UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Overview of RepLab 2012: Evaluating Online Reputation Management Systems

Amigó, E.; Corujo, A.; Gonzalo, J.; Meij, E.; de Rijke, M.

Publication date
2012
Document Version
Final published version
Published in
Working Notes for CLEF 2012 Conference

Link to publication

Citation for published version (APA):
Amigó, E., Corujo, A., Gonzalo, J., Meij, E., & de Rijke, M. (2012). Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In P. Forner, J. Karlgren, C. Womser-Hacker, & N. Ferro (Eds.), *Working Notes for CLEF 2012 Conference: Rome, Italy, September 17-20, 2012* (CEUR Workshop Proceedings; Vol. 1178). CEUR-WS. http://ceur-ws.org/Vol-1178/CLEF2012wn-RepLab-AmigoEt2012.pdf

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Oct 2026

Overview of RepLab 2012: Evaluating Online Reputation Management Systems

Enrique Amigó*, Adolfo Corujo **,
Julio Gonzalo ***, Edgar Meij †, and Maarten de Rijke ‡

enrique@lsi.uned.es,acorujo@llorenteycuenca.com,
julio@lsi.uned.es,edgar.meij@uva.nl,derijke@uva.nl

Abstract. This paper summarizes the goals, organization and results of the first RepLab competitive evaluation campaign for Online Reputation Management Systems (RepLab 2012). RepLab focused on the reputation of companies, and asked participant systems to annotate different types of information on tweets containing the names of several companies. Two tasks were proposed: a *profiling* task, where tweets had to be annotated for relevance and polarity for reputation, and a *monitoring* task, where tweets had to be clustered thematically and clusters had to be ordered by priority (for reputation management purposes). The gold standard consisted of annotations made by reputation management experts, a feature which turns the RepLab 2012 test collection in a useful source not only to evaluate systems, but also to reach a better understanding of the notions of polarity and priority in the context of reputation management.

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

1 Introduction

Reputation management has already become an essential part of corporate communication [1]. It comprises activities aiming at building, protecting and repairing the image of people, organizations, products, or services. It is vital for companies (and public figures) to maintain the good name and preserve their "reputational capital".

Current technology applications provide users a wide access to information, enabling them to share it instantly and 24 hours a day due to constant connectivity. Information, including users' opinions about people, companies or products, is quickly spread over large communities. In this setting every move of a company, every act of a public figure are subject, at all times, to the scrutiny of a

* nlp.uned.es research group, UNED, Madrid, Spain
** Llorente y Cuenca, Madrid, Spain
*** nlp.uned.es research group, UNED, Madrid, Spain
† ISLA research group, University of Amsterdam, The Netherlands
‡ ISLA research group, University of Amsterdam, The Netherlands

powerful global audience. The control of information about public figures and organizations at least partly has moved from them to users and consumers [2]. For effective Online Reputation Management (ORM) this constant flow of online opinions needs to be watched.

While traditional reputation analysis is mostly manual, online media 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 already witnessing an unprecedented demand for text mining software for ORM. Although opinion mining has made significant advances in the last few years, most of the work has been 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[1] which aims at vertebrating research on reputation management as a "living lab": a series of evaluation campaigns in which task design and evaluation methodologies are jointly carried out by researchers and the target user communities (reputation management experts). Given the novelty of the topic (as compared with opinion mining on product reviews and mainstream topic tracking), an evaluation campaign should maximize the use of data collections built within LiMoSINe, the academic interest on tasks with practical relevance, and the standardization of evaluation methodologies and practices in the field.

RepLab has been, therefore, set out to bring together the Information Access research community with representatives from the ORM industry, aiming at:

- establishing a five-year roadmap that includes a description of the language technologies required in terms of resources, algorithms, and applications;
- specifying suitable evaluation methodologies and metrics; and
- developing test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial systems.

CLEF 2012 RepLab has been coordinated by three LiMoSINe partners: Llorente & Cuenca, University of Amsterdam, and Universidad Nacional de Educación a Distancia (UNED). In the next sections we will first define the tasks related to the ORM and then summarize the pilot tasks addressed in RepLab 2012.

2 Task Definition

Following the methodology applied by Llorente & Cuenca, one of the main Public Relations consultancies in Spain and Latin America, we distinguish between two practical ORM scenarios: *monitoring* and *profiling*. As their name suggest, the

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

former consists of a constant monitoring of online media, searching and analyzing every mention of the company, while the latter aims at distilling the reputation of the company in online media at a certain point in time.

2.1 Monitoring

In the context of ORM, *monitoring* refers to a constant (e.g. daily) scrutiny of online (and, in particular, social) media searching for information related to the entity. It focuses on the opinions and news related to a given company and aims at early detection of any potential menace to its reputation, that is, issues and opinions that could damage the company's public image. That implies a frequent inspection of the most recent online information. Microblogs and, especially, Twitter, are key sources for this task.

Proper handling of the stream of information related to an entity (we will use "company" in the discussion that follows, as it is the most typical case in reputation management) involves a number of challenging information access problems, including (but not limited to):

- *Company name disambiguation*: as monitoring is strongly recall-oriented (nothing relevant to the company should be missed), ambiguous company names may generate a lot of noise (consider Blackberry, Orange and Apple, just to mention a few fruits that are also company names). An automatic solution to this initial filtering problem would already have a major impact on the budget needed to monitor online information. An evaluation campaign focused on company name disambiguation in Twitter (WePS-3) already proved that this is not a trivial problem: the best fully automatic system had a performance of 75% accuracy, which is not impressive considering that a random baseline gets 50%.
- *Topic detection and tracking*: the ability of distinguishing what are the different issues in which a company is involved, grouping together texts that refer to the same issue, tracking issues along time, detecting novel topics, etc., is crucial for automatic reputation management and also for assisting reputation experts and facilitating their analysis task.
- *Polarity for reputation*: Does the information (facts, opinions) in the text have positive, negative, or neutral implications for the image of the company? This problem is related to sentiment analysis and opinion mining, but has substantial differences with the mainstream research in that areas: polar facts are ubiquitous (for instance, "Lehmann Brothers goes bankrupt" is a fact with negative implications for reputation), perspective plays a key role. The same information may have negative implications from the point of view of clients and positive from the point of view of investors, negative sentiments may have positive polarity for reputation (for example, "R.I.P. Michael Jackson. We'll miss you" has a negative associated sentiment - sadness -, but a positive implication for the reputation of Michael Jackson.).
- *Impact prediction*. Early detection of issues that may have a snowball effect is crucial for reputation management.

– *Focus.* What is the role of the company in a given issue? Is the company central to the topic or peripheral?

There are general underlying challenges to the issues above, such as how to process social texts (consider Twitter sublanguage, for instance) in real time; and there are also problems that build on solutions to each of the issues above, such as the assigning priority to a number of automatically detected topics at a given point in time, considering their polarity, focus, potential impact, novelty, etc.

2.2 Profiling

Profiling, as one of the aspects of ORM, refers to a single or periodic (e.g., monthly) revision of the ORM of a company as it distills from news, opinions and comments expressed in social media or online press. Unlike monitoring, which is crucially a real-time problem, profiling consists of a static survey of opinions and polar facts concerning a certain company and extracted for a given period. Normally, this information is contrasted with what has been said in the same period of time about the company's potential competitors, and with what has been said about the company in earlier periods of time. The main goal of profiling is to assess the company's positioning with respect to different aspects of its activity and with respect to its peer companies. It comprises a comparative analysis of the content related to that company, aiming at finding out what image the company projects in such dimensions as commercial, financial, social, labour or sectoral or with respect to certain topics (e.g., sustainability, innovations or leadership) and how the company's image compares to that of other companies within the same sector.

Adequate profiling implies harvesting documents from a variety of online sources, and annotating them for reputation-related aspects. Typical annotations are:

– Is the document related to the company? (see "company name disambiguation" above).
– If so, what dimensions (commercial, labour, institutional, social, financial...) of the company's activity are affected by certain content?
– Is the opinion holder a client, a citizen, an activist, a professional, etc.?
– Does the document have positive or negative implications for the reputation of the company along those dimensions? (see "polarity for reputation" above)

The types of opinion holder and the company dimensions are standard annotations (RepTrack guidelines[2]). But, of course, there is much more that can be extracted from texts. For example, detecting opinion targets (aspects of a company which are subject to opinion) is a relevant and challenging problem. Opinions about products usually have a limited and predictable set of aspects: opinions about a smartphone involve its screen, battery, camera, etc. Opinions

[2] `http://www.reputationinstitute.com`

about companies and people, on the other hand, are harder to map into a fixed set of aspects, and the relevant aspects vary quickly with time. It is, therefore, necessary to automatically identify company aspects which are subject to opinions. Computing text similarity is also relevant to group equivalent opinion targets and equivalent opinions, in order to automatically build an accurate "opinion-based entity profile".

In its first year, RepLab has addressed two pilot tasks on companies and Twitter data, each targeting one of the above scenarios (monitoring and profiling). For monitoring, Twitter is an essential source of real-time information. For profiling, Twitter is just one of the many sources that must be considered, and perhaps not the most important. But, in order to avoid an excessive complexity in the first year, we decided to focus on one single type of online social media.

A distinctive feature of both tasks is that manual annotations have been provided by online reputation management experts from a major Public Relations consultancy (Llorente & Cuenca). Such annotations are much more costly than a crowdsourcing alternative, but they have the crucial advantage that data will not only serve to evaluate systems, but also to understand the concept of reputation from the perspective of professional practitioners.

2.3 RepLab Tasks in 2012

In the *profiling task*, systems were asked to work on Twitter data (tweets containing a company name, for several companies) and annotate two kinds of basic information on tweets:

- Ambiguity: Is the tweet related to the company? For instance, a tweet containing the word "subway" may refer to the fast food company or to the underground city transport. Manual assessments have been provided by reputation management experts, with three possible values: relevant/irrelevant/undecidable. Tweets annotated as relevant/irrelevant have been used to evaluate systems.
- Polarity for Reputation: Does the tweet content have positive or negative implications for the company's reputation? Manual assessments were: positive/negative/neutral/undecidable. Tweets in the first three categories will be used to assess systems' performance.
  Note that, as discussed above, polarity for reputation is substantially different from standard sentiment analysis (polar facts, sentiment polarity different from reputation polarity). Therefore, systems were not explicitly asked to classify tweets as factual vs. opinionated: the goal is finding polarity for reputation, regardless of whether the content is opinionated or not.

For this first RepLab campaign we did not consider the problems of annotating the type of opinion holder, the dimension of the company affected by a text, the opinion targets, etc.

In the *monitoring task*, systems received a stream of tweets containing the name of an entity, and their goal was to (i) cluster the most recent tweets thematically, and (ii) assign relative priorities to the clusters. A cluster with high

priority represents a topic which may affect the reputation of the entity and deserves immediate attention.

Manual assessments included:

- Suitable topical clusters for the tweets.
- A four-level graded assessment of the priority of each cluster. Our reputation experts used four explicit priority levels: alert > average priority > low priority > irrelevant (not about the company). In addition, there is one more implicit priority level which comes from the "other" cluster. This cluster is used for tweets that are about the company, but do not qualify as topics and are negligible for the sake of monitoring purposes. Therefore, in the gold standard there are up to five priority levels:
  alert > average priority > low priority > tweets in the "other" cluster > irrelevant

These annotations have been used to evaluate the output of the systems, which is expected to be a rank of clusters containing topically similar tweets. As we mentioned above, some of the factors that may play a role in the priority assessments are:

- Novelty. Monitoring is focused on early discovery of issues that might affect the reputation of the client (the company in RepLab data); in general, already known issues are less likely to fire an alert.
- Polarity. Topics with polarity (and, in particular, with negative polarity, where action is needed) usually have more priority.
- Focus. A high priority topic is very likely to have the company as the main focus of the content ("focus" corresponds to the classical notion of relevance in Document Retrieval).
- Trendiness (actual and potential). Topics with a lot of twitter activity are more likely to have high priority. Note that experts also try to estimate how a topic will evolve in the near future. For instance, it may involve a modest amount of tweets, but from people which are experts in the topic and have a large number of followers. A topic likely to become a trend is particularly suitable to become an alert and therefore to receive a high priority.

Note, however, that the priority of a topic is determined by online reputation experts according to their expertise and intuitions; therefore, priority assessments will not always necessarily have a direct, predictable relationship with the factors above. This is precisely one of the issues that we want to investigate with this test collection.

3 Measures and Evaluation Methodology

The *monitoring task* combines two problems: clustering (into topically-related texts) and ranking (clusters must be ranked by priority). To our knowledge, there is no standard evaluation measure for this type of combined problem. We have,

therefore, dedicated part of our efforts to design a suitable evaluation measure for this problem. We have started by defining a general "document organization problem" that subsumes clustering, retrieval and filtering. We have defined an evaluation measure for this combined problem that satisfies all desirable properties from each of the subsumed tasks (expressed as formal constraints). This measure is the combination (via a weighted harmonic mean) of *Reliability* and *Sensitivity*, defined as Precision and Recall of the binary document relationships predicted by a system on the set of relationships established in the gold standard, with a specific weighting scheme.

These measures are discussed in detail in [3]. Since these metrics have the best formal properties, we can apply them to all suitable tasks in RepLab: not only to the *monitoring task*, but also to the filtering and polarity for reputation problems in the *profiling task*. Polarity for reputation cannot be evaluated with *Reliability* and *Sensitivity* if we see it as a ternary classification problem (positive, neutral, negative), but if we see the polarity annotation as a ranking problem (with positive texts first, then neutral, then negative), reliability and sensitivity also fit that problem. In fact, in other sentiment classification tasks (such as the one in Semeval), polarity annotation is seen as a ranking problem.

In the profiling problem, however, it is also needed a combined filtering / polarity measure that gives an indication of the overall success of the task. Both tasks combined can be seen as a classification problem with four classes: irrelevant (not about the company), negative, neutral and positive. These four classes cannot be mapped into a ranking problem, and therefore we cannot use *Reliability* and *Sensitivity*. We have therefore decided to use the measure that best correlates with the usefulness of the output for reputation experts. If we assume that every misclassified instance has the same impact on the expert task (i.e., fixing any automatic classification mistake takes the same effort), then the most appropriate measure is *Accuracy*: the cost of producing a manually annotated output starting from the output of the system is inversely proportional to the accuracy, i.e. the proportion of correctly classified instances.

4 Evaluation Test Beds

Both monitoring and profiling in RepLab will use Twitter data in English and Spanish. The balance between both languages depends on the availability of data for each of the companies included in the dataset.

4.1 Trial dataset

Trial data consists of at least 30,000 tweets crawled per each company name, for six companies (Apple, Lufthansa, Alcatel, Armani, Marriott, Barclays) using the company name as query, in English and Spanish. The time span and the proportion between English and Spanish tweets depends on the company.

In order to respect the Twitter terms of service, both in the trial and in the test data sets, the tweet content is not provided directly. Instead, the tweet

identifiers are given which enables to retrieve the contents by means of the "Twitter corpus tool" provided by TREC Microblog Track organizers. As for the content of the URLs mentioned in the tweets, it is given directly, since it may vary along time (tweets can disappear but not be altered).

For each company's timeline, 300 tweets (approximately in the middle of the timeline) have been manually annotated by reputation management experts. This is the "labelled" dataset. The rest (around 15,000 unannotated tweets before and after the annotated set, for each company), is the "background" dataset. Tweets in the background set have not been annotated.

The 300 tweets corresponding to each company are annotated as follows:

1. Each tweet is annotated with two fields: *related* (is the tweet about the company?) and *polarity for reputation* (does the tweet content have positive/neutral/negative implications for the company's reputation?).
2. Tweets are clustered *topically* (using topic labels)
3. Clusters are annotated for *priority* (does the cluster topic demands urgent attention from the point of view of Reputation Management?)

Tweet annotations, that is, tags for relatedness and polarity for reputation, are used to evaluate systems in the *profiling task* (see Section 2.2). Cluster annotations (tweet-cluster relations and cluster priority) are used to evaluate systems in the monitoring task (see Section 2.1).

The annotated files include the following information:

- id: id of the tweet
- user_screen_name: the Twitter username of the tweet's author.
- tweet_url: the complete URL of the tweet.
- entity_id: the id of the entity which the tweet is associated to.
- language: the language filter used in the query to retrieve the tweet.
- related: 'yes' if the tweet is related to the entity, 'no' otherwise.
- polarity: polarity for reputation: "positive"/"negative" if the content of the tweet has positive/negative implications for the reputation of the entity, otherwise it is "neutral". It is worthy to mention that this is not the same as polarity in sentiment analysis. Note also that only related tweets of identified clusters have been annotated for polarity.
- cluster: label of the cluster (topic) which the tweet belongs to.
- priority: priority of the cluster (topic) for reputation monitoring purposes. 0 - not related to the entity; 1 - related but with little relevance - 2 related, average priority; 3 - Alert (high priority). Priority for the cluster "Other topics" (related tweets on topics with too little content or relevance to have their own cluster) is not annotated.
- original_urls: list of external links included in the content of the tweet.
- unshorten_urls: list of original (unshorted) URLs in the tweet.
- md5_unshorten_urls: md5 hash of the unshorted URLs. This field can be used to get the content of the URL from a local directory that includes the content of the external links mentioned in the tweets. Each subdirectory caches the content of one URL.

For "background" tweets, that is for the tweets that have not been annotated by experts, the corpus contains the following information:

- id: id of the tweet
- user_screen_name: the Twitter username of the tweet's author.
- tweet_url: the complete URL of the tweet.
- entity_id: the Id of the entity which the tweet is associated to.
- language: the language filter used in the query to retrieve the tweet.
- original_urls: list of external links included in the content of the tweet.
- unshorten_urls: list of original (unshorted) URLs in the tweet.
- md5_unshorten_urls: md5 hash of the unshorted URLs.

4.2 Test dataset

Test data are identical to trial data, for a different set of 31 companies (Telefonica, BBVA, Repsol, Indra, Endesa, BME, Bankia, Iberdrola, "Banco Santander", Mediaset, IAG, Inditex, Mapfre, Caixabank, "Gas Natural", Yahoo, Bing, Google, ING, "Bank of America", Blackberry, BMW, BP, Chevrolet, Ferrari, Fiat, VW, Wilkinson, Gillette, Nivea, Microsoft). The tweets have been crawled using the company identifier as query. There are between 19,400 and 50,000 tweets per company name, in English and Spanish. Similarly to the trial dataset, the time span, and the proportion between English and Spanish tweets here depend on the company.

Note that: (i) unlike many test collections, in RepLab 2012 the test set is significantly larger than the trial set, which is too small to be used as proper training corpora; (ii) companies in the trial and test collections are different; therefore, systems cannot individually learn features for each company; they must learn features at a higher level of generalization. Both design decisions are intended to avoid a large set of systems that blindly apply machine learning machinery (via Weka or other similar software packages), and to push participants into creative solutions to the problem. Of course, this is a decision that will be revised and potentially changed for RepLab 2013, depending on the feedback received from participants.

For each company's timeline, approximately in the middle of it, a set of tweets has been extracted to be annotated by reputation management experts. "Unlabelled" tweets will be used to evaluate systems. For each company the "background" dataset contains the tweets before and after the annotated test set.

The "labelled" dataset with the annotations done by experts will be available for research purposes once the evaluation campaign is over.

The information associated to the entities of the test dataset includes:

- entity_id: id of the entity (e.g., RL2012E06)
- entity_name: complete name of the entity (e.g. Gas Natural SDG, S.A.)
- query: Query used to retrieve the tweets (e.g. "gas natural")
- dataset: dataset which the entity belongs to (in this case, test).

- homepage: URL of the entity's homepage (e.g., `http://www.gasnaturalfenosa.com/`).
- wikipedia_page_en: URL of the entity's Wikipedia page in English (e.g., `http://en.wikipedia.org/wiki/Gas\_Natural`)
- wikipedia_page_es: URL of the entity's Wikipedia page in Spanish (e.g., `http://es.wikipedia.org/wiki/Gas\_Natural\_Fenosa`)
- languages: list of languages used to retrieve tweets (i.e., language filters added to the query, 'lang' parameter of the Twitter Search API).
- md5_homepage: md5 hash of the homepage URL.
- md5_wikipedia_page_en: md5 hash of the entity's Wikipedia page URL in English.
- md5_wikipedia_page_es: md5 hash of the entity's Wikipedia page URL in Spanish.

5 Participants

Being a pilot exercise, RepLab 2012 attracted a remarkable number of groups – 39 – that registered for one or both tasks. Broadly speaking, the main focus of interest was the polarity for reputation problem, with a mainstream approach of using sentiment polarity detection software adapted to the reputation scenario and/or the textual source (tweets). Only 13 groups, however, finally submitted runs to the profiling task, and 3 to the monitoring task, which seemed more complex.

5.1 Monitoring

CIRGDISCO applied "Concept term expansion in tweets" to alleviate the noise and shortness of tweets. This consisted of extracting concept terms with syntactic patterns, assigning priority to terms with training data, then assigning priority to each tweet, and then voting (between tweets in the cluster) to decide the priority of the cluster. As for multilinguality, they used Bing translator to map all tweets into English.

OPTAH applied a multilingual clustering method initially developed for news, adapting it to tweets; and investigated how far polarity and intensity of sentiments can help assigning priority to topics. They concluded that sentiments are useful, but other signals are needed, in particular "good" and "bad" news terms which affect polarity for reputation but not sentiment polarity ("negative sentiment alone is not enough to detect high priority clusters")

UNED tested three approaches to the clustering part of the problem: (i) a strategy that is based on first clustering terms – instead of tweets – to deal with the short size of tweets, (ii) a clustering method that considers wikified tweets, where each tweet is represented with a set of Wikipedia entries that are

semantically related to it; and (iii) Twitter-LDA, a topic modeling approach that extends LDA by considering some of the intrinsic properties of Twitter data. For the ranking problem, UNED relied on the idea that the priority of a topic depends on the sentiment expressed in the subjective tweets that refer to it.

5.2 Profiling

BM/Yahoo! addressed both problems (ambiguity and polarity reputation) using Support Vector Machines (SVM) classifiers and lexicon-based techniques, using automatically built company profiles and bootstrapping background data (Freebase). They expanded term-based representations of tweets using Freebase and Wikipedia graphs for filtering, extracting "Related Concepts" associated to the freebase company page; and "Non Related Concepts" associated to all other freebase entries with a similar name. Finally, they submitted a combination run jointly with UNED.

CIRGDISCO participated in the filtering subtask only, with a two-pass algorithm for company name disambiguation in tweets. The algorithm makes use of Wikipedia as a primary knowledge resource in the first pass of the algorithm, and the tweets are matched across Wikipedia terms. The matched terms are then used for score propagation in the second pass of the algorithm, that also makes use of multiple sources of evidence.

Daedalus used "Stylus", their previously existing multilingual sentiment analysis software. Their approach to the polarity classification is based on (i) the information provided by a semantic model that includes rules and resources (polarity units, modifiers, stopwords) annotated for sentiment analysis, (ii) a detailed morphosyntactic analysis of the input text that permits controlling the scope of semantic units and perform a fine-grained detection of negation in clauses, and (iii) the use of an aggregation algorithm to estimate the global polarity value of the text based on the local polarity values of the different segments, which includes an outlier detection. For the filtering step, they applied named entity recognition based on dictionaries and hand coded rules. They obtained significant differences in English and Spanish, which suggests that resources matter to handle the problem well.

ILPS used the Wikipedia pages of the source entity to filter tweets, semanticising the tweets with Wikipedia pages and disambiguating on the grounds of these pages. For each entity, they automatically assembled sets of Wikipedia pages that, if linked to a tweet, indicate the relevance of the tweet for the entity.

For determining polarity, they used sentiment baseline models that aggregate polarity of terms (and iteratively expand the set of terms); models of polarity for reputation are then based on the assumption that "The impact on the reputation

of an entity as represented in a tweet is based on the sentiment the tweet causes in other users". In other words, they analyze sentiment in retweets and replies to a tweet to determine polarity for reputation of the original tweet.

Gavagai used their Ethersource software, which defines "semantic poles" via term sets. For RepLab, they considered the "customer satisfaction" semantic pole, which basically consists of hundreds of manually selected words in EN and ES, plus semiautomatic enlargement via a semantic model built from text streams. Each tweet is compared with two opposite semantic poles, using a manually set threshold to classify.

GATE used string similarity, structural similarity, contextual similarity and commonness (most frequent sense) for the filtering step. For the polarity for reputation subtask, they employed standard GATE processing plus an emoticon processor, plus Machine Learning. Notably, words were excluded as features as they led to a 5% performance drop, perhaps due to small amount of training data.

CIRGDISCO used the previously existing Opal system modified to deal with opinions in tweets, testing whether rules involving language use in social media (emoticons, slang, colloquial expressions etc.) help polarity classification, even in a language for which the polarity lexicon is small. The results were applied to determining priority in the monitoring task.

OXYME used a machine learning approach for both subtasks. Features used include query dependent features, relevancy features, tweet features and sentiment features; an important component of the relevancy features are manually provided positive and negative feedback terms.

SEOUL used a correlation coefficient to assign polarity scores to relevant words within a tweet, and then used aggregated term scores to determine the polarity of a tweet.

UIOWA built a Google Adwords based filter for ambiguity, and used several approaches (SentiWordNet, Happiness Score and Machine Learning) for Polarity.

UNED studied the feasibility of applying complex sentiment analysis methods to classifying polarity for reputation. They adapted an existing emotional concept-based system for sentiment analysis to determine the polarity of tweets, extending the system to work with English and Spanish texts and including a module for filtering tweets according to their relevance to each company. Finally, they submitted a combined run with BM/Yahoo! that uses heterogeneity-based voting techniques.

6 Results and Discussion

6.1 Profiling

Filtering Subtask Results Table 1 displays system results for the filtering subtask. Systems are ranked by the harmonic mean of their Reliability and Sensitivity (F(R,S)), and Accuracy is also reported. Note that systems that take all tweets as relevant have an accuracy of 0.71 (because in average 71% of the tweets are relevant) but an F(R,S) of 0, because they do not capture any priority relationship between pairs of tweets. The top three systems by Daedalus, however, have an accuracy which is similar to the "all relevant" baseline, but they are more informative and receive the highest $F(R, S)$ score.

Looking at the top performing systems in terms of $F(R, S)$ $(0, 26)$ and accuracy $(0, 81$ for a baseline of $0, 71)$, it seems that there is still a wide margin to improve system performance. Note that the Replab setting is, however, the most challenging setting for filtering algorithms, because the training set is small and does not use the same set of entities as the test set.

Polarity for Reputation Subtask Results Table 2 displays system results for the polarity for reputation subtask. Again, systems are ranked by the harmonic mean of their Reliability and Sensitivity (F(R,S)), and Accuracy is also reported.

The "all positive", "all neutral" and "all negative" baselines have accuracies of $0, 44$, $0, 33$ and $0, 23$, respectively. This is very different from the typical sentiment analysis estimations on Twitter, where only a small percentage (around 15%) of tweets have sentiment polarity. For reputation experts, tweets in our collection have positive or negative polarity in 67% of the cases, and only 33% do not have implications for the reputation of the companies in the sample.

In terms of $F(R, S)$, the top performing system is Daedalus_1, which performs better than the second ranked system by a margin of 18% $(0, 40$ vs $0, 34)$. This system is also very close to the best accuracy $(0, 48$ vs $0, 49$ of the top accuracy, from UNED).

Using $F(R, S)$ to compare tasks, detecting polarity seems to be - surprisingly - less challenging than the filtering task $(0, 48$ is the top result for polarity and $0, 26$ the top result for filtering). Note that accuracy tells a very different story, because it rewards baseline "all positive" behavior in the filtering task, while the polarity task, as it has three relatively balanced classes, gives lower results for baseline behaviors.

Profiling Overall Results Table 3 displays system results for the overall profiling task. Accuracy, for the combined filtering + polarity subtasks, is defined as the fraction of tweets that are correctly classified both for relevance and polarity, and it can be seen as a classification problem with four categories: irrelevant, negative, neutral and positive.

The top performing system is OXY_2, which gives the correct classification in 41% of the cases. Two other systems reach a similar performance (Gavagai

Table 1. Filtering Subtask: System Results

system	*accuracy*	R	S	$F(R,S)$
filtering_Daedalus_2	0,72	0,24	0,43	**0,26**
filtering_Daedalus_3	0,70	0,24	0,42	0,25
filtering_Daedalus_1	0,72	0,24	0,40	0,25
filtering_CIRGDISCO_1	0,70	0,22	0,34	0,23
profiling_kthgavagai_1	0,77	0,25	0,36	0,22
profiling_OXY_2	**0,81**	0,23	0,27	0,20
profiling_uiowa_1	0,68	0,18	0,29	0,18
profiling_uiowa_3	0,68	0,18	0,29	0,18
profiling_ilps_4	0,60	0,16	0,22	0,16
profiling_ilps_3	0,66	0,16	0,26	0,16
profiling_OXY_1	0,80	0,22	0,17	0,14
profiling_ilps_2	0,61	0,14	0,20	0,13
profiling_uned_3	0,71	0,17	0,25	0,13
profiling_uned_4	0,71	0,17	0,25	0,13
profiling_ilps_5	0,49	0,12	0,17	0,12
profiling_uned+BMedia_1	0,72	0,18	0,21	0,11
profiling_BMedia_1	0,74	0,17	0,12	0,10
profiling_BMedia_2	0,74	0,17	0,12	0,10
profiling_BMedia_3	0,74	0,17	0,12	0,10
profiling_BMedia_4	0,74	0,17	0,12	0,10
profiling_BMedia_5	0,74	0,17	0,12	0,10
profiling_OXY_3	0,79	0,20	0,13	0,10
profiling_OXY_4	0,79	0,20	0,13	0,10
profiling_uned_1	0,69	0,16	0,15	0,09
profiling_uned_2	0,69	0,16	0,15	0,09
profiling_GATE_1	0,52	0,12	0,13	0,09
profiling_GATE_2	0,52	0,12	0,13	0,09
profiling_OXY_5	0,78	0,16	0,11	0,08
all relevant	0,71	0	0	0
profiling_ilps_1	0,71	0	0	0
profiling_uiowa_2	0,71	0	0	0
profiling_uiowa_4	0,71	0	0	0
profiling_uiowa_5	0,71	0	0	0

Table 2. Polarity Subtask: System Results

system	*accuracy*	R	S	$F(R,S)$
polarity_Daedalus_1	0,48	0,39	0,45	**0,40**
profiling_uned+BMedia_1	0,45	0,34	0,37	0,34
profiling_BMedia_2	0,41	0,33	0,37	0,34
profiling_uiowa_2	0,35	0,31	0,39	0,33
profiling_uned_2	**0,49**	0,33	0,31	0,31
profiling_uned_4	**0,49**	0,33	0,31	0,31
profiling_BMedia_3	0,38	0,29	0,35	0,31
profiling_OPTAH_1	0,36	0,33	0,31	0,30
profiling_OPTAH_2	0,37	0,40	0,27	0,30
profiling_BMedia_5	0,41	0,29	0,32	0,29
profiling_ilps_3	0,42	0,31	0,27	0,28
profiling_kthgavagai_1	0,38	0,35	0,26	0,28
profiling_ilps_5	0,43	0,36	0,24	0,27
profiling_BMedia_1	0,43	0,29	0,27	0,27
profiling_GATE_2	0,33	0,27	0,28	0,26
profiling_uned_1	0,44	0,32	0,24	0,26
profiling_uned_3	0,44	0,32	0,24	0,26
profiling_OXY_4	0,35	0,29	0,25	0,26
profiling_uiowa_1	0,27	0,32	0,25	0,26
profiling_BMedia_4	0,39	0,27	0,26	0,25
profiling_ilps_4	0,40	0,26	0,25	0,25
profiling_OXY_2	0,36	0,28	0,27	0,25
profiling_uiowa_4	0,40	0,26	0,25	0,24
profiling_ilps_1	0,36	0,30	0,22	0,24
profiling_OXY_1	0,37	0,24	0,22	0,22
profiling_uiowa_5	0,43	0,38	0,18	0,21
profiling_OXY_5	0,38	0,29	0,20	0,21
profiling_ilps_2	0,38	0,25	0,18	0,20
profiling_OXY_3	0,38	0,27	0,18	0,19
profiling_uiowa_3	0,32	0,23	0,14	0,15
polarity_HJHL_1	0,14	0,24	0,09	0,10
polarity_HJHL_4	0,16	0,32	0,08	0,10
polarity_HJHL_2	0,13	0,20	0,07	0,08
polarity_HJHL_3	0,14	0,21	0,06	0,07
All positive	0,44	0	0	0
profiling_GATE_1	0,44	0	0	0
All neutral	0,33	0	0	0
All negative	0,23	0	0	0

and UNED with 0, 40 and 0, 39). All of them perform substantially better than the "all relevant and positive" baseline which reaches 0, 27.

Note that, from the point of view of automatic reputation analysis, making a wrong classification in 59% of the cases is, for any purpose, impractical. At least in the RepLab scenario (where systems are required to analyze entities for which they do not have specific annotated data), the profiling task is far from being solved automatically.

Table 3. Profiling Task (filtering + polarity detection): System Results

system	*accuracy*
profiling_OXY_2	**0,41**
profiling_kthgavagai_1	0,40
profiling_uned_4	0,39
profiling_OXY_1	0,39
profiling_uned_3	0,39
profiling_uned+BMedia_1	0,37
profiling_ilps_2	0,36
profiling_OXY_4	0,36
profiling_BMedia_4	0,36
profiling_ilps_3	0,35
profiling_OXY_3	0,35
profiling_GATE_1	0,35
profiling_OXY_5	0,34
profiling_ilps_4	0,34
profiling_uned_2	0,34
profiling_BMedia_5	0,34
profiling_BMedia_2	0,33
profiling_GATE_2	0,33
profiling_BMedia_1	0,33
profiling_uiowa_1	0,33
profiling_BMedia_3	0,33
profiling_uiowa_3	0,32
profiling_uned_1	0,32
profiling_ilps_5	0,29
profiling_ilps_1	0,28
profiling_uiowa_5	0,28
all relevant and positive	0,27
profiling_uiowa_2	0,27
profiling_uiowa_4	0,26
all relevant and neutral	0,26
all relevant and negative	0,18

Language Issues In RepLab 2012, all systems were required to handle two languages: English and Spanish. In principle, Spanish was more challenging be-

cause of the comparative lack of language analysis tools (for sentiment analysis and for twitter language specificities) in Spanish. Tables 4 and 5 show how the performance of systems differs between tweets in English and Spanish, respectively.

6.2 Monitoring Task

Monitoring systems in RepLab 2012 can be evaluated according to (i) the quality of the clusters they produce; (ii) the quality of the priority relationships that they specify, and (iii) the combined quality of both processes. In all cases, we can use Reliability and Sensitivity (and their harmonic mean, F(R,S)) as evaluation measure.

As a baseline, we have implemented the Hierarchical Agglomerative Clustering algorithm or HAC [4] with single linkage, which has proved to be the most competitive for related problems [5]; we have used Jaccard Word Distance as document similarity measure, and we have used only two priority levels, assigning singletons (clusters with one document) to the second level (in other words, we have used size as the only indicator of priority). We have set the stopping threshold at various levels $(0, 10, 20, ..., 100)$ to get different versions of the baseline.

Table 6 shows the results of the five runs submitted, together with the baseline system in its eleven variants. R, S and F(R,S) are reported for clustering relationships first (where they map to BCubed Precision and Recall), priority relationships, and finally all relationships together. Systems are ranked by F(R,S) on all relationships (last column).

In terms of clustering, the three participant groups have similar performance (F(R,S) is between $0,38$ and $0,40$), below the baseline algorithm (HAC) with thresholds $0, 10, 20$. Remarkably, the best system is the baseline algorithm with a threshold of 0, which means that only one big cluster is produced. That is an indication that systems are not substantially contributing to solve the problem yet.

In terms of priority relationships, the best systems (UNED and CIRGDISCO) are close but still below the best baseline.

Finally, using the overall quality measure with all relationships, the top performing system (UNED_3) is well below the best baseline ($0,29$ versus $0,41$). Of course this difference has to be put in perspective: we have implemented the baseline for eleven different values of the stopping threshold, which means that the best performing baseline has an "oracle" effect, i.e., it is using the optimal threshold setting for the test corpus. While participants were only allowed to submit five runs (and, in fact, two groups only used one run and the third group used only three).

Note that the combined measure, by pulling all relationships in one bag, is weighting cluster relationships more than priority relationships. In fact, the top performing system is the "all tweets in one cluster" baseline, which has a F(R,S) of 0 in terms of priority relationships. This happens because clusters in the test set tend to be big, and therefore produce more "same topic" relations

Table 4. Profiling: System results for English Tweets

system	*accuracy Filtering*	*accuracy Polarity*	*accuracy Profiling*	R	S	$F(R,S)$ Filt.	R	S	$F(R,S)$ Pol.
ALL NEGATIVE	0,69	0,13	0,09	0,00	0,00	0,00	0,00	0,00	0,00
ALL NEUTRAL	0,69	0,33	0,25	0,00	0,00	0,00	0,00	0,00	0,00
ALL POSITIVE	0,69	0,54	0,35	0,00	0,00	0,00	0,00	0,00	0,00
polarity_Daedalus_1	-	0,40	-	-	-	-	0,35	0,37	0,33
polarity_HJHL_1	-	0,38	-	-	-	-	0,26	0,28	0,26
polarity_HJHL_2	-	0,34	-	-	-	-	0,23	0,19	0,20
polarity_HJHL_3	-	0,35	-	-	-	-	0,24	0,18	0,19
polarity_HJHL_4	-	0,40	-	-	-	-	0,36	0,27	0,28
profiling_BMedia_1	0,72	0,45	0,35	0,18	0,16	0,13	0,28	0,23	0,24
profiling_BMedia_2	0,72	0,44	0,36	0,18	0,16	0,13	0,34	0,37	**0,35**
profiling_BMedia_3	0,72	0,36	0,34	0,18	0,16	0,13	0,32	0,33	0,31
profiling_BMedia_4	0,72	0,37	0,34	0,18	0,16	0,13	0,26	0,23	0,23
profiling_BMedia_5	0,72	0,40	0,36	0,18	0,16	0,13	0,30	0,32	0,30
profiling_GATE_1	0,52	0,54	0,39	0,15	0,17	0,13	0,00	0,00	0,00
profiling_GATE_2	0,52	0,35	0,35	0,15	0,17	0,13	0,26	0,30	0,26
profiling_ilps_1	0,69	0,38	0,28	0,00	0,00	0,00	0,28	0,21	0,22
profiling_ilps_2	0,58	0,38	0,34	0,18	0,26	0,20	0,20	0,20	0,19
profiling_ilps_3	0,61	0,46	0,37	0,21	0,32	0,23	0,25	0,25	0,24
profiling_ilps_4	0,57	0,46	0,35	0,21	0,29	0,23	0,21	0,19	0,19
profiling_ilps_5	0,45	0,53	0,31	0,16	0,21	0,16	0,24	0,11	0,14
profiling_kthgavagai_1	0,72	0,45	0,42	0,28	0,38	0,27	0,34	0,32	0,32
profiling_OPTAH_1	0,00	0,32	0,00	0,00	0,00	0,00	0,31	0,29	0,28
profiling_OPTAH_2	0,00	0,33	0,00	0,00	0,00	0,00	0,41	0,25	0,29
profiling_OXY_1	**0,78**	0,41	0,42	0,24	0,22	0,17	0,28	0,22	0,23
profiling_OXY_2	**0,78**	0,44	**0,47**	0,23	0,32	0,22	0,27	0,23	0,22
profiling_OXY_3	0,77	0,46	0,42	0,20	0,17	0,12	0,31	0,19	0,21
profiling_OXY_4	0,77	0,40	0,40	0,20	0,17	0,12	0,31	0,30	0,29
profiling_OXY_5	0,77	0,45	0,40	0,20	0,17	0,12	0,30	0,19	0,20
profiling_uiowa_1	0,71	0,35	0,37	0,23	0,30	0,21	0,31	0,28	0,28
profiling_uiowa_2	0,69	0,40	0,30	0,00	0,00	0,00	0,32	0,38	0,33
profiling_uiowa_3	0,71	0,40	0,38	0,23	0,30	0,21	0,28	0,17	0,20
profiling_uiowa_4	0,69	0,46	0,30	0,00	0,00	0,00	0,32	0,26	0,27
profiling_uiowa_5	0,69	**0,55**	0,35	0,00	0,00	0,00	0,26	0,15	0,18
profiling_uned_1	0,67	0,52	0,35	0,18	0,11	0,09	0,24	0,14	0,17
profiling_uned_2	0,67	0,54	0,36	0,18	0,11	0,09	0,27	0,20	0,21
profiling_uned_3	0,72	0,52	0,44	0,21	0,26	0,17	0,24	0,14	0,17
profiling_uned_4	0,72	0,54	0,45	0,21	0,26	0,17	0,27	0,20	0,21
profiling_uned+BMedia_1	0,70	0,47	0,37	0,20	0,19	0,12	0,37	0,35	0,35
filtering_CIRGDISCO_1	0,72	0,00	0,16	0,30	0,38	**0,29**	0,00	0,00	0,00
filtering_Daedalus_1	0,53	0,00	0,17	0,24	0,33	0,23	0,00	0,00	0,00
filtering_Daedalus_2	0,67	0,00	0,19	0,30	0,44	0,32	0,00	0,00	0,00
filtering_Daedalus_3	0,65	0,00	0,19	0,29	0,43	0,30	0,00	0,00	0,00

Table 5. Profiling: System Results for Spanish Tweets

System	Accuracy Filtering	Accuracy Polarity	Accuracy Profiling	R Filt.	S Fil.	F(R,S) Filt.	R Pol.	S Pol.	F(R,S) Pol.
All negative	0,73	0,26	0,21	0,00	0,00	0,00	0,00	0,00	0,00
All neutral	0,73	0,34	0,26	0,00	0,00	0,00	0,00	0,00	0,00
All positive	0,73	0,40	0,26	0,00	0,00	0,00	0,00	0,00	0,00
polarity_Daedalus_1	0,00	**0,48**	0,00	0,00	0,00	0,00	0,41	0,45	**0,41**
polarity_HJHL_1	-	0,00	-	-	-	-	0,00	0,00	0,00
polarity_HJHL_2	-	0,00	-	-	-	-	0,00	0,00	0,00
polarity_HJHL_3	-	0,00	-	-	-	-	0,00	0,00	0,00
polarity_HJHL_4	-	0,00	-	-	-	-	0,00	0,00	0,00
profiling_BMedia_1	0,74	0,39	0,30	0,16	0,06	0,07	0,26	0,23	0,23
profiling_BMedia_2	0,74	0,42	0,32	0,16	0,06	0,07	0,30	0,31	0,28
profiling_BMedia_3	0,74	0,42	0,30	0,16	0,06	0,07	0,28	0,33	0,29
profiling_BMedia_4	0,74	0,41	0,34	0,16	0,06	0,07	0,27	0,24	0,24
profiling_BMedia_5	0,74	0,38	0,29	0,16	0,06	0,07	0,28	0,29	0,27
profiling_GATE_1	0,48	0,40	0,30	0,08	0,08	0,07	0,00	0,00	0,00
profiling_GATE_2	0,48	0,36	0,30	0,08	0,08	0,07	0,29	0,29	0,27
profiling_ilps_1	0,73	0,35	0,27	0,00	0,00	0,00	0,26	0,19	0,20
profiling_ilps_2	0,64	0,33	0,35	0,24	0,28	0,21	0,27	0,14	0,16
profiling_ilps_3	0,69	0,38	0,35	0,25	0,31	0,23	0,24	0,20	0,21
profiling_ilps_4	0,63	0,35	0,34	0,19	0,21	0,16	0,25	0,22	0,23
profiling_ilps_5	0,58	0,37	0,33	0,19	0,20	0,15	0,31	0,18	0,22
profiling_kthgavagai_1	**0,83**	0,37	0,41	0,31	0,37	0,28	0,28	0,14	0,18
profiling_OPTAH_1	0,00	0,44	0,00	0,00	0,00	0,00	0,30	0,29	0,28
profiling_OPTAH_2	0,00	0,44	0,00	0,00	0,00	0,00	0,37	0,26	0,28
profiling_OXY_1	0,84	0,37	0,40	0,22	0,17	0,16	0,26	0,23	0,23
profiling_OXY_2	0,86	0,37	0,42	0,25	0,26	0,24	0,29	0,28	0,27
profiling_OXY_3	0,82	0,34	0,35	0,21	0,15	0,14	0,21	0,17	0,18
profiling_OXY_4	0,82	0,36	0,36	0,21	0,15	0,14	0,26	0,20	0,22
profiling_OXY_5	0,73	0,31	0,25	0,00	0,00	0,00	0,23	0,19	0,20
profiling_uiowa_1	0,71	0,21	0,31	0,26	0,33	0,24	0,29	0,20	0,21
profiling_uiowa_2	0,73	0,29	0,23	0,00	0,00	0,00	0,29	0,34	0,30
profiling_uiowa_3	0,71	0,29	0,36	0,26	0,33	0,24	0,22	0,13	0,14
profiling_uiowa_4	0,73	0,37	0,27	0,00	0,00	0,00	0,25	0,28	0,24
profiling_uiowa_5	0,73	0,34	0,26	0,00	0,00	0,00	0,13	0,01	0,02
profiling_uned_1	0,76	0,37	0,36	0,17	0,28	0,18	0,33	0,17	0,22
profiling_uned_2	0,76	0,43	0,38	0,17	0,28	0,18	0,29	0,28	0,28
profiling_uned_3	0,71	0,37	0,37	0,14	0,31	0,15	0,33	0,17	0,22
profiling_uned_4	0,71	0,43	0,38	0,14	0,31	0,15	0,29	0,28	0,28
profiling_uned+BMedia_1	0,77	0,43	**0,39**	0,17	0,30	0,19	0,31	0,31	0,30
filtering_CIRGDISCO_1	0,72	-	-	0,31	0,46	**0,32**	-	-	-
filtering_Daedalus_1	0,79	-	-	0,25	0,38	0,25	-	-	-
filtering_Daedalus_2	0,71	-	-	0,20	0,34	0,21	-	-	-
filtering_Daedalus_3	0,69	-	-	0,20	0,33	0,21	-	-	-

than priority ("this tweet has more priority than this other tweet") relations. It seems that a more refined version of F(R,S) that takes into account the balance between different types of relationships is required.

In any case, it seems obvious that the monitoring problem is a complex one, which probably cannot be solved with current state of the art techniques. At least, certainly not with the ones tested at RepLab.

Table 6. Monitoring Task: Overall Results

System	CLUSTERING			PRIORITY			ALL		
	R (BCubed P)	S (BCubed R)	F(R,S)	R	S	F(R,S)	R	S	F(R,S)
baseline0%	0,40	1	**0,50**	0	0	0	0,40	0,43	**0,41**
baseline10%	0,50	0,70	0,49	0,35	0,16	0,16	0,34	0,33	0,33
baseline20%	0,89	0,32	0,42	0,32	0,33	**0,28**	0,38	0,26	0,30
UNED_3	0,72	0,32	0,40	0,25	0,30	0,26	0,32	0,26	0,29
baseline30%	0,95	0,26	0,35	0,32	0,28	0,27	0,37	0,23	0,27
baseline40%	0,97	0,23	0,34	0,31	0,31	0,27	0,35	0,21	0,26
baseline50%	0,97	0,22	0,33	0,31	0,31	0,27	0,35	0,21	0,26
baseline60%	0,97	0,21	0,32	0,30	0,30	0,27	0,34	0,20	0,25
baseline70%	0,98	0,20	0,31	0,30	0,30	0,27	0,33	0,20	0,25
cirgdisco_1	0,95	0,24	0,35	0,24	0,30	0,24	0,29	0,22	0,25
baseline80%	0,98	0,19	0,29	0,30	0,30	0,26	0,33	0,19	0,24
baseline90%	0,98	0,17	0,27	0,29	0,29	0,25	0,31	0,17	0,22
OPTAH_1	0,70	0,34	0,38	0,19	0,16	0,16	0,37	0,19	0,22
baseline100%	0,98	0,17	0,26	0,28	0,27	0,24	0,30	0,16	0,20
UNED_2	0,85	0,34	0,39	0	0	0	0,85	0,09	0,14
UNED_1	0,90	0,20	0,30	0	0	0	0,90	0,05	0,10

6.3 Comparative analysis of measures

Being $F(R, S)$ a novel evaluation measure which is used in all RepLab tasks, it is interesting to compare its behavior with other standard measures.

Filtering $F(R, S)$ and Accuracy return different results although, in general, a high $F(R, S)$ implies a high Accuracy (but not viceversa). We can measure to what extent an improvement between two systems is verified with both measures using UIR [6]. UIR computes a statistic on the number of test cases (companies in our test collection) in which a system is better than the other for both measures. $UIR > 0.25$ is an indicator of a robust improvement when two measures are considered.

Table 7 shows, for the ten best runs according to UIR, the number of systems which improve the run with $UIR > 0.25$ and the number of systems which the

run improves with $UIR > 0.25$. Results show that Daedalus improves over a majority of the runs without being robustly improved by any other.

Table 7. Results of UIR Analysis on the Filtering Subtask: Accuracy and $F(R, S)$

system	*is improved by* (UIR(Acc,F(R,S))< 0.25)	*improves* (UIR(Acc,F(R,S))≥ 0.25)
Daedalus_2	0	28
Daedalus_1	1	27
Daedalus_3	0	27
uiowa_3	0	26
uiowa_1	0	26
kthgavagai_1	0	26
CIRGDISCO_1	0	22
OXY_2	0	20
ilps_3	6	19
uned_4	0	15

Polarity . In this case, Table 8 shows that Daedalus robustly improves on the rest of runs, and UNED+Bmedia_1, Bmedia_2 are only improved by Daedalus.

Table 8. Results of UIR Analysis on the Polarity for Reputation Subtask: Accuracy and $F(R, S)$

system	*is improved by* (UIR(Acc,F(R,S))< 0.25)	*improves* (UIR(Acc,F(R,S))≥ 0.25)
Daedalus_1	0	37
uned+Bmedia_1	1	32
BMedia_2	1	30
uned_2	1	23
uned_4	1	23
BMedia_1	3	17
BMedia_3	2	16
BMedia_5	2	14
OPTAH_1	5	12
ilps_4	3	12

Monitoring For the clustering problem there is a clear trade-off between Reliability and Sensitivity. Therefore, the ranking results may vary substantially depending on the weight that we give to each measure (in our case, we assigned equal weights by using the harmonic mean of both). Using UIR we can estimate how robust is the final ranking with respect to variations in the relative

weight given to R and S. With $UIR > 0.25$, Table 9 show that CIRGDISCO and UNED_2 improve over UNED_1, while OPTAH and UNED_3 do not improve on anyone or are improved by anyone, which means that the comparison between them and the rest of the systems will always be dependent on the relative weight given to R and S.

Table 9. Results of UIR Analysis on the Clustering Subtask: Reliability and Sensitivity

system	*is improved by* (UIR(R_{clus},S_{clus})< 0.25)	*improves* (UIR(R_{clus},S_{clus})≥ 0.25)
cirgdisco_1	0	UNED_1
UNED_2	0	UNED_1
UNED_1	cirgdisco_1 UNED_2	0
OPTAH_1	0	0
UNED_3	0	0

For the priority problem, Table 10 shows that there are three levels which are independent of the relative weight between R and S: CIRGDISCO and UNED_3 are the top performers, then OPTAH is in a second level, and finally UNED_1 and UNED_2 form the third level.

Table 10. Results of UIR Analysis on the Priority Subtask: Reliability and Sensitivity

system	*is improved by* (UIR(R_{prior},S_{prior})< 0.25)	*improves* (UIR(R_{prior},S_{prior})≥ 0.25)
cirgdisco_1	-	UNED_1 OPTAH_1 UNED_2
UNED_3	-	UNED_1 OPTAH_1 UNED_2
OPTAH_1	cirgdisco_1 UNED_3	UNED_1 UNED_2
UNED_1	cirgdisco_1 OPTAH_1 UNED_3	-
UNED_2	cirgdisco_1 OPTAH_1 UNED_3	-

7 Discussion

RepLab 2012 has used task specifications which are particularly challenging for systems: first, the training set and the test set use different companies; second, the training set is small (six companies), especially compared with the test set (31 companies). Our intention was to foresee an scenario where a reputation analysis web service has to deal with any query (company name) posed by any user at any time. At the same time, the lack of training material was meant to prevent submissions that consisted simply of a random choice of Machine Learning (ML) algorithms and parameters using some ML toolkit, which contribute little to the understanding of the challenges underlying the proposed tasks. With these

specifications, the tasks have turned out to be particularly challenging and well beyond the current state of the art of participant systems.

An scenario where plenty of training material is available is also realistic; in fact, this is the most common situation with clients of Public Relations agencies. Monitoring, for instance, is typically performed on a daily basis, and after a few days there is already a large volume of annotated material to work with. An optimal system, in this setting, should be able to constantly learn from the stream of annotated texts and adapt to a continuously changing stream of reputation threaths, with new events and entities appearing continuously. Focusing the next RepLab exercise in this scenario may be a natural evolution from this year's setting.

We have also observed that, being a relatively new discipline, annotations made by reputation management experts, even coming from the same agency, sometimes differ in how they classify certain types of statements. For instance, some annotators tend to think that a plain mention (without associated sentiments) is positive, because being mentioned contributes to reinforce your online profile. Other annotators keep both dimensions (popularity, or being mentioned, and notoriety, or being praised) strictly separated. Differences seem to arise from the fact that reputation analysts work with different clients that have different needs: ultimately, the notion of what is good or bad for the reputation of a company is a subjective matter, where the company has the last word. Being RepLab a close collaboration between research and industry, we expect that this pilot exercise will also contribute to create more consistent guidelines to produce reputation management reports for different companies, and also across different types of entities.

A note has to be made with respect to the reusability of test collections that use Twitter data. According to the current Twitter Terms of Service, the organization cannot distribute the tweets themselves, but rather the link to the tweets, so that participants have to download the tweets themselves. But the set of available tweets changes over time: users cancel their accounts, change their privacy settings or remove specific tweets. That means that, over time, the RepLab 2012 test collection will be continuously shrinking in size. That makes using the test collection and comparing with the state of the art more challenging than with other sources. This confirms that, in spite of Twitter being more open in nature than other platforms, working with social media poses significant challenges that go beyond the skills of computer scientists.

References

1. Hoffman, T.: Online reputation management is hot? but is it ethical? Computerworld (44) (February 2008)
2. Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., Tomokiyo, T.: Deriving marketing intelligence from online discussion. In: Proceedings of 11-th ACM International Conference on Knowledge Discovery and Data Mining, Chicago (August 2005)

3. Amigo, E., Gonzalo, J., Verdejo, F.: Reliability and sensitivity: Generic evaluation measures for document organization tasks. Technical report, UNED (2012)
4. Zhao, Y., Karypis, G., Fayyad, U.: Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery **10** (2005) 141–168 10.1007/s10618-005-0361-3.
5. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: overview of the web people search clustering task. In: 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. (2009)
6. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, M.: Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. Journal of Artificial Intelligence Research **42** (2011) 689–718