Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems

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

[1] UNED NLP & IR Group
Juan del Rosal, 16. 28040 Madrid, Spain, http://nlp.uned.es
[2] Llorente & Cuenca
Lagasca, 88. 28001 Madrid, Spain, http://llorenteycuenca.com
[3] Yahoo! Research
Diagonal 177, 08018 Barcelona, Spain, http://research.yahoo.com/
[4] ISLA, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, http://isla.science.uva.nl

Abstract. We summarize the goals, organization, and results of the second RepLab competitive evaluation campaign for Online Reputation Management systems (RepLab 2013). RepLab 2013 focuses on the process of monitoring the reputation of companies and individuals, and asks participating systems to annotate different types of information on tweets containing the names of several companies. First, tweets have to be classified as related or unrelated to the entity; relevant tweets have to be classified according to their polarity for reputation (Does the content of the tweet have positive or negative implications for the reputation of the entity?), clustered in coherent topics, and clusters have to be ranked according to their priority (potential reputation problems had to come first). The gold standard consists of more than 140,000 tweets annotated by a group of trained annotators supervised and monitored by reputation experts.

Keywords: RepLab, Reputation Management, Evaluation Methodologies and Metrics, Test Collections, Text Clustering, Sentiment Analysis

1 Introduction

In a world of online networked information, where its control has moved to users and consumers, every move of a company and every act of a public figure are subject, at all times, to the scrutiny of a powerful global audience. While traditional reputation analysis is mostly manual, online media one allow to process, understand and aggregate large streams of facts and opinions about a company or individual. In this context, natural language processing plays a key, enabling role and we are witnessing an unprecedented demand for text mining software for ORM. Although opinion mining has made significant advances in recent years,

most of the work has focused on products. However, mining and interpreting opinions about companies and individuals is, in general, a much harder and less understood problem, since unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modeling of these entities.

RepLab is an initiative promoted by the EU project LiMoSINe[5] which aims at enabling research on reputation management as a "living lab": a series of evaluation campaigns in which task design and evaluation are jointly carried out by researchers and the target user communities (reputation management experts). Like its first edition in 2012 [2], RepLab 2013 has been organized as a CLEF lab, and the results of the exercise are discussed at CLEF 2013 in Valencia, Spain, on 23–26th September.

RepLab 2013 has been focused on the task of monitoring the reputation of entities (companies, organizations, celebrities, etc.) on Twitter. The monitoring task for analysts consists of searching the stream of tweets for potential mentions to the entity, filtering those that do refer to the entity, detecting topics (i.e., clustering tweets by subject) and ranking them based on the degree to which they are potential reputation alerts (i.e., issues that may have a substantial impact on the reputation of the entity, and must be handled by reputation management experts).

2 Tasks

2.1 Task Definition

Following the outline given above, the RepLab 2013 task is defined as (multilingual) topic detection combined with priority ranking of the topics, as input for reputation monitoring experts. The detection of polarity for reputation (does the tweet have negative/positive implications for the reputation of the entity?) is an essential step to assign priority, and is evaluated as a standalone subtask.

Participants were welcome to present systems that attempt the full monitoring task (filtering + topic detection + topic ranking) or modules that contribute only partially to solve the problem. Subtasks that are explicitly considered in RepLab 2013 are:

- *Filtering.* Systems are asked to determine which tweets are related to the entity and which are not. For instance, distinguishing between tweets that contain the word "Stanford" referring to the University of Stanford and filtering out tweets about Stanford as a place. Manual annotations are provided with two possible values: related/unrelated.
- *Polarity for reputation classification.* The goal is to decide if the tweet content has positive or negative implications for the company's reputation. Manual annotations are: positive/negative/neutral.

[5] http://www.limosine-project.eu

– *Topic detection.* Systems are asked to cluster related tweets about the entity by topic with the objective of grouping together tweets referring to the same subject/event/conversation.
– *Priority assignment.* The full task involves detecting the relative priority of topics. So as to be able to evaluate priority independently from the clustering task, we will evaluate the subtask of predicting the priority of the cluster a tweet belongs to.

A substantial difference between RepLab 2013 and its first edition in 2012 is that, in 2013, the training and test entities are the same, and therefore conventional machine learning techniques are readily applicable. RepLab 2013 models a scenario where reputation experts are constantly tracking and annotating information about a client (entity), and therefore it is likely to have manual annotations for data related to the entity of interest. RepLab 2012, on the other hand, modeled the scenario of a web application that can be used by anyone, at any time, using any entity name as keyword. In that case, training material was referred to entities other than those in the training set.

In RepLab 2013 it was possible to present systems that address only filtering, only polarity identification, only topic detection or only priority assignment. Another difference with 2012 is that in its second edition, the RepLab organization provided baseline components for all of the four subtasks. This way any participant was able to participate in the full task regardless of his particular contribution or expertise.

Some relevant details on the polarity for reputation and topic detection tasks follow. *Polarity for reputation* is substantially different from standard sentiment analysis. First, when analyzing polarity for reputation, both facts and opinions have to be considered. For instance, "Barclays plans additional job cuts in the next two years" is a fact with negative implications for reputation. Therefore, systems will not be explicitly asked to classify tweets as factual vs. opinionated: the goal is to find polarity for reputation, that is, what implications a piece of information might have on the reputation of a given entity, regardless of whether the content is opinionated or not. Second, negative sentiments do not always imply negative polarity for reputation and vice versa. For instance, "R.I.P. Michael Jackson. We'll miss you" has a negative associated sentiment (sadness, deep sorrow), but a positive implication for the reputation of Michael Jackson. And the other way around, a tweet such as "I LIKE IT..... NEXT...MITT ROMNEY...Man sentenced for hiding millions in Swiss bank account," has a positive sentiment (joy about a sentence) but has a negative implication for the reputation of Mitt Romney.

As for the *topic detection + topic ranking* process, a three-valued classification was applied to assess the priority of each entity-related topic: alert (the topic deserves immediate attention of reputation managers), mildly relevant (the topic contributes to the reputation of the entity but does not require immediate attention) and unimportant (the topic can be neglected from a reputation management perspective). Some of the factors that play a role in the priority assessments are:

– *Polarity.* Topics with polarity (and, in particular, with negative polarity, where action is needed) usually have more priority.
– *Centrality.* A high priority topic is very likely to have the company as the main focus of the content.
– *User's authority.* A topic promoted by an influential (for example, in terms of the number of followers or the expertise) user has better chances of receiving high priority.

2.2 Baselines

The baseline approach consists of tagging tweets (in the test set) with the same tags of the closear tweet in the (entity) training set according to the Jaccard word distance. The baseline is, therefore, a simple version of memory-based learning. We have selected this approach for several reasons: (i) it is easy to understand; (ii) it can be applied to every subtask in RepLab 2013; (iii) it keeps the coherence between tasks: if a tweet is annotated as non-related, it will not receive any priority or topic tag; (iv) it exploits the training data set per entity.

2.3 Evaluation Measures

All subtasks consist of tagging single tweets according to their relatedness, priority, polarity or topic. However, each one corresponds to a particular artificial intelligence problem: binary classification (relatedness), three-level classification (polarity and priority), clustering (topic detection), and their concatenation (full task). A common feature for all tasks is that the classes, levels or clusters can be unbalanced. This entails challenges for the definition of our evaluation methodology. First, in classification tasks, a non informative system (i.e., all tweets to the same class) can achieve high scores without providing useful information. Second, in three-level classification tasks, a system could sort tweets correctly without a perfect correspondence between predicted and true tags. Third, an unbalanced cluster distribution across entities produces an important trade-off between precision/recall oriented evaluation metrics (precision or cluster entropy versus recall or class entropy) and that makes the measure combination function crucial for system ranking.

In evaluation, there is a hidden trade-off between interpretability and strictness. For instance, Accuracy is easy to interpret: it simply reports how frequently the system makes the correct decision. However, it is also easy to be cheated under unbalanced test sets. For instance, returning all tweets in the same class, cluster or level, may have high accuracy if the set is unbalanced. Other measures based on information theory are more strict when penalizing non informative outputs, but at the cost of interpretability. In this evaluation campaign we employ Accuracy as a high interpretable measure, and the combination of Reliability and Sensitivity (R&S) as a strict and theory grounded measure [4].

Basically, R&S assumes that any organization task consists of a bag of relationships between documents. In our tasks, two documents are related if they have different priority, polarity or relatedness level, or when they appear in the

same cluster. In brief, R&S computes the precision and recall of relationships produced by the systems with respect to the goldstandard. In order to avoid the quadratic efect of document pairwise, R&S is computed for each document relationships and averaged in a second step. Reliability and Sensitivity are computed as, being $\mathcal{I}$ the set of tweets considered in the evaluation:

$$R(system) = Avg_{i \in \mathcal{I}} R(i) \quad S(system) = Avg_{i \in \mathcal{I}} S(i)$$

$$R(i) = P_{j \in \mathcal{I}}(rel_{gold}(i,j) = rel_{sys}(i,j) | rel_{sys}(i,j))$$
$$S(i) = P_{j \in \mathcal{I}}(rel_{gold}(i,j) = rel_{sys}(i,j) | rel_{gold}(i,j)),$$

where $rel_{gold}(i,j)$ represents that i has a higher or lower polarity, priority or relatedness than i, or that i and j belong to the same cluster. $Rel_{sys}(i,j)$ is analogous but applied to the system output.

R&S has three main strengths. First, it can be applied to ranking, filtering, organization by levels and grouping tasks. This matches all the RepLab 2013 tasks. In addition, it gives the possibility to evaluate the full task as a whole. Second, it covers simultaneously the desirable formal properties satisfied by other measures in each particular task [4]. Third, according to experimental results that we corroborate with RepLab 2013 data, R&S is strict with respect to other measures: a high score according to R&S ensures a high score according to any traditional measure. In other words, a low score according to one particular traditional measure produces a low R&S score, even when the system is rewarded by other measures.

R and S are combined with the F measure, i.e., a weighted harmonic mean of R and S. This combining function is grounded in measure theory and satisfies a set of desirable constraints. One of the most useful is that a low score according to one of the two measures strongly penalizes the combined score. However, specially in clustering tasks, the F measure is seriously affected by the relative weight of partial measures (the α parameter). In order to solve this, we complement the evaluation results with the Unanimous Improvement Ratio, which has been proved to be the only weighting independent combining criterion [3]. UIR is computed over the test cases (entities in RepLab) in which all measures corroborates a difference between runs. Let S_1 and S_2 be two runs and $N_{>\forall}(S_1, S_2)$ the amount of test cases for which S_1 improves S_2 for all measures, then:

$$UIR(S_1, S_2) = \frac{N_{>\forall}(S_1, S_2) - N_{>\forall}(S_2, S_1)}{Amount\ of\ cases}$$

3 Dataset

RepLab 2013 uses Twitter data in English and Spanish. The balance between both languages depends on the availability of data for each of the entities included in the dataset. The collection comprises tweets about 61 entities from four domains: automotive, banking, universities and music. The domain selection was done to offer a variety of scenarios for reputation studies. To this aim

we included entities whose reputation largely relies on their products (automotive), entities for which transparency and ethical side of their activity are the most decisive reputation factors (banking), entities for which the reputation of which depends on a very broad and intangible set of products (universities) and, finally, entities where the reputation is based almost equally on their products and personal qualities (music bands and artists). Table 1 summarizes the description of the corpus, as well as the number of tweets for both training and test sets, and the distribution by language.

Crawling was performed from 1 June, 2012 until 31 Dec, 2012 using each entity's canonical name as query. For each entity, at least 2,200 tweets were collected: the first 700 were reserved for the training set and the last 1,500 for the test collection. This distribution was set in this way to obtain a temporal separation (ideally of several months) between the training and test data. The corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets situated between the training (earlier tweets) and test material (the latest tweets) in the timeline.

Table 1. RepLab 2013 dataset.

	All	**Autom.**	**Banking**	**Univers.**	**Music/Artist**
Entities	61	20	11	10	20
Training No. Tweets	45,679	15,123	7,774	6,960	15,822
Test No. Tweets	96,848	31,785	16,621	14,944	33,498
Total No. Tweets	142,527	46,908	24,395	21,904	49,320
No. Tweets EN	113,544	38,614	16,305	20,342	38,283
No. Tweets ES	28,983	8,294	8,090	1,562	11,037

These data sets were manually labelled by thirteen annotators who were trained, guided and constantly monitored by experts in ORM. Each tweet is annotated as follows:

- RELATED/UNRELATED: the tweet is/is not about the entity.
- POSITIVE/NEUTRAL/NEGATIVE: the information contained in the tweet has positive, neutral or negative implications for the entity's reputation.
- Identifier of the topic cluster the tweet has been assigned to.
- ALERT/MILDLY IMPORTANT/UNIMPORTANT: the priority of the topic cluster the tweet belongs to.

Table 2 shows statistics about the filtering subtask. The collection contains 110,352 tweets related with the entities, out of which 34,882 are in the training set and 75,470 are in the test set. The 32,175 unrelated tweets of the dataset are distributed as follows: 10,797 tweets in the training set and 21,378 in the test set. The table also shows the distributions by domain.

Table 3 shows the distribution of polarity classes in the RepLab 2013 dataset. The RepLab 2013 dataset contains 63,442 tweets classified as positive by the annotator, 30,493 classified as neutral and 16,415 classified as negative. The

Table 2. RepLab 2013 dataset for the Filtering Task.

	All	Autom.	Banking	Univers.	Music/Artist
Training No. Related	34,882	11,356	5,753	3,412	14,361
Training No. Unrelated	10,797	3,767	2,021	3,548	1,461
Test No. Related	75,470	24,415	12,053	7,715	31,287
Test No. Unrelated	21,378	7,370	4,568	7,229	2,211
Total No. Related	110,352	35,771	17,806	11,127	45,648
Total No. Unrelated	32,175	11,137	6,589	10,777	3,672

Table 3. RepLab 2013 dataset for the Polarity Task.

	All	Autom.	Banking	Univers.	Music/Artist
Training No. Positive	19,718	5,749	2,195	2,286	9,488
Training No. Neutral	9,753	4,616	767	894	3,476
Training No. Negative	5,409	991	2,791	232	1,395
Test No. Positive	43,724	24,415	12,053	7,715	31,287
Test No. Neutral	20,740	9,512	1,407	2,443	7,378
Test No. Negative	11,006	2,101	4,994	820	3,091
Total No. Positive	63,442	12,802	5,652	4,452	20,818
Total No. Neutral	30,493	14,128	2,174	3,337	10,854
Total No. Negative	16,415	3,092	7,785	1,052	4,486

distribution in the training set is 19,718 tweets classified as positive, 9,753 as neutral and 5,409 as negative, while the test set contains 63,442 positive tweets, 30,493 neutral tweets and 16,415 negatives.

Table 4 displays the number of topics per set as well as the average number of tweets per topic, which is 17.77 for the whole collection but goes from 14.40 in the training set to 21.14 in the test set. The training set contains 3,813 different topics, the test set 5,757 different topics, for a total of 9,570 different topics in the RepLab 2013 dataset.

Finally, Table 5 summarizes the distributions of tweets in priority classes. The less representative class is *alert*, with 4,780 tweets classified as a possible reputation alert in the whole corpus. *Mildly_Important* has 56,578 tweets and *Unimportant* receives 48,992 tweets.

In order to determine inter-annotator agreement we perform two different experiments. First, 14 entities (4 automotive, 3 banking, 3 universities, 4 music) have been labeled by two annotators. This subset contains 31,381 tweets that represent 22% of the RepLab 2013 dataset covering all domains. Second, three annotators labeled 3 entities of the automotive domain. Table 6 shows the results of the first experiment of agreement using percentage of agreement and Kappa metrics (both Cohen and Fleiss) for filtering, polarity and priority detection tasks, and F measure of Reliability and Sensitivity for topic detection

Table 4. RepLab 2013 dataset for the Topic Detection Task.

	All	Autom.	Banking	Univers.	Music/Artist
Training No. Topics	3,813	1,389	831	503	1,090
Training Avg. No. Tweets/Topic	14.40	12.36	11.35	17.57	16.53
Test No. Topics	5,757	1,959	1,121	1,035	1,642
Test Avg. No. Tweets/Topic	21.14	18.42	18.95	21.78	24.74
Total No. Topics	9,570	3,348	1,952	1,538	2,732
Total Avg. No. Tweets/Topic	17.77	15.39	15.15	19.67	20.64

Table 5. RepLab 2013 dataset for the Priority Detection Task.

	All	Autom.	Banking	Univers.	Music/Artist
Training No. Alert	1,540	226	841	88	385
Training No. Mildly_Important	17,961	5,388	2,509	1,949	8,115
Training No. Unimportant	15,379	5,742	2,403	1,375	5,859
Test No. Alert	3,240	483	2,195	102	460
Test No. Mildly_Important	38,617	10,967	5,429	4,441	17,780
Test No. Unimportant	33,613	12,965	4,429	3,172	13,047
Total No. Alert	4,780	709	3,036	190	845
Total No. Mildly_Important	56,578	16,355	7,938	6,390	25,895
Total No. Unimportant	48,992	18,707	6,832	4,547	18,906

task. As can be observed, the percentage of agreement for the filtering subtask is near 100%, while taking into account the class distribution with the kappa metrics the inter agreement between annotator decreases. The values obtained for reputation polarity in terms of percentage of agreement are quite similar to other studies over sentiment analysis task. As in the filtering subtask, the value obtained with kappa in the reputation polarity subtask decrease with respect of percentage of agreement. For the topic detection subtask, we do not compute inter agreement between annotators for the whole RepLab 2013 dataset. This is due to the organization of the labeling process. The annotators consider the training and test set as two different sets, so cannot group tweets of both sets. The agreement for the topic detection task is higher than expected, taking into account the complexity of the subtask.

As expected, the results obtained in the experiment with three annotators are lower. As can be seen in Table 7, the inter agreement for the filtering task is quite similar to that obtained in the experiment with two annotators, while the results for the reputational polarity decrease considerably in all metrics. Concerning the topic detection subtask, the table shows the average of F measure over all combinations between annotators. Notably, this task is the one with a lower decrease with respect to the experiment with two annotators, even if this subtask depends on the organization behavior of the annotators. Similarly to the previous

Table 6. RepLab 2013 agreement: analysis of 14 entities labeled by two annotators.

	% Agreement	Cohen κ	Fleiss κ	$\mathbf{F_1}(\mathbf{R}, \mathbf{S})$
Training Filtering	94.80	70.01	68.84	–
Training Polarity	68.27	41.04	38.93	–
Training Topic Detection	–	–	–	49.59
Training Priority Detection	58.41	23.96	15.96	–
Test Filtering	96.46	68.00	67.86	–
Test Polarity	68.81	42.26	39.92	–
Test Topic Detection	–	–	–	48.07
Test Priority Detection	60.88	29.29	20.91	–
Total Filtering	95.94	66.69	66.35	–
Total Polarity	68.59	41.93	39.79	–
Total Topic Detection	–	–	–	–
Total Priority Detection	60.07	28.04	20.24	–

experiments of two annotators, as the training and test are considered as two sets by the annotator, the topic detection inter agreement for the whole RepLab 2013 dataset is not computed. Finally, the values obtained for the priority task for three annotators decrease more than for topic detection comparing with the previous experiment, but are still similar.

4 Participation

44 groups signed up for RepLab 2013, although only 15 of them submitted runs to the official evaluation.[6] This year the task was defined in such a way that using the baselines provided by the organizers, every group, besides participating in a concrete subtask, could submit its system to the full task. Nevertheless, only 4 systems explicitly used this possibility.[7] Overall, 5 groups participated in the topic detection subtask, 11 in the reputation polarity classification subtask, 14 in the filtering subtask and 4 in the priority assignment subtask. Below we list the participants and briefly describe the approaches used by each group. Table 8 shows the acronyms and affiliations of the research groups that took part in RepLab 2013.

CIRGDISCO participated in the filtering subtask. They exploited "context phrases" found in tweets and Wikipedia disambiguated articles for a particular entity in an SVM classifier that utilizes features extracted from the Wikipedia graph structure, i.e. incoming and outcoming links from and to Wikipedia articles. They used, in addition, features derived from term-specificity and term-collocation features derived from the Wikipedia article of the analysed entity.

[6] One additional group sent their results two days after the deadline, and their runs are reported here as "unofficial." An asterisk in tables indicates an unofficial result.
[7] Daedalus, GAVKTH, SZTE_NLP, and UNED_ORM.

Table 7. RepLab 2013 agreement analysis of 3 entities labeled by three annotators.

	% Agreement	Fleiss κ	Average($\mathbf{F_1}(\mathbf{R}, \mathbf{S})$)
Training Filtering	92.46	56.63	–
Training Polarity	48.81	36.75	-p
Training Topic Detection	–	–	48.11
Training Priority Detection	46.89	27.23	–
Test Filtering	91.54	59.60	–
Test Polarity	51.98	39.11	–
Test Topic Detection	–	–	51.33
Test Priority Detection	53.93	36.04	–
Total Filtering	91.83	59.59	–
Total Polarity	51.03	38.59	–
Total Topic Detection	–	–	–
Total Priority Detection	51.72	33.38	–

Daedalus submitted specific runs for the filtering and polarity subtasks, apart from the full task. Their approach to the filtering subtask is based on the use of linguistic processing modules to detect and disambiguate named entities at several levels. The 4 submitted runs are defined by a combination of morphosyntactic-based vs. semantic disambiguation and a case sensitive/insensitive processing of the tweets. On the other hand, the polarity classification uses a lexicon-based approach to sentiment analysis, improved with a full syntactic analysis and detection of negation and polarity modifiers, which also provides the polarity at entity level.

DIUE applied a supervised Machine Learning (ML) approach for the polarity classification subtask. The Python NLTK has been used for preprocessing, including file parsing, text analysis and feature extraction. The best run combines bag-of-words with a set of 18 features related to presence of the polarized term, negation before the polarized expression, as well as entity reference based on sentiment lexicons and shallow text analysis.

GAVKTH used its commercially available system for the filtering and reputation polarity subtasks. The system, designed for large scale analysis of streaming text and measuring the public attitude towards targets of interest, has been used with no adjustment for the specific subtasks. The basic approach relies on distributional semantics represented in a semantic space by means of a patented implementation of the Random Indexing processing framework.

LIA applied a large variety of ML methods mainly based on exploiting tweet contents to filtering, polarity classification, topic detection, and priority assignment. In several experiments some metadata were added and a fewer number of runs incorporated external information by using provided links to Wikipedia and entities' official web sites.

Table 8. List of participants: acronyms and affiliation.

Acronym	Affiliation	Country
CIRGDISCO	National University of Ireland, Galway	Ireland
Daedalus	Daedalus, S.A.	Spain
DIUE	Universidade de Évora	Portugal
GAVKTH	Gavagai	Sweden
IE	National University of Singapore	Singapore
LIA	University of Avignon	France
NLP&IR_GROUP_UNED	UNED	Spain
POPSTAR	Universidade Porto	Portugal
REINA	Reina Research Group, University of Salamanca	Spain
SZTE_NLP	University of Szeged	Hungary
UAMCLYR	Universidad Autónoma Metropolitana Cuajimalpa	Mexico
UNED_ORM	UNED	Spain
UNED-READERS*	UNED	Spain
UNEDTECNALIA	Tecnalia Research And Innovation, UNED	Spain
UVA_UNED	University of Amsterdam, UNED	The Netherlands, Spain
volvam	Volvam Analytics and University of Alicante	Ireland, Spain

NLP&IR_GROUP_UNED focused on addressing filtering and reputation polarity classification using an IR method. Viewing these two subtasks as the same problem, i.e. finding the most relevant class to annotate a given tweet, a classical IR approach was applied, using the tweet content as query against an index with the models of the classes used to annotate tweets. The classes were modelled by means of the Kullback Leibler Divergence (KLD), in order to extract their most representative terminology. For topic detection, instead of a clustering based technique, this group resorted to Formal Concept Analysis (FCA) to represent the contents in a lattice structure. Topics were extracted from the lattice using a FCA concept, *stability*.

popstar participated in the filtering and reputation polarity classification subtasks. For filtering, these researchers explored different learning algorithms considering a variety of features describing the relationship between an entity and a tweet, such as text, keyword similarity scores between entities metadata and tweets, the Freebase entity graph and Wikipedia.

REINA used classical systems for the similarity matrix and community detection techniques for topic detection. No distinction was made between languages of the tweets, doing a uniform lexical analysis of all tweets, applying a simple s-stemmer and removing the words with less than 4 characters. Additionally, the

discarded emoticons were considered as well as hashtags and some entities terms. The urls shared by two tweets were deemed as another important feature of the tweets, assuming this is indicative of topic similarity.

SZTE_NLP presented a system to tackle the filtering and reputation polarity classification subtasks using supervised ML techniques. Several Twitter specific text preprocessing and features engineering methods were applied. Besides supervised methods, they also experimented with incorporating clustering information.

UAMCLYR adopted Distributional Term Representations (DTR) to tackle the filtering and reputation polarity classification subtasks. Terms were represented by means of contextual information given by the term co-occurrence statistics. For topic detection and priority assignment, these researchers explored clustering and classification methods as well as term selection techniques working with two settings: single tweets and tweets extended with derived posts.

UNED_ORM submitted runs to the full task and all the subtasks testing several approaches. First, Instance-based learning using Heterogeneity Based Ranking to combine seven different similarity measures was applied to all the subtasks. The filtering subtask was also tackled by automatically discovering positive and negative filter keywords, i.e. terms present in a tweet that reliably predict the relatedness or non-relatedness of the message to the analysed entity. The topic detection subtask was attempted with three approaches: agglomerative clustering over Wikified tweets, co-occurrence term clustering and an LDA-based model that uses temporal information. Finally, the polarity subtask was tackled by generating domain specific semantic graphs in order to automatically expand the general purpose lexicon SentiSense.

*UNED-READERS** applied an unsupervised knowledge-based approach to filter relevant tweets for a given entity. The method exploits a new way of contextualizing entity names from relatively large collections of texts using probabilistic signature models, i.e., discrete probability distributions of words lexically related to the knowledge or topic underlying the set of entities in background text collections. The contextualization is intended to recover relevant information about the entity, particularly, lexically related words, from background knowledge.

UNEDTECNALIA submitted a filtering algorithm that takes advantage of the Web of Data in order to create a context for every entity. The semantic context of the analysed entities is generated by querying different data sources (modelled by a set of ontologies) provided by the Linked Open Data Cloud. The extracted context is then compared to the terms contained in the tweet.

UVA_UNED, a collaborative participation of UvA and UNED, focused on applying an active learning approach to the filtering subtask. It consisted of exploiting features based on the detected semantics in the tweet (using Entity Linking with

Wikipedia), as well as tweet-inherent features such as hashtags and usernames. The tweets manually inspected during the active learning process were at most 1% of the test data.

volvam participated in polarity classification and applied one supervised and two unsupervised approaches, combining ML and lexicon-based techniques with an emotional concept model. These methods had been properly adapted to English and Spanish depending on the resources available for each language. The first, unsupervised, approach made use of fuzzy lexicons in order to catch informal variants that are common in Twitter texts. The supervised method extended the first approach with ML techniques and an emotion concept model, while the last one also employed ML but incorporating the bag-of-concepts approach using SenticNet common-sense affective knowledge.

5 Evaluation Results

5.1 Polarity

Polarity has been evaluated according to Accuracy and R&S. Only entity-related tweets in the test set have been assessed. In order to keep evaluation independent from the filtering task, we do not penalize polarity annotations made on non-related tweets. That is, only related tweets are considered in the Accuracy and R&S computation. The related tweets without system response are penalized. The system results, sorted by accuracy are shown in Table 9. The table includes only the best system, according to R&S or Accuracy, for each team. The second column contains the ratio of tweets for which the output gives results.

The majority class is the dataset is "POSITIVE". The baseline approach appears in the middle of the ranking. SZTE and POPSTAR teams improve, in general, most systems according to both accuracy and R&S. Note that some systems achieve a low accuracy (below the baseline) but with competitive R&S. As R&S only looks at the relative ordering between tweets (rather than the

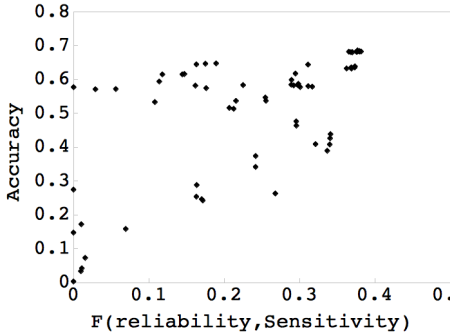

Fig. 1. Polarity: Accuracy versus R&S

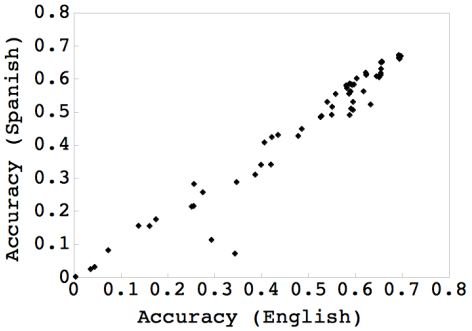

Fig. 2. Polarity: Accuracy EN vs ES.

Table 9. Accuracy, ratio of processed tweets, correlation at entity level, Reliability and Sensitivity for polarity task.

Run	Acc.	Processed Tweet ratio	Corr. Ent. Level	R	S	F
SZTE NLP 8	0.69	1.00	0.88	0.48	0.34	0.38
LIA 7	0.65	1.00	0.82	0.50	0.15	0.19
POPSTAR 5	0.64	0.98	0.89	0.43	0.34	0.37
UAMCLYR 2	0.62	1.00	0.82	0.38	0.27	0.29
UNED ORM 2	0.62	1.00	0.70	0.36	0.10	0.15
LIA 3	0.60	1.00	0.64	0.37	0.27	0.29
UNED ORM 1	0.59	1.00	0.87	0.32	0.29	0.30
Baseline	0.58	1.00	0.87	0.32	0.29	0.30
NLP IR UNED 1	0.58	1.00	0.79	0.33	0.31	0.32
UAMCLYR 05	0.58	1.00	0.78	0.33	0.29	0.30
IE 6	0.58	1.00	0.22	0.94	0.00	0.00
ALL POSITIVE	0.58	1.00	0.00	1.00	0.00	0.00
DIUE 1	0.55	1.00	0.21	0.33	0.22	0.25
VOLVAM 3	0.54	1.00	0.36	0.32	0.23	0.26
IE 5	0.52	1.00	0.18	0.29	0.22	0.21
Daedalus 3	0.44	1.00	0.52	0.31	0.40	0.34
volvam 2	0.41	1.00	0.38	0.31	0.39	0.34
GAVKTH 6	0.37	0.98	0.49	0.30	0.21	0.24
ALL NEUTRAL	0.27	1.00	0.00	1.00	0.00	0.00
GAVKTH 2	0.26	0.82	0.21	0.37	0.21	0.27
ALL NEGATIVE	0.15	1.00	0.00	1.00	0.00	0.00

actual tags), a possible reason is that, while many tags are not correct, the ordinal polarity relationship between them is correct. Figure 1 illustrates the correspondence between Accuracy and R&S. Note that a high R&S tends to be associated with a high accuracy.

Another important aspect of polarity detection for ORM is the ability to predict the average polarity of an entity with respect to other entities. To evaluate this ability, we have computed the Pearson correlation between the average estimated and real polarity levels across entities.[8] An interesting result is that some approaches are able to estimate the average polarity reputation for an entity with a 0.9 correlation with the ground truth.

Finally, Figure 2 shows the correlation between Accuracy scores over English versus Spanish tweets. In most cases there is a high correspondence. The accuracy for Spanish seems to be upper bounded by the accuracy over English tweets.

5.2 Filtering

In this task, tweets must be classified as related or unrelated to the entity of interest. R&S in filtering tasks (two levels) correspond with the products of

[8] For the correlation computation, we assign 0, 1 and 2 for each class respectively.

Table 10. Results for the Filtering Subtask.

Run	R	S	F	ACC
POPSTAR 2	0.73	0.45	0.49	0.91
SZTE NLP 7	0.60	0.44	0.44	0.93
LIA 1	0.66	0.36	0.38	0.87
UAMCLYR 04	0.56	0.4	0.38	0.91
LIA 6	0.62	0.33	0.34	0.88
UNED ORM 2	0.43	0.38	0.34	0.86
BASELINE	0.49	0.32	0.33	0.87
Daedalus 1	0.35	0.45	0.32	0.85
UNED-READERS 2	0.38	0.33	0.28	0.55
CIRG IRDISCO 4	0.34	0.33	0.27	0.84
IE 4	0.45	0.23	0.26	0.44
CIRG IRDISCO 1	0.5	0.24	0.25	0.87
Uva UNED 6	0.68	0.22	0.21	0.82
UNEDTECNALIA 1	0.28	0.29	0.18	0.46
NLP IR GROUP UNED 9	0.29	0.22	0.17	0.78
IE 2	0.46	0.16	0.17	0.53
CIRG IRDISCO 2	0.82	0.16	0.17	0.86
NLP IR GROUP UNED 8	0.31	0.19	0.16	0.79
GAVKTH 1	0.81	0.07	0.05	0.76
ALL RELATED	0	0	0	0.77
ALL UNRELATED	0	0	0	0.23

precision in both classes and the product or recall scores respectively. Table 10 shows the Accuracy and R&S results for the filtering task. Again, we have included only the best run according to Accuracy or R&S for each team. Most tweets are related (77%). As in the polarity tasks, the baseline approach appears in the middle of the ranking for both R&S and Accuracy. Figure 3 shows the correspondence between Accuracy and R&S. As in the polarity task, a high R&S score ensures a high Accuracy score. As in the polarity task, there are no important differences in system scores when considering the Spanish vs. English tweets. There is a 0.94 Pearson Correlation) between scores over both kind of tweets. In general, the top scores are much higher than in RepLab 2012; this is explained by the fact that in this new dataset the training and test entities are the same.

5.3 Priority

The Priority task consists of classifying tweets into three levels. Reliability represents the ratio of correct priority relationships per tweet, while Sensitivity represents the ratio of captured relationships per tweet. In this case, as well as in polarity, only the related tweets (according to assessors) are considered in the evaluation process. Table 11 shows the results. Only the best Accuracy and R&S score per team is included. Not all systems have annotated all tweets (see the

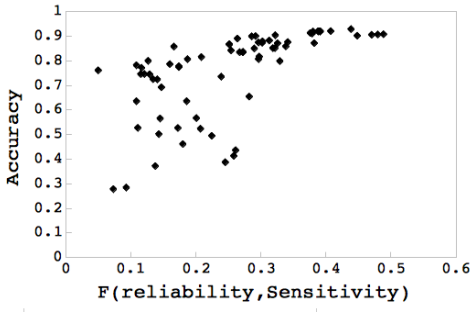

Fig. 3. Accuracy versus R&S in the Filtering Task.

Table 11. Accuracy, Reliability and Sensitivity Results for the Priority Subtask.

Run	R	S	F	ACC	Amount of processed tweets
LIA 5	0.39	0.32	0.34	0.63	0.97
UNED ORM 1	0.31	0.31	0.3	0.6	1
BASELINE	0.3	0.3	0.3	0.6	1
GAVKTH 2	0.36	0.19	0.25	0.37	0.82
UAMCLYR 2	0.24	0.2	0.2	0.46	1
GAVKTH 7	0.37	0.09	0.13	0.41	0.83
UAMCLYR 3	0.58	0.07	0.09	0.57	1
ALL MILDLY IMPORTANT	0	0	0	0.52	1
ALL UNIMPORTANT	0	0	0	0.44	1
ALL ALERT	0	0	0	0.04	1

last column). The best run achieves a high score for both R&S and Accuracy measures. The baseline approach is improved substantially for both measures.

5.4 Topic Detection

Topic detection is a clustering task which has been evaluated according to R&S, which correspond with the popular measures Bcubed precision and Recall [1]. Table 12 displays the results. Only the best F measure is considered for each team. Figure 4 shows that there is an important trade-off between R and S in this task. In these circumstances, the F measure weighted with $\alpha = 0.5$ rewards the runs located in the diagonal axis. But this choice of α is, to some extent, arbitrary. For this reason, we check the evaluation results according to UIR (see previous section). UIR is a complementary measure that indicates to what extent run improvements are sensitive to variations in the measure weighting scheme (i.e. in α). Table 13 shows for all runs, the other runs which are improved by the first with UIR$\geq 0,2$. This implies that there is a difference higher than 0.2

Table 12. Reliability and Sensitivity in the Topic Detection Task.

Run	S	R	F	Ratio proc. tweets
UNED_ORM_2	0.46	0.32	0.33	0.99
REINA_2	0.32	0.43	0.29	0.79
LIA_3	0.22	0.35	0.25	1.00
UAMCLYR_7	0.35	0.50	0.24	0.97
REINA_1	0.16	0.52	0.23	0.99
Baseline	0.15	0.22	0.17	1.00
NLP_IR_UNED_1	0.67	0.11	0.17	0.53
ALLINONE	0.07	1.00	0.12	1.00
ALLINALL	1.00	0.04	0.07	1.00

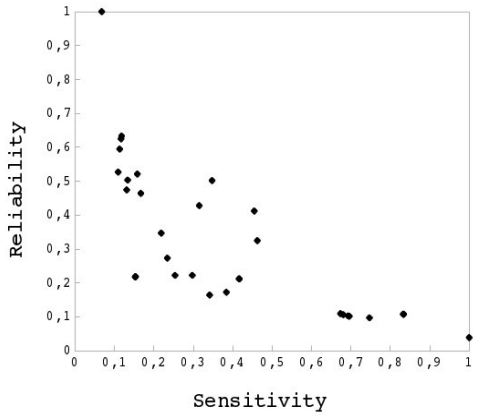

Fig. 4. Reliability vs. Sensitivity in the Topic Detection Task.

between the cases in which the first run improves the other for R and S and vice versa. Interestingly, although UAMCLYR_7 is not the best system in the $F_{\alpha=0.5}$ ranking, it improves robustly a great amount of runs. Some team runs like LIA are not comparable to each other. Probably, they have different grouping thresholds.

5.5 Full Task

The full task joins filtering, priority and topic detection tasks. The use of R&S allows us to apply the same evaluation criterion to all subtasks and therefore, to combine all of them. It is possible to apply R&S directly over the set of relationships (priority, filtering and clustering) but then the most frequent binary relationships dominate the evaluation results (in our case, priority relationships would dominate). We decided to use a weighted harmonic mean (F measure) of

Table 13. UIR analysis for the Topic Detection Task.

RUN	Improves runs UIR ≥ 0.2	Number of improved runs
UAMCLYR_07	UAMCLYR_1,2,3,4,5,6 LIA_2,3,4 REINA_1 BASELINE UNED_ORM_1	12
UNED_ORM_2	LIA_1,2,3,4 UNED_ORM_1,3,4,5,6,7 BASELINE	11
REINA_2	LIA_1,2,3,4 BASELINE UAMCLYR_4 UNED_ORM_1,6,7	9
UAMCLYR_8	LIA_2,4 UAMCLYR_1,2,3,4 BASELINE UNED_ORM_1	8
UNED_ORM_4	BASELINE UNED_ORM_1,6 LIA_1,4	5
UNED_ORM_5	BASELINE UNED_ORM_1,6 LIA_1,4	5
REINA 1	UAMCLYR_3,4 BASELINE UNED_ORM_1	4
UNED_ORM_3	BASELINE UNED_ORM_1 LIA_1 UNED_ORM_6	4
UNED_ORM_7	LIA_2,4 BASELINE UNED_ORM_1	4
UAMCLYR_6	BASELINE UAMCLYR_4 UNED_ORM_1	3
UAMCLYR_3	BASELINE UAMCLYR_4 UNED_ORM_1	3
NLP_IR_UNED_10	NLP_IR_UNED_3,4,5	3
UAMCLYR_5	BASELINE UAMCLYR_04 UNED_ORM_1	3
NLP_IR_UNED_8	NLP_IR_UNED_3,4,5	3
NLP_IR_UNED_9	NLP_IR_UNED_3,4,5	3
LIA_2	BASELINE UNED_ORM_1	2
LIA_3	BASELINE UNED_ORM_1	2
LIA_4	BASELINE UNED_ORM_1	2
UNED_ORM_6	BASELINE UNED_ORM_1	2
UAMCLYR_01	UAMCLYR_02	1
UAMCLYR_04	BASELINE	1
NLP_IR_UNED_6	NLP_IR_UNED_4	1
NLP_IR_UNED_7	NLP_IR_UNED_4	1

the six Reliability and Sensitivity measures corresponding to the three subtasks embedded in the full task. In cases of empty partial outputs, we have completed runs with the baseline approach as specified in the guidelines.

Table 14 shows the team ranking in terms of F. However, this evaluation is highly sensitive to the relative importance of measures in the combining function. For this reason, we have also computed UIR between each pair of runs. Here we consider as an unanimous improvement of system A over system B to those test cases (entities) for which all the six measures are better for A than for B. Results of the UIR analysis are shown in Table 15. The third and fourth columns represent how many entities one run improves or is improved by the other. It only includes those run pairs for which UIR is bigger than 0.2. As the table shows, actually, runs from different teams are not comparable to each other: improvements in F are dependent on the relative weighting scheme. However,

Table 14. Full Task Results.

Run	F measure
UNED_ORM_2	0.19
UNED_ORM_7	0.18
UNED_ORM_4	0.17
UNED_ORM_6	0.17
DAEDALUS_1..8	0.16
UNED_ORM_1	0.16
UNED_ORM_8	0.12
UNED_ORM_3	0.11
UNED_ORM_5	0.11
SZTE_NLP_1..10	0.03

Table 15. UIR Analysis for the Full Task.

Run 1	Run 2	Imp.	Is imp.	UIR
UNED_ORM_2	UNED_ORM_4	24	1	0.38
UNED_ORM_2	UNED_ORM_6	15	0	0.25
UNED_ORM_3	UNED_ORM_5	14	1	0.21
SZTE_7	SZTE_4	44	15	0.47
SZTE_7	SZTE_3	43	15	0.46
SZTE_7	SZTE_6	44	17	0.44
SZTE_7	SZTE_1	42	17	0.41
SZTE_7	SZTE_2	40	15	0.41
SZTE_7	SZTE_5	43	19	0.39
SZTE_7	SZTE_9	40	18	0.36
SZTE_7	SZTE_8	37	17	0.33
SZTE_7	SZTE_10	35	18	0.28
SZTE_10	SZTE_9	37	22	0.25

there are a number of significant improvements (in terms of UIR) between runs from the same teams.

6 Conclusions

Perhaps the main outcome of RepLab 2013 is its dataset, which comprises more than 142,000 tweets in two languages with four types of high-quality manual annotations, covering all essential aspects of the reputation monitoring process. We expect this dataset to become a useful resource for researchers not only in the field of reputation management, but also for researchers in Information Retrieval and Natural Language Processing in general. Just to give an example, the topics (tweet clusters) together with their relative ranking can be directly mapped into a test collection to evaluate search with diversity algorithms over Twitter.

Compared to RepLab 2012, availability of training data for the entities in the test set naturally improves system results and also allows for a more straight-

forward application of machine learning techniques. But the tasks themselves are still far from solved; even with plenty of entity-specific training material the RepLab tasks—polarity, topic detection, and ranking—have proved challenging for state-of-the-art systems.

Acknowledgements

This research was partially supported by the European Community's FP7 Programme under grant agreement nrs 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe), the ESF Research Network Program ELIAS, the Spanish Ministry of Education (FPU grant AP2009-0507 and FPI grant BES-2011-044328), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), and the Regional Government of Madrid under MA2-VICMR (S2009/TIC-1542), the Netherlands Organisation for Scientific Research (NWO) under project nrs 640.004.802, 727.011.005, 612.001.116, HOR-11-10, 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 Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105 and the Yahoo! Faculty Research and Engagement Program.

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12(4), 461–486 (2009)
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)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. Journal of Artificial Intelligence Research 42(1), 689–718 (2011)
4. Amigó, E., Gonzalo, J., Verdejo, F.: A General Evaluation Measure for Document Organization Tasks. In: Proceedings of SIGIR 2013 (jul 2013)