Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management

Enrique Amigó[1], Jorge Carrillo-de-Albornoz[1], Irina Chugur[1], Adolfo Corujo[2], Julio Gonzalo[1], Edgar Meij[3], Maarten de Rijke[4], and Damiano Spina[1]

[1] UNED NLP & IR Group
Juan del Rosal, 16. 28040 Madrid, Spain, nlp.uned.es
[2] Llorente & Cuenca
Lagasca, 88. 28001 Madrid, Spain, www.llorenteycuenca.com
[3] Yahoo Labs
Avinguda Diagonal 177, 08018 Barcelona, Spain, labs.yahoo.com
[4] University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands, ilps.science.uva.nl

Abstract. This paper describes the organisation and results of RepLab 2014, the third competitive evaluation campaign for Online Reputation Management systems. This year the focus lied on two new tasks: reputation dimensions classification and author profiling, which complement the aspects of reputation analysis studied in the previous campaigns. The participants were asked (1) to classify tweets applying a standard typology of reputation dimensions and (2) categorise Twitter profiles by type of author as well as rank them according to their influence. New data collections were provided for the development and evaluation of systems that participated in this benchmarking activity.

Keywords: RepLab, Reputation Management, Evaluation Methodologies and Metrics, Test Collections, Reputation Dimensions, Author Profiling, Twitter

1 Introduction

RepLab is a competitive evaluation exercise supported by the EU project LiMoSINe.[5] It aims at encouraging research on Online Reputation Management and providing a framework for collaboration between academia and practitioners in the form of a "living lab": a series of evaluation campaigns in which task design and evaluation are jointly carried out by researchers and the target user community (in our case, reputation management experts). Similar to the previous campaigns [1,2], RepLab 2014 was organized as a CLEF lab.[6]

Previous RepLab editions focused on problems such as entity resolution (resolving name ambiguity), topic detection (what are the issues discussed about

[5] http://www.limosine-project.eu
[6] http://clef2014.clef-initiative.eu/

the entity?), polarity for reputation (which statements and opinions have negative/positive implications for the reputation of the entity?) and alert detection (which are the issues that might harm the reputation of the entity?). Although online monitoring pervades all online media (news, social media, blogosphere, etc.), RepLab has always been focused on Twitter content, as it is the key media for early detection of potential reputational issues.

In 2014, RepLab focused on two additional aspects of reputation analysis – reputation dimensions classification and author profiling – that complement the tasks tackled in the previous campaigns. As we will see below, reputation dimensions contribute to a better understanding of the topic of a tweet or group of tweets, whilst author profiling provides important information for priority ranking of tweets, as certain characteristics of the author can make a tweet (or a group of tweets) an alert, requiring special attention of reputation experts. Section 2 explains the tasks in more detail. A description of the data collections created for RepLab 2014 and chosen evaluation methodology can be found in Sections 3 and 4, respectively. In Section 5, we briefly review the list of participants and employed approaches. Section 6 is dedicated to the display and analysis of the results, based on which we, finally, draw conclusions in Section 7.

2 Tasks Definition

In 2014, RepLab offered its participants the following tasks: (1) classification of Twitter posts by reputation dimension and (2) classification and ranking of Twitter profiles.

2.1 Reputation Dimensions Classification

The aim of this task is to assign tweets to one of the seven standard reputation dimensions of the RepTrak Framework[7] developed by the Reputation Institute. These dimensions reflect the affective and cognitive perceptions of a company by different stakeholder groups. The task can be viewed as a complement to topic detection, as it provides a broad classification of the aspects of the company under public scrutiny. Table 1 shows the definition of each reputation dimension, supported by an example of a labelled tweet:

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

Table 1: RepTrak dimensions. Definitions and examples of tweets.

Dimension	Definition and Example
Performance	Reflects long term business success and financial soundness of the company. `Goldman Profit Rises but Revenue Falls: Goldman Sachs reported a second-quarter profit of $1.05 billion,... http://dlvr.it/bmVY4`
Products & Services	Information about the company's products and services, as well as about consumer satisfaction. `BMW To Launch M3 and M5 In Matte Colors: Red, Blue, White but no black...`
Leadership	Related to the leading position of the company. `Goldman Sachs estimates the gross margin on ACI software to be 95% O…o`
Citizenship	The company's acknowledgement of the social and environmental responsibility, including ethical aspects of business: integrity, transparency and accountability. `Find out more about Santander Universities scholarships, grants, awards and SME Internship Programme bit.ly/1mMl2OX`
Governance	Related to the relationship between the company and the public authorities. `Judge orders Barclays to reveal names of 208 staff linked to Libor probe via @Telegraph soc.li/mJVPh1R`
Workplace	Related to the working environment and the company's ability to attract, form and keep talented and highly qualified people. `Goldman Sachs exec quits via open letter in The New York Times, brands bank working environment ‘‘toxic and destructive’’ ow.ly/9EaLc`
Innovation	The innovativeness shown by the company, nurturing novel ideas and incorporating them into products. `Eddy Merckx Cycles announced a partnership with Lexus to develop their ETT Hme trial bike. More info at...http://fb.me/1VAeS3zJP`

2.2 Author Profiling

This task is composed of two subtasks that were evaluated separately.

Author Categorisation. The task was to classify Twitter profiles by type of author: Company (i.e., corporate accounts of the company itself), Professional (in the economic domain of the company), Celebrity, Employee, Stockholder, Investor, Journalist, Sportsman, Public Institution, and Non-Governmental Organisation (NGO). The systems' output was expected to be a list of profile identifiers with the assigned categories, one per profile.

Author Ranking. Using as input the same set of Twitter profiles as in the task above, systems had to find out which authors had more reputational influence

(who the influencers or opinion makers are) and which profiles are less influential or have no influence at all. For a given domain (e.g., automotive or banking), the systems' output was a ranking of profiles according to their probability of being an opinion maker with respect to the concrete domain, optionally including the corresponding weights. Note that, because the number of opinion makers is expected to be low, we modelled the task as a search problem (hence the system output is a ranked list) rather than as a classification problem.

Some aspects that determine the influence of an author in Twitter (from a reputation analysis perspective) can be the number of followers, number of comments on a domain or type of author. As an example, below is the profile description of an influential financial journalist:

> **Description:** New York Times Columnist & CNBC Squawk Box (@SquawkCNBC) Co-Anchor. Author, Too Big To Fail. Founder, @DealBook. Proud father. RTs endorsements
> **Location:** New York, New York nytimes.com/dealbook
> **Tweets:** 1,423
> **Tweet examples:**
> "Whitney Tilson: Evaluating the Dearth of Female Hedge Fund Managers http://nyti.ms/1gpClRq @dealbook"
> "Dina Powell, Goldman's Charitable Foundation Chief to Lead the Firm's Urban Investment Group http://nyti.ms/1fpdTxn @dealbook"

Shared PAN-RepLab Author Profiling: Participants were also offered the opportunity to attempt the shared author profiling task RepLab@PAN.[8] In order to do so, systems had to classify Twitter profiles by gender and age. Two categories, female and male, were used for gender. Regarding age, the following classes were considered: 18-24, 25-34, 35-49, 50-64, and 65+.

3 Data Sets

This section briefly describes the data collections used in each task. Note that the current amount of available tweets may be lower, as some posts may have been deleted or made private by the authors: in order to respect the Twitter's terms of service (TOS), we did not provide the contents of the tweets, but only tweet ids and screen names. Tweet texts can be downloaded using any of the following tools:

1. TREC Microblog Track[9]
2. SemEval-2013 Task 2 Download script[10]
3. A Java tool provided by the RepLab organisers[11]

[8] http://pan.webis.de/
[9] https://github.com/lintool/twitter-tools
[10] http://www.cs.york.ac.uk/semeval-2013/task2/index.php?id=data
[11] http://nlp.uned.es/replab2013/replab2013_twitter_texts_downloader_latest.tar.gz

3.1 Reputation Dimensions Classification Data Set

This data collection is based on the RepLab 2013 corpus[12] and contains over 48,000 manually labelled English and Spanish tweets related to 31 entities from the automotive and banking domains. The tweets were crawled from the 1st June 2012 to the 31st Dec 2012 using the entity's canonical name as query. The balance between languages depends on the availability of data for each entity. The distribution between the training and test sets was established as follows. The training set was composed of 15,562 Twitter posts and 32,446 tweets were reserved for the test set. Both data sets were manually labelled by annotators trained and supervised by experts in Online Reputation Management from the online division of a leading Public Relations consultancy Llorente & Cuenca.[13]

The tweets were classified according to the RepTrak dimensions[14] listed in Section 2. In case a tweet could not be categorised into any of these dimensions, it was labelled as "Undefined".

The reputation dimensions corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets temporally situated between the training (earlier tweets) and test material (the latest tweets) in the timeline.

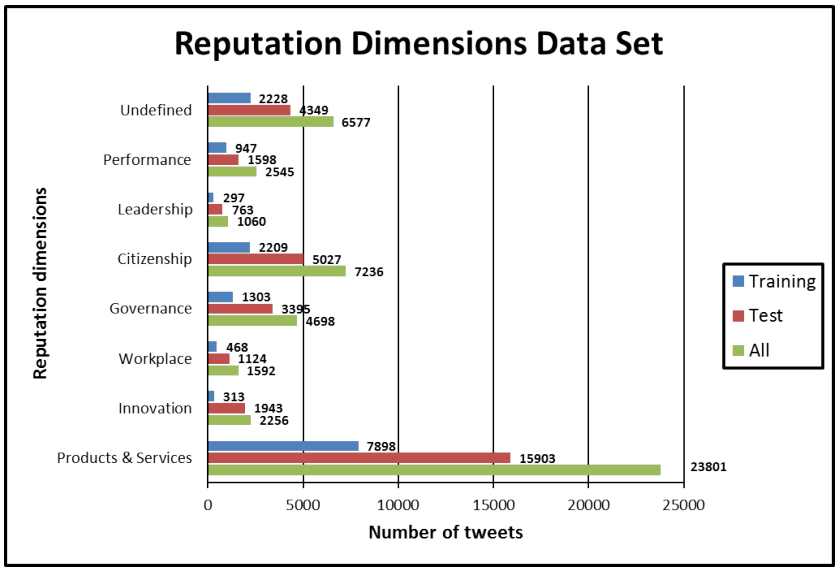

Fig. 1: Distribution of classes in the Reputation Dimensions data set.

[12] http://nlp.uned.es/replab2013
[13] http://www.llorenteycuenca.com/
[14] http://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework

Figure 1 shows the distribution of the reputation dimensions in the training and test sets, and in the whole collection. As can be seen, the Products & Services dimension is the majority class in both data sets, followed by the Citizenship and Governance. The large number of tweets associated with the Undefined dimension in both sets is noteworthy, which suggests the complexity of the task, as even human annotators could not specify the category of 6,577 tweets.

3.2 Author Profiling Data Set

This data collection contains over 7,000 Twitter profiles (all with at least 1,000 followers) that represent the automotive, banking and miscellaneous domains. The latter includes profiles from different domains. The idea of this extra set is to evaluate if approaches designed for a specific domain are suitable for a broader multi-domain scenario. Each profile contains (i) screen name; (ii) profile URL, and (iii) the last 600 tweets published by the author at crawling time.

The collection was split into training and test sets: 2,500 profiles in the training set and 4,991 profiles in the test set. Reputation experts performed manual annotations for two subtasks: *Author Categorisation* and *Author Ranking*. First, they categorised profiles as company (i.e., corporate accounts of companies), professional, celebrity, employee, stockholder, journalist, investor, sportsman, public institution, and non-governmental organisation (NGO). In addition, reputation experts manually identified the opinion makers (i.e., authors with reputational influence) and annotated them as "Influencer". The profiles that were not considered opinion makers were assigned the "Non-Influencer" label. Those profiles that could not be classified into one of these categories, were labelled as "Undecidable".

The distribution by classes in the Author Categorisation data collection is shown in Figure 2. As can be seen, Professional and Journalist are the majority classes in both training and test sets, followed by the Sportsman, Celebrity, Company and NGO. Surprisingly, the number of authors in the categories Stockholder, Investor and Employee is considerably low. One possible explanation is that such authors are not very active on Twitter, and more specialized forums need to be considered in order to monitor these types of users.

Regarding the distribution of classes in the Author Ranking dataset, Table ?? shows the number of authors labelled as Influencer and Non-Influencer in the training and test sets. The proportion of influencers is much higher than we expected, and calls for a revision of our decision to cast the problem as search (find the influentials) rather than classification (classify as influential or non-influential).

3.3 Shared PAN-RepLab Author Profiling Data Set

For the shared PAN-RepLab author profiling task, 159 Twitter profiles from several domains were annotated with gender (female and male) and age (18-24, 25-34, 35-49, 50-64, and 65+). The profiles were selected from the RepLab

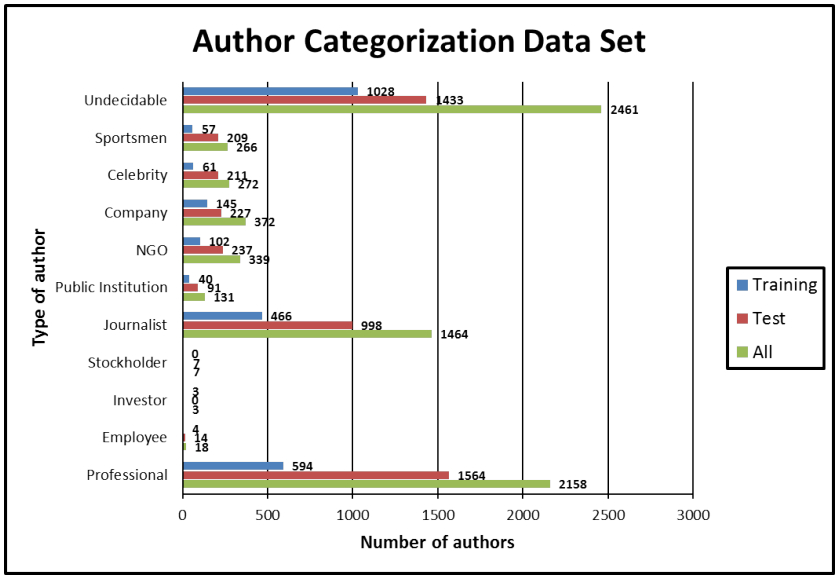

Fig. 2: Distribution of classes in the Author Categorisation data set.

Table 2: Distribution of classes in the Author Ranking data set.

	Influencer	**Non-Influencer**
Training	796	1704
Test	1563	3428
All	2359	5132

2013 test collection and from a list of influential authors provided by Llorente & Cuenca.

131 profiles were included into the miscellaneous data set of the RepLab author profiling data collection accompanied by the last 600 tweets published by the authors at crawling time. 28 users had to be discarded because more than 50% of their tweets were written in languages other than English or Spanish. The selected 131 profiles, in addition to age and gender, were manually tagged by reputation experts as explained in Section 3.2: with (1) type of author and (2) opinion-maker labels.

4 Evaluation Methodology

4.1 Baselines

For both classification tasks — Reputation Dimensions and Author Categorisation — a simple Bag-of-Words (BoW) classifier was proposed as official baseline.

We used Support Vector Machines,[15] with a linear kernel. The penalty parameter C was automatically adjusted by weights inversely proportional to class frequencies in the training data. We used the default values for the rest of parameters.

For the Reputation Dimensions task, a different multi-class tweet classifier was built for each entity. Tweets were represented as BoW with binary occurrence (1 if the word is present in the tweet, 0 if not). The BoW representation was generated by removing punctuation, lowercasing, tokenizing by white spaces, reducing multiple repetitions of characters (from n to 2) and removing stopwords.

For the Author Categorisation task, a different classifier was built for each domain in the training set (i.e., banking and automotive). Here, each Twitter profile was represented by the latest 600 tweets provided with the collection. Then, the built pseudo-documents were preprocessed as described before.

Finally, the number of followers of each Twitter profile has been used as baseline for the Author Ranking task.

4.2 Evaluation Measures

Reputation Dimensions Categorisation This task is a multi-class classification problem and its evaluation is an open issue. The traditional *Accuracy* measure presents drawbacks for unbalanced data. On the other hand, the commonly used *F-measure* over each of the classes does not allow to produce a global system ranking. In this evaluation campaign we chose *Accuracy* as the official measure for the sake of interpretability. It is worth mentioning that, in the Reputation Dimensions task, systems outperformed a most-frequent baseline which always selects the majority class labels (see Section 6.1).

Author Categorisation Similar to the Reputation Dimensions, the first subtask of Author Profiling is a categorization task. We also used *Accuracy* as the official evaluation measure. However, the obtained empirical results suggest that *Accuracy* is not able to discriminate system outputs from the majority class baseline. For this reason, the results were complemented with *Macro Average Accuracy* ($MAAC$), which penalizes non-informative runs.

Author Ranking The second subtask of Author Profiling is a ranking problem. Influential authors must be located at the top of the system output ranking. This is actually a traditional information retrieval problem, where relevant and irrelevant classes are not balanced. Studies on information retrieval measures can be applied in this context, although author profiling differs from information retrieval in a number of aspects. The main difference (which is a post-annotation finding) is that the ratio of relevant authors is much higher than the typical ratio of relevant documents in a traditional information retrieval scenario.

[15] http://scikit-learn.org/stable/modules/svm.html

Another differentiating characteristic is that the set of potentially influential authors is rather small, while information retrieval test sets usually consist of millions of documents. This has an important implication for the evaluation methodology. All information retrieval measures state a weighting scheme which reflects the probability of users to explore a deepness level in the system's output ranking. In the Online Reputation Management scenario, this deepness level is still not known. We decided to use *MAP* (*Mean Average Precision*) for two reasons. First, because it is a well-known measure in information retrieval. Second, because it is recall-oriented and also considers the relevance of authors at lower ranks.

5 Participation

49 groups signed in for RepLab 2014, although only 11 of them (from 9 different countries) finally submitted results in time for the official evaluation. Overall, 8 groups participated in the Reputation Dimensions task, and 5 groups submitted their results to Author Profiling (all of them submitted to the author ranking subtask, and all but one to the author categorization subtask).

Table 3 shows the acronyms and affiliations of the research groups that participated in RepLab 2014. In what follows, we list the participants and briefly describe the approaches they used.

Table 3: List of participants: acronyms and affiliations.

Acronym	**Affiliation**	**Country**
CIRGIRDISCO	National University of Ireland, Galway	Ireland
DAE	Daedalus, S.A.	Spain
LIA	University of Avignon	France
LyS	Departamento de Computación, Universidade da Coruña	Spain
ORM_UNED	Universidad Nacional de Educación a Distancia	Spain
STAVICTA	Linnaeus University,Växjö and Lund University	Sweden
UAMCLYR	Universidad Autónoma Metropolitana Cuajimalpa	Mexico
uogTr	School of Computing Science, University of Glasgow	United Kingdom
UTDBRG	University of Tehran	Iran
UvA	ISLA, University of Amsterdam	The Netherlands
SIBtex	SIB Swiss Institute of Bioinformatics, Genève University of Applied Sciences, Carouge	Switzerland

CIRGIRDISCO participated in the Reputation Dimensions task. They used dominant Wikipedia categories related to a reputation dimension in a Random Forest classifier. Additionally, they also applied tweet-specific, language-specific and similarity-based features. The best run significantly improved over the baseline accuracy.

DAE attempted the Reputation Dimensions Classification. Their initial idea was to evaluate the best combination strategy of a machine learning classifier with a rule-based algorithm that uses logical expressions of terms. However, the baseline experiment employing just Naive Bayes Multinomial with a term vector model representation of the tweet text was ranked second among runs from all participants in terms of *Accuracy*.

LIA carried out a considerable number of experiments for each task. The proposed approaches rely on a large variety of machine learning methods. The main accent was put on exploiting tweet contents. Several methods also included selected metadata. Marginally, external information was considered by using provided background messages.

LyS attempted all the tasks. For Dimensions Classification and Author Categorisation a supervised classifier was employed with different models for each task and each language. A NLP perspective was adopted, including preprocessing, PoS tagging and dependency parsing, relying on them to extract features for the classifier. For author ranking, their best performance was obtained by training a bag-of-words classifier fed with features based on the Twitter profile description of the users.

ORM_UNED proposed a learning system based on voting model for the Author Profiling task. They used a small set of features based on the information that can be found in the text of tweets: POS tags, number of hashtags or number of links.

SIBtex integrated several tools into a complete system for tweet monitoring and categorisation which uses instance-based learning (K-Nearest Neighbours). Dealing with the domain (automotive or banking) and the language (English or Spanish), their experiments showed that even with all data merged into one single Knowledge Base (KB), the observed performances were close to those with dedicated KBs. Furthermore, English training data in addition to the sparse Spanish data were useful for Spanish categorisation.

STAVICTA devised an approach based on the textual content of tweets without considering metadata and the content of URLs for the reputation dimensions classification. They experimented with different feature sets including bag of n-grams, distributional semantics features, and deep neural network representations. The best results were obtained with bag of bi-gram features with minimum frequency thresholding. Their experiments also show that semi-supervised recursive auto-encoders outperform other feature sets used in the experiments.

UAMCLYR participated in the Author Profiling task. For Author Categorisation they used a supervised approach based on the information found in Twitter users' profiles. Employing attribute selection techniques, the most representative attributes from each user's activity domain were extracted. For Author Ranking they developed a two-step chained method based on stylistics attributes (e.g., lexical richness, language complexity) and behavioural attributes (e.g., posts' frequency, directed tweets) obtained from the users' profiles and posts. These attributes were used in conjunction with a Markov Random Fields to improve an initial ranking given by the confidence of Support Vector Machine classification algorithm.

uogTr investigated two approaches to the Reputation Dimensions classification. Firstly, they used a term's Gini-index score to quantify the term's representativeness of a specific class and constructed class profiles for tweet classification. Secondly, they performed tweet enrichment using a web scale corpus to derive terms representative of a tweet's class, before training a classifier with the enriched tweets. The tweet enrichment approach proved to be effective for this classification task.

UTDBRG participated in the Author Ranking subtask. The presented system utilizes a Time-sensitive Voting algorithm. The underlying hypothesis is that influential authors tweet actively about hot topics. A set of topics was extracted for each domain of tweets and a time-sensitive voting algorithm was used to rank authors in each domain based on the topics.

UvA addressed the Reputation Dimensions task by using corpus-based methods to extract textual features from the labelled training data to train two classifiers in a supervised way. Three sampling strategies were explored for selecting training examples. All submitted runs outperformed the baseline, proving that elaborate feature selection methods combined with balanced datasets help improve classification performance.

6 Evaluation Results

This section reports and analyses the results of the RepLab 2014 tasks, except for the shared PAN-RepLab author profiling, for which no submissions were received.

6.1 Reputation Dimensions Classification

Eight groups participated in the Reputation Dimensions task. 31 runs were submitted. Most approaches employed different machine learning algorithms such as Support Vector Machine (UvA, uogTr), Random Forest (CIRGIRDISCO, uogTr), Naive Bayes (DAE, UvA, STAVICTA), distance to class vectors (uogTr), LibLinear (LyS). SIBtex focussed on instance based learning techniques.

Regarding the employed features, some approaches considered information beyond the tweet textual content. For instance, uogTr expanded tweets with pseudo-relevant document sets and Wikipedia entries, CIRGIRDISCO employed Wikipedia categories, LyS considered psychometric dimensions and linguistic information such as dependency trees and part of speech. STAVICTA applied Distributional Semantic Models to expand tweets.

Table 4 shows the final ranking for the Reputation Dimensions task in terms of *Accuracy*. The last column represents the ratio of classified tweets from the set of tweets that were available at the time of evaluation. Note that tweets manually tagged as "Undefined" were excluded from the evaluation and tweets tagged by systems as "Undefined" were considered as non-processed.

Table 4: Official ranking for the Reputation Dimensions task.

Run	**Accuracy**	**Ratio of processed tweets**
uogTr_RD_4	0.73	0.99
DAE_RD_1	0.72	0.96
LyS_RD_1	0.72	0.91
SIBtex_RD_1	0.70	0.95
CIRGIRDISCO_RD_3	0.71	0.95
SIBtex_RD_2	0.70	0.95
STAVICTA_RD_4	0.70	0.89
DAE_RD_4	0.70	0.98
LyS_RD_2	0.69	0.92
stavicta_RD_1	0.69	0.88
CIRGIRDISCO_RD_1	0.69	0.94
uogTr_RD_5	0.69	0.99
stavicta_RD_2	0.68	0.89
UvA_RD_4	0.67	0.95
stavicta_RD_3	0.66	0.86
DAE_RD_3	0.66	0.91
UvA_RD_5	0.66	0.96
UvA_RD_1	0.65	0.91
UvA_RD_2	0.65	0.95
UvA_RD_3	0.62	0.94
Baseline-SVM	0.62	0.86
uogTr_RD_2	0.62	1
LIA_DIM_2	0.618	0.96
uogTr_RD_3	0.61	1
LIA_DIM_5	0.61	0.98
CIRGIRDISCO_RD_2	0.61	0.94
LIA_DIM_4	0.60	0.98
DAE_RD_2	0.59	0.82
DAE_RD_5	0.59	0.82
LIA_DIM_1	0.55	0.91
uogTr_RD_1	0.50	1
LIA_DIM_3	0.36	0.99
Majority class baseline	0.56	1

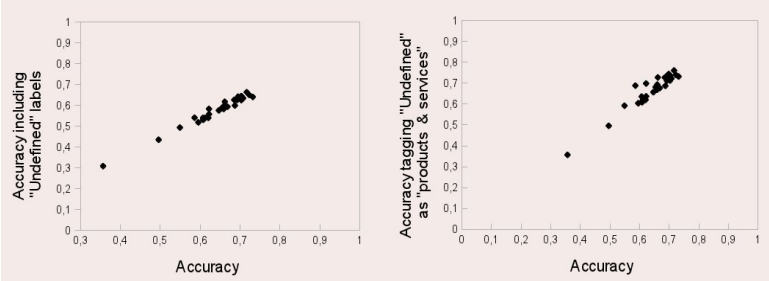

Fig. 3: Correspondence between the Accuracy results including "Undefined" or assigning them to the majority class. Each dot represents a run.

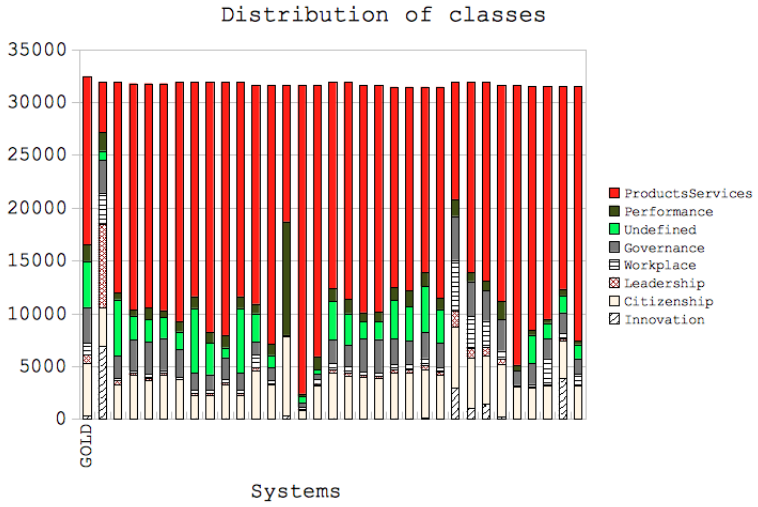

Fig. 4: Distribution of classes across the runs in the Reputation Dimensions task.

Besides participant systems, we included a baseline that employs Machine Learning (SVM) using words as features. Note that classifying every tweet as the most frequent class (majority class baseline) would get an accuracy of 56%. Most runs are above this threshold and provide, therefore, some useful information beyond a non-informative run.

There is no clear correspondence between performance and algorithms. The top systems used a variety of methods such as a basic Naive Bayes approach (DAE_RD_1), enrichment with pseudo-relevant documents (uogTR_RD_4) or multiple features including dependency relationships, POS tags, and psycometric dimensions (Lys_RD_1).

Given that tweets labelled as "Undefined" in the gold standard were not considered for evaluation purposes, systems that tagged tweets as "Undefined" had a negative impact on their performance. In order to check to what extent this affects the evaluation results, we computed *Accuracy* without considering this label. The leftmost graph in Figure 3 shows that there is a high correlation between both evaluation results across single runs. Moreover, replacing the "Undecidable" labels by "Product and Services" (majority class) also produces similar results (see rightmost graph in Figure 3).

Figure 4 illustrates the distribution of classes across the systems annotations and the goldstandard. As the figure shows, most of the systems tend to assign the majority class "Products and services" to a greater extent than the goldstandard.

6.2 Author Categorisation

Four groups participated in this task providing 10 official runs. Most of the runs are based on some kind of Machine Learning method over Twitter profiles. For instance, LIA employed Hiden Markov Models, Cosine distances with TF-IDF and Gini purity criteria, as well as Poisson modelling. UAMCLYR and LyS applied Support Vector Machine, and LyS used a combination of four algorithms: ZeroR, Random Tree, Random Forest and Naive Bayes.

As for features, the proposal of LyS includes term expansion with WordNet. ORM_UNED considered different metadata (e.g., profile domain, number of mentions, hashtags), and LyS included psychometric properties related to psychological dimensions (e.g., anger, happiness) and to topics such as money, sports, or religion.

Table 5: Accuracy of systems for the Author Categorisation task per domain.

Run	Automotive	Banking	Miscellaneous	Average (Aut.&Bank.)
LIA_AC_1	0.45	0.5	0.46	0.47
Baseline-SVM	0.43	0.49	-	0.46
Most frequent	0.45	0.42	0.51	0.44
UAMCLYR_AC_2	0.38	0.45	0.39	0.41
UAMCLYR_AC_1	0.39	0.42	0.42	0.4
ORM_UNED_AC_1	0.37	0.41	0.39	0.39
UAMCLYR_AC_3*	0.37	0.41	0.22	0.39
ORM_UNED_AC_3	0.39	0.39	0.18	0.39
UAMCLYR_AC_4*	0.36	0.41	0.19	0.39
LIA_AC_2	0.36	0.4	0.38	0.38
ORM_UNED_AC_2	0.35	0.39	0.3	0.37
LIA_AC_3	0.29	0.31	0.37	0.3
LyS_AC_1	0.14	0.15	0.25	0.15
LyS_AC_2	0.13	0.14	0.22	0.13

Table 5 shows the ranking for the Author Categorisation task. Two unofficial runs (submitted shortly after the deadline) are marked with an asterisk (*). The

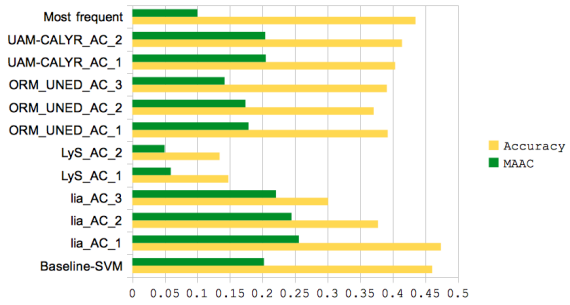

Fig. 5: Accuracy and MAAC for the Author Categorisation task.

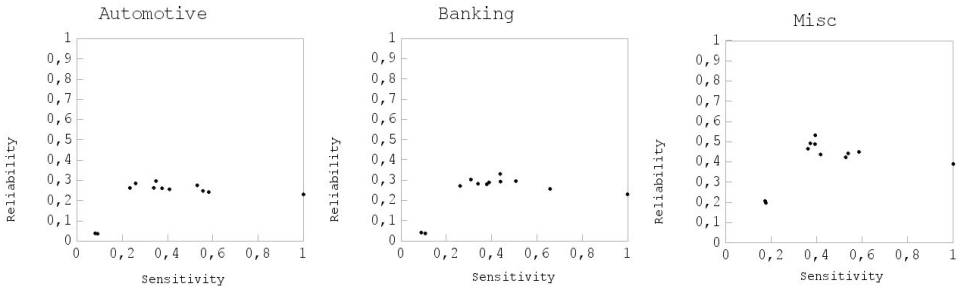

Fig. 6: Reliability and Sensitivity in the Author Categorisation task.

Accuracy values were computed separately for each domain (automotive, banking and miscellaneous). We included two baselines: Machine Learning (SVM) using words as features, and a baseline that assigns the most frequent class (in the training set) to all authors. Average *Accuracy* of the banking and automotive domains was used to rank systems.

Interestingly, there is a high correlation between system scores in the automotive vs. banking domains (0.97 *Pearson coefficient*). The low *Accuracy* values in the case of LyS are due to the fact that more than half of the authors were not included in the output file.

The most relevant aspect of these results is that, in terms of *Accuracy*, assigning the majority class outperforms most runs, although, of course, this output is not informative. The question, then, is how much information the systems are able to produce. In order to answer this question we have computed the *Macro Average Accuracy* (*MAAC*), which has the characteristic of assigning the same low score to any non informative classifier (e.g., random classification or one label for all instances). Figure 5 shows that most systems are able to improve the majority class baseline according to *MAAC*. This means that systems are able to produce information about classes, although they reduce the number of accurate decisions with respect to the majority class baseline.

From the grouping point of view, the majority class baseline relates all author to each other in the same class. On the other hand, systems try to identify more classes, increasing the correctness of grouping relationships at the cost of losing relationships. In order to analyse this aspect, we also calculated *Reliability* and *Sensitivity* of author relationships (*Bcubed Precision* and *Recall*), as if it was a clustering problem. *Reliaiblity* reflects the correctness of grouping relationships between authors. *Sensitivity* shows how many of these relationships are captured by the system.

The graphs in Figure 6 show the relationship between grouping precision and recall (R and S). The majority class baseline achieves the maximum *Sensitivity*, given that all authors are assigned to one class and, therefore, all relationships are captured, but including noisy relationships. As the graphs show, in general, systems are able to increase slightly the correctness of the produced relationships (*Reliability* increase), but at the cost of losing relationships (lower *Sensitivity*).

6.3 Author Ranking

Five groups participated in this task, for a total of 14 runs. The author influence estimation is grounded on different hypotheses. The approach proposed by LIA assumes that influencers tend to produce more opinionated terms in tweets. UTDBRG assumed that influential authors tweet more about hot topics. This requires a topic retrieval step and a time sensitive voting algorithm to rank authors. Some participants trained their systems over the biography text (LyS, UAMCLYR), binary profile metadata such as the presence of URLs, verified account, user image (LyS), quantitative profile metadata such as the number of followers (LyS, UAMCLYR), style-behaviour features such as the number of URLs, hashtags, favourites, retweets etc. (UAMCLYR).

Table 6: Mean Average Precision of systems in the Author Ranking task.

Run	Automotive	Banking	Miscellaneous	Average (Banking and Automotive)
UTDBRG_AR_4	0.72	0.41	0.00	0.57
LyS_AR_1.txt	0.60	0.52	0.68	0.56
UTDBRG_AR_1	0.70	0.40	0.00	0.55
UTDBRG_AR_5	0.69	0.32	0.00	0.50
UTDBRG_AR_3	0.68	0.32	0.00	0.50
LIA	0.50	0.45	0.65	0.48
UAM-CALYR_AR_5	0.44	0.49	0.77	0.47
UAM-CALYR_AR_1	0.45	0.42	0.77	0.44
UAM-CALYR_AR_2	0.45	0.42	0.77	0.44
UTDBRG_AR_2	0.46	0.37	0.00	0.41
LyS_AR_2	0.36	0.45	0.80	0.40
UAM-CALYR_AR_3	0.39	0.38	0.78	0.38
UAM-CALYR_AR_4	0.39	0.38	0.78	0.38
Followers	0.37	0.39	0.90	0.38
ORM_UNED_AR_3	0.38	0.32	0.65	0.35

Table 6 shows the results for the Author Ranking task produced with the TREC_EVAL tool. In the table, systems are ordered according to the average *MAP* between the automotive and banking domains. Unfortunately, some participants returned their results in the gold standard format (binary classification as influencers or non influencers) instead of using the prescribed ranking format. We did not discard those submissions and turned their results into the official format by locating profiles marked as influencers at the top, otherwise respecting the original list order.

The followers baseline simply ranks the authors by descending number of followers. It is clearly outperformed by most runs, indicating that additional signals provide useful information. The exception is the miscellaneous domain, where probably additional requirements over the number of followers, such as expertise in a given area, do not clearly apply.

On the other hand, runs from three participants exceeded 0.5 *MAP*, using very different approaches. Therefore, current results do not clearly point to one particular technique.

Figure 7 shows the correlation between the *MAP* values achieved by the systems in the automotive vs. banking domains. There seems to be little correspondence between results in both domains, suggesting that the performance of systems is highly biased by the domain. For future work, it is probably necessary to consider multiple domains to extract robust conclusions.

Figures 8, 9 and 10 illustrate the precision recall curves in the automotive, banking and miscellaneous data sets respectively. We tried to group systems in three levels (black, grey and discontinuous lines) according to their performance. The baseline approach based on followers is represented by the thick dashed line.

Systems improve the followers based baseline in both the automotive and banking domains in all recall levels. This suggests that the number of followers is not the most determinant feature even for the most followed authors. However,

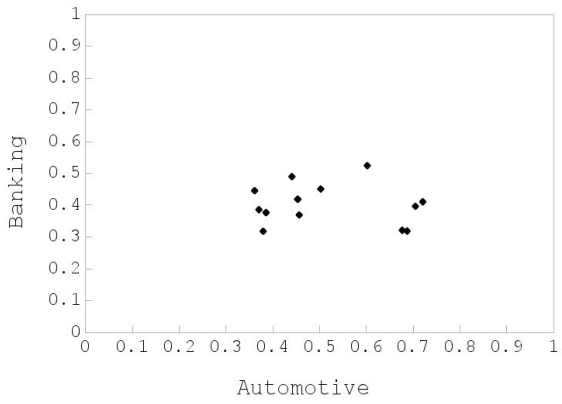

Fig. 7: Correlation of MAP values: Automotive vs. Banking.

this is not the case of the miscellaneous data set, in which the author compilation were biased to high popular writers.

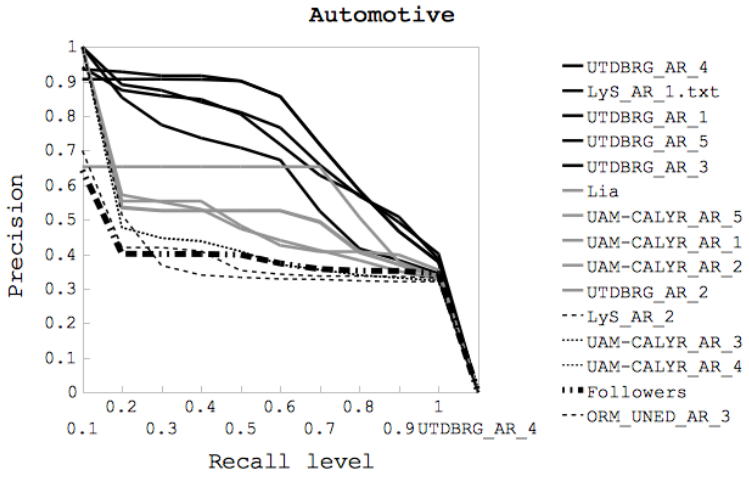

Fig. 8: Precision/Recall curves for Author Ranking in the automotive domain.

7 Conclusions

After two evaluation campaigns on core Online Reputation Management tasks (name ambiguity resolution, reputation polarity, topic and alert detection), Rep-

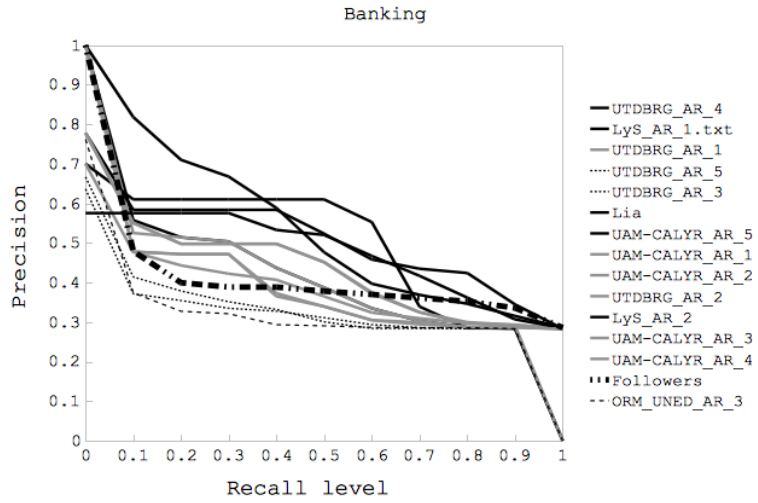

Fig. 9: Precision/Recall curves for Author Ranking in the banking domain.

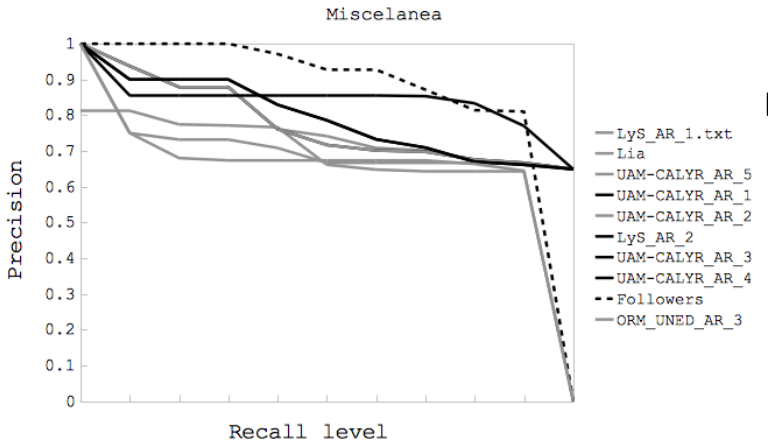

Fig. 10: Precision/Recall curves for Author Ranking in the miscellaneous domain.

Lab 2014 developed an evaluation methodology and test collections for two different reputation management problems: (1) classification of tweets according to the reputation dimensions, and (2) identification and categorisation of opinion makers. Once more, the manual annotations were provided by reputation experts from Llorente & Cuenca (48,000 tweets and 7,000 author profiles annotated).

Being the first shared evaluation on these tasks, participants explored a wide range of approaches in each of them. The classification of tweets according to their reputation dimensions seems to be feasible, although it is not yet clear which are the best signals and techniques to optimally solve it. Author categorisation, on the other hand, proved to be challenging in this initial approximation.

Current results represent simply a first attempt to understand and solve the tasks. Nevertheless, we expect that the data set we are releasing will allow for further experimentation and for a substantial improvement of the state of the art in the near future, as has been the case with the RepLab 2012 and RepLab 2013 data sets.

Acknowledgements. This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 288024 (LiMoSINe) and nr 312827 (VOX-Pol), ESF grant ELIAS, the Spanish Ministry of Education (FPU grant AP2009-0507), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), the Regional Government of Madrid under MA2VICMR (S2009/TIC-1542), Google Award (Axiometrics), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

References

1. Amigó, E., Carrillo-de-Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating Online Reputation Management Systems. In: CLEF 2013 Working Notes (Sep 2013)
2. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M.: Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In: CLEF 2012 Labs and Workshop Notebook Papers (2012)