Ranking and Reputation Classification on Twitter

David Vilares, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez and Jesús Vilares

Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, España
{david.vilares, miguel.hermo, miguel.alonso, carlos.gomez, jvilares}@udc.es

**Abstract.** This paper describes our participation at RepLab 2014, a competitive evaluation for reputation monitoring on Twitter. The following tasks were addressed: (1) categorisation of tweets with respect to standard reputation dimensions and (2) characterisation of Twitter profiles, which includes: (2.1) identifying the type of those profiles, such as journalist or investor, and (2.2) ranking the authors according to their level of influence on this social network. We consider an approach based on the application of natural language processing techniques in order to take into account part-of-speech, syntactic and semantic information. However, each task is addressed independently, since they respond to different requirements. The official results confirm the competitiveness of our approaches, which achieve the 2nd place, tied in practice with the 1st place, at the author ranking task; and 3rd place at the reputation dimensions classification tasks.

**Keywords**: Reputation Monitoring, Author Ranking, Twitter, Natural Language Processing, Machine Learning

## 1 Introduction

In recent years, Twitter has become a wide information network, where millions of users share their views about products and services. This microblogging social network is an important source of information for companies and organisations, which aim to know what people think about their articles. In this way, identifying how people relate aspects and traits such as performance, services or leadership with their business, is a good starting point for monitoring the perception of the public via sentiment analysis applications. In a similar line, companies are interested in user profiling: identifying the profession, cultural level, age or the level of influence of authors in an specific domain may have potential benefits when making decisions with respect to advertisement policies, for example.

The RepLab 2014 on Twitter [1] focusses on these challenges, providing standard metrics and test collections where both academic and commercial systems can be evaluated. The collections contain tweets written in English and Spanish. Two main tasks were proposed: (1) categorisation of tweets with respect to standard reputation dimensions and (2) characterisation of Twitter profiles. The first task consisted of classifying tweets into the standard reputation dimensions: *products&services*, *innovation*, *workplace*, *citizenship*, *governance*, *leadership*, *performance* and *undefined*. The characterization of Twitter profiles is composed of two subtasks: (2.1) author categorisation and (2.2) author ranking. The author categorisation task covers up to 7 user types: *journalist*, *professional*, *authority*, *activist*, *investor*, *company* or *celebrity*. With respect to the author ranking task, the goal is to detect *influential* and *non-influential* users, ranking them according to this aspect (from the most to the least influential). Our approaches achieve state-of-the-art results for the classification on reputation dimensions and author ranking.

The remainder of the paper is structured as follows. Section 2 describes the main features of our methods. Sections 3, 4 and 5 show how we tackle the proposed tasks, illustrating and discussing the official results. Finally, we present our conclusions in Section 6.

## 2 System description

The major part of our models rely on natural language processing (NLP) approaches which include steps such as: preprocessing, part-of-speech (PoS) tagging and parsing. The obtained syntactic trees act as a starting point for extracting the features which feed the supervised classifier employed for tasks 1 (reputation dimensions classification) and 2.1 (author categorisation). We built different models for each task and for each language considered in the evaluation campaign. With respect to task 2.2 (author ranking), a simple but effective method was used. Differences between tasks and languages are explained in the following sections. We describe below the high level architecture of our NLP pipeline.

### 2.1 NLP for online reputation

**Preprocessing** We carry out an *ad-hoc* preprocessing to normalise some of the most common features of the Twitter jargon, which may have an influence on the perfomance of the tasks proposed at RepLab 2014:

- *Replacement of URL's*: References to external links and resources are replaced by the string *'URL'*.
- *Hashtags*: The use of hashtags may be helpful for classification tasks, since they are often used to label tweets. In this way, we only delete the symbol *'#'* in order to give to these elements the same treatment as words.
- *Twitter usernames*: In this social network, the usernames are preceded by the symbol *'@'*. In order not to cause confusion at the tokenisation or tagging steps, we delete that symbol, to then capitalise the first character and give these elements the same treatment as actual proper names.

**Part-of-speech tagging** In order to be able to obtain the syntactic structure of tweets, we first need to label each token of the message with its respective part-of-speech tag. We used the Ancora [2] and the Penn Treebank [3] corpora to train the Spanish and the English taggers, respectively. The Spanish tagger relies on the Brill tagger [4] implementation included with NLTK[1], following the configuration described at [5]. With respect to English we used an averaged perceptron discriminative sequence model [6] which presents state-of-the-art results for the Penn Treebank. Specifically, we took the trained model provided with the TextBlob[2] framework. During the process of PoS tagging we also obtain the lemma of each word.

**Dependency parsing** Given a sentence $S = w_1...w_n$, where $w_i$ represents the word at the position $i$ in the sentence, a dependency parser returns a *dependency tree*, a set of triplets $\{(w_i, arc_{ij}, w_j)\}$ where $w_i$ is the *head* term, $w_j$ is the *dependent* and $arc_{ij}$ represents the *dependency type*, which denotes the syntactic function that relates the head and the dependent. In this way, the phrase *'best performance'* could be represented syntactically as *(performance, modifier, best)*. We rely on MaltParser [7], a data-driven dependency parser generator, to build our parsers. We used again the Ancora and the Penn Treebank corpora to train the Spanish and the English parser, respectively. Our aim is to employ dependency parsing to capture the non-local relations between words that lexical-based approaches cannot handle properly.

### 2.2 Feature extraction

Our classifiers are fed with three different types of features:

- *N-grams*: This type of features detect the presence of sequences of contiguous words, where $n$ is the number of concatenated terms. In this paper, we consider both 1-grams and 2-grams (which make it possible to capture some contextual information based on word proximity). Simple normalisation techniques such as converting words to their lowercase form are applied. In addition to n-grams of words, we also consider n-grams of lemmas[3]. The aim is to reduce sparsity and training more accurate classifiers, specially for Spanish language, where verbs, adjectives and nouns present gender and number declensions.
- *Psycometric properties*: The LIWC [8] is a software that can be used to identify psychometric word properties present in a text. Among other languages, it provides dictionaries for both Spanish and English. We use those dictionaries in this work to relate words with psychological features such as *insight*, *anger* or *happiness*, but also with topics such as *money*, *sports* or *religion*.

[1] http://www.nltk.org/
[2] http://textblob.readthedocs.org/en/dev/
[3] Lemmas are the canonical forms of words. For example, the lemma of *'walking'* is *'walk'*

In this way, we match the words of a text, returning all their psychometric dimensions.

– *Generalised dependency triplets*: In this paper, we apply an enriched approach presented at [9] of the initial method described at [10]. Given a dependency triplet of the form $(w_i, arc_{ij}, w_j)$ a generalised triplet has the form $(g(w_i, x), d(arc_{ij}), g(w_j, x))$, where $g$ is a generalisation function and $x$ the desired type of generalisation, which can be: the *word* itself, its *lemma*, its *psychometric properties*, its *part-of-speech tag* or *none*, if we decide to completely delete the content of the token. On the other hand, the function $d$ can be defined to keep or remove the dependency type of a triplet. For example, the triplet *(performance, modifier, best)* can be generalised as *(optimism, modifier, adjective)* by applying the generalisation functions *(g(performance, psychometric properties), modifier, g(best, part-of-speech tag))*. The goal is to reduce the sparsity of standard dependency triplets, generalising concepts and ideas in a homogeneous way.

In all cases, we use the number of occurrences as the weighting factor for the supervised classifier.

### 2.3 Classifier

We use the WEKA [11] framework for building our classifiers. For each task, we tuned the weights and the kernel of the classifier in order to maximise performance, as detailed in the following sections.

## 3 Task 1: Reputation Dimensions Categorisation

The task consisted on relating tweets with the standard reputation dimensions proposed by the Reputation Institute and the RepTrak model[4]: *products&services, innovation, workplace, citizenship, governance, leadership, performance and undefined* (if a tweet is no assigned to any of the other dimensions).

**Dataset** The RepLab 2014 corpus is composed of English and Spanish tweets extracted from the RepLab 2013 corpus, which contained a collection of tweets referring to up to 61 entities. The RepLab 2014 only takes into account those who refer to banking or automotive entities, where each one is labelled with one of the standard reputation dimensions. To create the collection the canonical name of the entity was used as a query to retrieve the tweets which talk about it. Thus, each tweet contains the name of an entity. In addition, the corpus provides information about the author of each tweet, the content of external links that appear in a message and a flag to know if the tweet is written in English or Spanish.

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

**Evaluation metrics** This task is evaluated as a multi-class categorisation problem. Thus, precision, recall and accuracy are the official metrics:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where TP and TN refer to the true positives and negatives and FP and FN indicate the false positives and negatives, respectively. The organisers sorted the official results by accuracy.

**Runs** We sent two runs. For each run, we trained two different LibLinear classifiers [12]: one for English and another one for Spanish language. We tuned the weights for the majority classes (*products, citizenship, undefined and governance*) using a value of 0.75, giving the less frequent categories a weight of 1. In both cases, our approaches only handle the content of a tweet, discarding the user information and the content of the external links. In the latter case, we think processing the content of the web pages referred to in a tweet may excessively increase the cost of analysing a tweet. In addition, we believe the tweet reputation dimensions are not necessarily to be related with the content of the link, where probably many concepts and ideas appear. The results presented below these lines seem to confirm our hypothesis since we ranked 3rd, very close to the best-performing system. More specifically, our contributions were:

- *Run 1*: The English model took as features: unigrams of lemmas, bigrams of lemmas, and word psychometric properties. With respect to the Spanish classifier, the experimental setup showed that the best-performing model over Spanish messages was composed of: unigrams of lemmas, bigrams of lemmas and generalised triplets of the form (_, dependency type, lemma), *i.e.*, dependency triplets where the head is omitted. In both cases, we tried to obtain the best sets of features via greedy search on the training corpus and a 5-fold cross-validation.
- *Run 2*: This model uses the same classifier and the same sets of features as run 1, but excluding those which include the name of any of the entities used to create the train corpus. Our main aim was protecting our model from a possible bias on the training corpus. We observed that many tweets belonging to certain entities were labelled mainly only into a single reputation dimension. We were concerned that this fact could create an overfitted model which would not work properly on the test set. In this respect, this run also allowed us to measure the impact on performance of using the name of entities on the test set.

**Results** Table 3 shows the ranking of the systems for the reputation dimension task, based on their accuracy. The baseline of the RepLab organisation is a naive bag-of-words approach trained on a Support Vector Machine (SVM). Our run 1 ranked 3rd, confirming the effectiveness of our perspective. The second run also worked acceptably, although performance dropped by almost two percentage points. This confirms a slight bias on the test set, since it contains tweets that refer to the same entities as the training set and they were collected in the same interval of time. Table 3 show the detailed performance for our best run. Our model obtains both an acceptable recall and precision for the most prevalent classes, but the same is not true for minority classes, due to the small number of samples in the training set. The majority of the participants exhibited this same weakness.

## 4 Task 2.1: Author categorisation

The goal of the task was to assign Twitter profiles to one of these categories: *journalist, professional, authority, activist, investor, company* and *celebrity*. An additional class *undecidable* was proposed to place all those users that did not match any of the proposed categories.

**Dataset** The training and the test set are composed of the authors who wrote the automotive and banking tweets that we mentioned previously. In addition to user information, the organizers included the identifiers of the last 600 tweets of each user at the moment of the creation of the corpus. Due to the lack of time, we decided to download only 100 tweets for each author. In order to obtain these tweets faster, we used the capabilities of the Twitter API to download the timeline of an author instead of downloading the tweets one by one. However, that API method only allows the user to obtain the 3 200 most recent tweets of each author, so we were unable to find the tweets included in the corpus for many of them (the most active ones). More specifically, we could retrieve no tweets for around 1 000 authors.

**Evaluation metrics** The official results are the average accuracy between the categories corresponding to automotive and banking. Only the authors categorised as *influential* in the gold standard of task 2.2 are taken into account.

**Runs** This task is addressed as follows: given a set of tweets for an author, they are collected into a single file, which is used to finally classify the user according to the proposed categories. Since many of the categories in the training corpus only contained a few authors, we discarded those classes in order to avoid confusing machine learning algorithm. We trained two classifiers, one for each language. After, testing different Support Vector Machine implementations, we obtained the best performance on the training set (5-fold cross-validation) using an SMO [13].

**Table 1.** Ranking for task 1: Reputation Dimensions Categorisation

| Team | Run | Accuracy |
|---|---|---|
| uogTr | 4 | 0,731 |
| DAE | 1 | 0,723 |
| **LyS** | **1** | **0,717** |
| SIBtex | 1 | 0,707 |
| CIRGIRDISCO | 3 | 0,707 |
| SIBtex | 2 | 0,705 |
| stavicta | 4 | 0,703 |
| DAE | 4 | 0,703 |
| **LyS** | **2** | **0,699** |
| stavicta | 1 | 0,695 |
| CIRGIRDISCO | 1 | 0,692 |
| uogTr | 5 | 0,687 |
| stavicta | 2 | 0,685 |
| UvA | 4 | 0,668 |
| stavicta | 3 | 0,662 |
| UvA | 5 | 0,659 |
| UvA | 1 | 0,654 |
| UvA | 2 | 0,647 |
| UvA | 3 | 0,622 |
| baseline-replab | | 0,622 |
| uogTr | 2 | 0,621 |
| lia | 2 | 0,613 |
| uogTr | 3 | 0,609 |
| lia | 5 | 0,607 |
| CIRGIRDISCO | 2 | 0,607 |
| lia | 4 | 0,596 |
| DAE | 2 | 0,586 |
| DAE | 5 | 0,586 |
| lia | 1 | 0,549 |
| uogTr | 1 | 0,496 |
| lia | 3 | 0,357 |

**Table 2.** Detailed performance for our best run on the Reputation Dimensions Categorisation task

| Category | Recall | Precision | #tweets | % tweets |
|---|---|---|---|---|
| Innovation | 0.085 | 0.271 | 306 | 1.09 |
| Citizenship | 0.732 | 0.848 | 5027 | 17.89 |
| Leadership | 0.200 | 0.484 | 744 | 2.65 |
| Workplace | 0.274 | 0.527 | 1124 | 4.00 |
| Governance | 0.461 | 0.697 | 3395 | 12.08 |
| Performance | 0.404 | 0.499 | 1598 | 5.69 |
| Products&Services | 0.879 | 0.702 | 15903 | 56.60 |

To identify which authors are Spanish and which ones are English, for each author we counted the number of his last 600 tweets included at the corpus that are written in each language, assigning the author to the most frequent one. This information is provided by the RepLab 2014 organisation, without any need to download the tweets. More specifically, as we did in task 1, we sent two runs:

- *Run 1*: Both the Spanish and the English models use unigrams of lemmas and psychometric properties as features. We selected these features via greedy search on our processed training corpus (where all tweets of a user are merged into a single file). Since we did not have any tweet for many authors, we trained a back-off machine learner: a bag-of-words classifier which categorises these authors according to their profile description.
- *Run 2*: The only difference with respect to run 1 is the back-off classifier. Authors for which we have not downloaded any tweet are always assigned to the majority class in the training corpus: *undecidable*.

**Results** Table 4 shows the performance of the systems participating in this task. We think that our poor performance is due to the small site of the training corpus that we were able to collect and process. The baseline proposed by the RepLab organisers reinforces our hypothesis, since they used an SVM approach based on a bag-of-words. They also included another baseline which assigns all authors to the majority class in the training corpus.

**Table 3.** Ranking for task 2.1: Author Categorisation

| Team | Run | Automotive | Banking | Miscelaneous | Average |
|---|---|---|---|---|---|
| lia | 1 | 0,445 | 0,503 | 0,462 | 0,474 |
| baseline-replab | | 0,426 | 0,495 | - | 0,461 |
| baseline-most frequent | | 0,45 | 0,42 | 0,51 | 0,435 |
| UAM-CALYR | 2 | 0,382 | 0,446 | 0,392 | 0,414 |
| UAM-CALYR | 1 | 0,386 | 0,421 | 0,415 | 0,404 |
| ORM_UNED | 1 | 0,374 | 0,41 | 0,392 | 0,392 |
| ORM_UNED | 3 | 0,389 | 0,392 | 0,177 | 0,391 |
| lia | 2 | 0,357 | 0,398 | 0,377 | 0,377 |
| ORM_UNED | 2 | 0,352 | 0,389 | 0,300 | 0,371 |
| lia | 3 | 0,293 | 0,308 | 0,369 | 0,301 |
| **LyS** | **1** | **0,142** | **0,153** | **0,254** | **0,147** |
| **LyS** | **2** | **0,131** | **0,137** | **0,223** | **0,134** |

## 5 Task 2.2: Author ranking

The task focusses on classifying authors a *influential* and *non-influential*, as well as ranking them according to that level of influence.

**Dataset** It is the same that the employed at task 2.1: Author Categorisation. The proportion in the training set is about 30% of influential users, with the remaining 70% being non-influential.

**Evaluation metrics** The organisers address the problem as a traditional ranking information problem using the Mean Average Precision (MAP) as standard metric. The experimental results are ordered according to the average of automotive and banking MAP measures.

**Runs** Classification of influential and non-influential users is made via a LibLinear classifier, following a machine learning perspective. To rank the authors we take as the starting point the confidence factor reported by the classifier for each sample. The confidence is then used to rank the users according to their level of influence. A higher confidence should indicate a higher influence. With respect to non-influential users, we firstly negate that factor, obtaining in this way lower values for the least influential authors. We again sent two models to evaluate this task, although in this case the runs present significant differences:

- *Run 1*: A bag-of-words model which takes each word of the Twitter profile descriptions to feed the supervised classifier. The weights of the classes were tuned taking 1.8 and 1.3 for influential an non-influential users, respectively. Since the corpus is domain-dependent (automotive and banking tweets) we hypothesise that the brief biography of the user may be an acceptable indicator of influence. We observed that words such as *'car'*, *'business'* or *magazine* were some of the most relevant tokens in terms of information gain.
- *Run 2*: This run follows a meta-information perspective, taking the information provided by the Twitter API for any user. More specifically, we used binary features such as: *URL in the Twitter profile, verified account, profile user background image, default profile, geo enabled, default profile image, notifications, is translation enabled* and *contributors enabled*. In addition the following numeric features are taken into account: *listed count, favourites count, followers count, statuses count, friends count* and *following*.

**Results** Table 5 illustrates the official results for this task. The baseline of the RepLab organisers ranks the authors by their number of followers. Our run 1 achieved the 2nd place, tied in practice with the 1st place, reinforcing the validity of the proposal for a specific domain. On the other hand, our second run did not work as expected, although it outperformed the baseline.

## 6 Conclusions

This paper describes the participation of the LyS research group at RepLab 2014. We sent runs for all tasks proposed. The classification for the reputation dimensions task is addressed from a NLP perspective, including preprocessing,

**Table 4.** Ranking for task 2.2: Author ranking

| Team | Run | MAP |
|---|---|---|
| UTDBRG | 4 | 0.565 |
| **LyS** | **1** | **0.563** |
| UTDBRG | 1 | 0.550 |
| UTDBRG | 5 | 0.503 |
| UTDBRG | 3 | 0.499 |
| Lia | 1 | 0.476 |
| UAM-CALYR | 5 | 0.465 |
| UAM-CALYR | 1 | 0.436 |
| UAM-CALYR | 2 | 0.436 |
| UTDBRG | 2 | 0.413 |
| **LyS** | **2** | **0.403** |
| UAM-CALYR | 3 | 0.381 |
| UAM-CALYR | 4 | 0.381 |
| baseline-replab | | 0.378 |
| ORM_UNED | 3 | 0.349 |

part-of-speech tagging and dependency parsing. We use the output obtained by our NLP pipeline for extracting lexical, psychometric and syntactic-based features, which are used to feed a supervised classifier. We ranked 3rd, very close to the best performing system, confirming the effectiveness of the approach.

The author categorisation task is addressed from the same perspective. However, we could not properly exploit the approach due to problems to obtain much of the content of the training corpus.

On the other hand, the author ranking challenge was addressed from a different perspective. We obtained the second best-performing system, tied in practice with the 1st place, by training a bag-of-words classifier which takes the Twitter profile description of the users as features. This model clearly outperformed our second run based on metadata such as the number of favourited tweets or followers.

## Acknowledgements

Research reported in this paper has been partially funded by Ministerio de Economía y Competitividad and FEDER (Grant TIN2010-18552-C03-02) and by Xunta de Galicia (Grant CN2012/008).

## References

1. E. Amigó, J. Carrillo-de-Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, 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, 2014.

2. M. Taulé, M. A. Martí, M. Recasens, AnCora: Multilevel Annotated Corpora for Catalan and Spanish, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.
3. M. P. Marcus, M. A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of English: The Penn treebank, Computational linguistics 19 (2) (1993) 313–330.
4. E. Brill, A simple rule-based part of speech tagger, in: Proceedings of the workshop on Speech and Natural Language, HLT'91, Association for Computational Linguistics, Stroudsburg, PA, USA, 1992, pp. 112–116. doi:10.3115/1075527.1075553.
5. D. Vilares, M. A. Alonso, C. Gómez-Rodríguez, A syntactic approach for opinion mining on Spanish reviews, Natural Language Engineering. Available on CJO2013. doi:10.1017/S1351324913000181.
6. M. Collins, Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 1–8. doi:10.3115/1118693.1118694.
URL http://dx.doi.org/10.3115/1118693.1118694
7. J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, E. Marsi, Maltparser: A language-independent system for data-driven dependency parsing., Natural Language Engineering 13 (2) (2007) 95–135.
8. J. Pennebaker, M. Francis, R. Booth, Linguistic inquiry and word count: LIWC 2001, Mahway: Lawrence Erlbaum Associates (2001) 71.
9. D. Vilares, M. A. Alonso, C. Gómez-Rodríguez, On the usefulness of lexical and syntactic processing in polarity classification of twitter messages, Journal of the Association for Information Science Science and Technology to appear.
10. M. Joshi, C. Penstein-Rosé, Generalizing dependency features for opinion mining, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 313–316.
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, SIGKDD Explorations 11 (1) (2009) 10–18. doi:10.1145/1656274.1656278.
URL http://doi.acm.org/10.1145/1656274.1656278
12. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, The Journal of Machine Learning Research 9 (2008) 1871–1874.
13. J. C. Platt, Advances in kernel methods, MIT Press, Cambridge, MA, USA, 1999, Ch. Fast training of support vector machines using sequential minimal optimization, pp. 185–208.