RepLab: An Evaluation Campaign for Online Monitoring Systems

Jorge Carrillo-de-Albornoz, Julio Gonzalo, and Enrique Amigó

Abstract Over a period of 3 years, RepLab was a CLEF initiative where computer scientists and online reputation experts worked together to identify and formalize the computational challenges in the area of online reputation monitoring. Two main results emerged from RepLab: a community of researchers engaged in the problem, and an extensive Twitter test collection comprising more than half a million expert annotations, which cover many relevant tasks in the field of online reputation: named entity resolution, topic detection and tracking, reputational alerts identification, reputational polarity, author profiling, opinion makers identification and reputational dimension classification. It has probably been one of the CLEF labs with a larger set of expert annotations provided to participants in a single year, and one of the labs where the target user community has been more actively engaged in the evaluation campaign. Here we summarize the design and results of the Replab campaigns, and also report on research that has built on RepLab datasets after completion of the 3-year competition cycle.

1 Introduction

Corporate reputation has been an intense subject of study in the last 30 years. It has been shown to be one of the most valuable assets of companies and organizations (Doorley and Garcia 2011). Research confirms its great influence on the behavior of all the stakeholders. To begin with, companies with better reputations engender loyalty in consumers across several generations and countries (Alsop 2006). Second, a solid reputation adds value to the actual worth of a company and awakens the interest of investors (Kreps and Wilson 1982). Finally, having a good reputation is crucial to attract highly qualified employees and thereby become more efficient and productive (Chong and Tan 2010). It is only logical that companies and

J. Carrillo-de-Albornoz · J. Gonzalo (✉) · E. Amigó
NLP & IR Group at UNED, Madrid, Spain
e-mail: jcalbornoz@lsi.uned.es; julio@lsi.uned.es; enrique@lsi.uned.es

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,
https://doi.org/10.1007/978-3-030-22948-1_20

organizations dedicate considerable resources to the management of such a key component of their business development.

Reputation management involves activities that aim at building and preserving a company's reputation. In the past, it was predominantly static, and mainly comprised building an attractive image via marketing campaigns and carefully planned corporate messages. Nowadays, social media have radically changed the traditional reputation management model, giving rise to new channels of communication between companies and their audience. Current technology applications provide users with a wide access to information, enabling them to share it instantly and 24 h 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 and every act of a public figure, are subject, at all times, to the scrutiny of a powerful global audience. The control of information about public figures and organizations has at least partly moved from them to users and consumers (Hoffman 2008; Jansen et al. 2009b; Glance et al. 2005). So that, for an effective Online Reputation Management (ORM), this constant flow of online opinions needs to be watched.

While traditional reputation analysis is mostly manual, online media make it possible to process, understand and aggregate large streams of facts and opinions about companies and individuals in an automatic manner. 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 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 modelling of these entities.

RepLab was an initiative promoted by the EU project LiMoSINe,[1] and aimed at structuring research on reputation management as a series of evaluation campaigns in which task design and evaluation methodologies are jointly developed by researchers and the target user communities (reputation management experts). The focus was on detecting challenges and opportunities for language technologies in online reputation monitoring problems, to define appropriate evaluation methodologies, build evaluation test collections with reference annotations provided by reputation experts, and run shared tasks on these collections with research labs from academia and industry.

Replab focused on Twitter data, and was designed to run in a 3-year cycle. The first evaluation campaign was held as a CLEF 2012 activity, and focused on a pilot task around the daily work of reputation experts. The monitoring task for analysts, as studied in RepLab, essentially consisted of searching the stream of tweets for potential mentions to the entity, filtering those that do refer to the entity,

[1] http://cordis.europa.eu/fp7/ict/language-technologies/project-limosine_en.html.

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 positive or negative impact on the reputation of the entity, and must be handled by reputation management experts). RepLab 2013 kept the same tasks and worked on producing a much larger, expert annotated dataset which comprises more than half a million manual annotations on tweets related to companies, universities and music bands. Finally, RepLab 2014 focused on two additional aspects of reputation analysis (reputation dimensions classification and author profiling) that complemented the tasks tackled in the previous campaigns.

In this chapter, we summarize the organization and results of RepLab evaluation campaigns, explore how RepLab datasets have been used to advance the state of the art from the end of RepLab up to now (2019), and discuss the lessons learnt along the way.

The chapter is organized as follows: Sect. 2 summarizes the three evaluation campaigns, including the participants, datasets and evaluation methodologies. Section 3 describes the tasks and their outcome. Section 4 summarizes post-RepLab research. Finally, Sect. 5 discusses the main lessons learned.

2 RepLab Evaluation Campaigns

RepLab was a competitive evaluation exercise supported by the EU project LiMoSINe. It aimed at encouraging research on Online Reputation Management and providing a framework for collaboration between academia and practitioners. A crucial feature of RepLab was that task design was jointly carried out by researchers and the target user community (reputation management experts). All evaluation campaigns were co-organized by three members of the Limosine project: Universidad Nacional de Educacion a Distancia (UNED) and University of Amsterdam (UvA) as academic partners, and the reputational experts of the consultancy firm Llorente & Cuenca and Yahoo! Research as industrial partners. The RepLab evaluation campaigns were carried out during years 2012, 2013 and 2014.

2.1 Problem Setup: Tasks and Metrics

The working scenario for RepLab is that of reputation experts constantly tracking and annotating information about a client (an entity that can be an organization, brand, individual, etc.). We focused on Twitter data for two reasons: it is a primary source to be tracked by online reputation experts, as it tends to be the online place where things happen first; it has a more open nature than other social networks (such as facebook), and therefore there are less privacy issues when downloading and working with Twitter data. Although it would have been great to work on several

social media, it proved too complex for the scope of our 3-year evaluation cycle and the resources available.

In the basic workflow of an online reputation expert working for a client, RepLab organizers identified several relevant subtasks where automation could substantially speed up the process: finding out whether tweets containing the entity name were actually about the entity (*filtering* or *disambiguation* task), annotating their reputation polarity (does the content have negative or positive implications for the reputation of the entity?), finding out which are the topics discussed about the entity, which of these topics are reputation alerts, what are the reputational dimensions of the entity involved in a topic, identifying whether tweet authors were influencers in the activity domain of the entity, etc.

Each task corresponds to a particular abstract problem, as for example binary classification (filtering), three-level classification (polarity and priority), clustering (topic detection) or ranking (author influence). A common feature of the data for all tasks is that the classes, levels or clusters tend to be unbalanced. This entails challenges both for the systems and for the definition of the 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 multi-class 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.

We also wanted to have a measure of the quality of a reputation monitoring system as a whole, i.e. as a result of the combination of all the above individual tasks. We focused on our so-called "full monitoring task" as a combination of filtering (classify relatedness content), clustering (into topically-related texts) and ranking (clusters must be ranked by priority). To our knowledge, there was no standard evaluation measure for this type of combined problem. We dedicated part of our efforts to design a suitable evaluation measure for this problem. We started by defining a general "document organization problem" that subsumes clustering, retrieval and filtering. We defined an evaluation measure for this combined problem that satisfies all desirable properties for each of the subsumed tasks (expressed as formal constraints). This measure is the combination (via a weighted harmonic mean) of *Reliability* and *Sensitivity* (Amigó et al. 2013), 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.

In evaluation, there is usually a 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 of little use with 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 stricter when penalizing non informative outputs, but at the cost of interpretability. In the RepLab evaluation campaigns we employed Accuracy as a

highly interpretable measure, and the combination of Reliability and Sensitivity (R&S) as a strict, theoretically sound measure.

R and S are combined with the F measure, i.e. a weighted harmonic mean of R and S. This combining function is grounded on measurement theory, and satisfies a set of desirable constraints. One of the most useful is that a low score according to any individual measure 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 (Amigó et al. 2011). UIR is computed over the test cases (entities in RepLab) in which all measures corroborate a difference between runs. Being S_1 and S_2 two runs and $N_{>\forall}(S_1, S_2)$ the amount of test cases for which S_1 improves S_2 for all measures:

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

Finally, we also dealt with the problem of identifying influencers in a given activity domain. This can be modeled as a binary classification task (each Twitter author must be categorized as influencer or non influencer) or as a ranking task (the system must return a list of authors with decreasing probability of being influencers). The main difference with a standard retrieval task is that the ratio of relevant authors turned out to be higher than the typical ratio of relevant documents in IR. Another differentiating characteristic is that the set of potentially influential authors is rather small, while information retrieval data sets usually consist of millions of documents. This has implications for the evaluation methodology. Most Information Retrieval measures reflect the fact that users are less likely to explore items which are deeper in the results list. It is not trivial to estimate how deep in the ranking reputation experts are expected to go; but it is obviously deeper than in a typical search, as their goal is to find as many opinion makers as possible. Hence, we decided to use *MAP* (*Mean Average Precision*), which is recall oriented and also considers the relevance of authors at lower ranks.

2.2 RepLab Datasets

RepLab comprises three different datasets built in the three evaluation campaigns (2012, 2013 and 2014):

- RepLab 2012 focused on the scenario of an online application where the user types in an entity name, and the system retrieves and organizes textual information about the entity. In this scenario, it cannot be assumed that there is entity-specific training material for the system. Therefore, training and test sets refer to different entities, and systems must be able to properly generalize on the

training data. Tweets in English and Spanish, containing the name of an entity of interest, were annotated according to several subtasks: whether the tweet talks about the entity or not, what is the reputational polarity of the tweet, which are the tweets talking about the same issue, and what is the relative importance of each issue from a reputational perspective.

- RepLab 2013 focused on the scenario where systems must help online reputation experts, who are constantly tracking and annotating information about a client (an organization, brand, individual, etc.). In this case, it is reasonable to assume that systems have previously annotated material about each entity. Tasks were the same as in 2012, and the main difference in design with respect to the 2012 dataset is that in this case, training and test materials refer to the same set of entities.
- RepLab 2014 used the same set of tweets as in 2013, expanding the annotations to two additional tasks: author profiling (who are the opinion makers and what type of activity do they have) and dimension categorization (what reputational dimension of the entity is affected by a tweet?).

RepLab datasets focus on 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 main reason for choosing Twitter is that it currently constitutes the first source for the latest news (Krishnamurthy et al. 2008), due to its ubiquitous and real-time nature, and had been little studied for automating the ORM process (Li and Li 2013; Jansen et al. 2009a).

The **RepLab 2012** manual annotations were provided by online reputation management experts from the Public Relations consultancy Llorente & Cuenca. Such annotations are much more costly than a crowdsourcing alternative, but they have the crucial advantage that data serves not only to evaluate systems, but also to understand the concept of reputation from the perspective of professional practitioners. The RepLab 2012 training dataset consists of at least 30,000 tweets crawled per each company name, for six companies[2] 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. For each company's timeline, 300 tweets (approximately in the middle of the timeline) were 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.

Test data are identical to training data, for a different set of 31 companies.[3] The tweets were crawled using the company identifier as query. There are between 19,400 and 50,000 tweets per company name, in English and Spanish. Similarly

[2]Training set: Apple, Lufthansa, Alcatel, Armani, Marriott, Barclays.

[3]Test set: 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.

to the training set, the time span, and the proportion between English and Spanish tweets here depends on the company. For each company's timeline, approximately in the middle, between 190 and 400 tweets are annotated by reputation management experts. The actual size for each entity depends on the availability of tweets at evaluation time for each company. "Labelled" tweets will be used to evaluate systems. Again, for each company the "background" dataset contains the tweets before and after the annotated test set.

The labelled data is annotated as follows by the ORM experts:

- Each tweet is first annotated with relatedness information (*yes*, if the tweet refers to the entity analysed, *no* otherwise).
- Those tweets related with the company are then labelled according to its polarity for reputation (does the tweet content have *positive/neutral/negative* implications for the company's reputation?).
- Tweets are clustered topically (using topic labels).
- Clusters are annotated for priority (does the cluster topic demand urgent attention from the point of view of reputation management?), in three levels (reputation alert, mildly important, unimportant).

Note that: (1) 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; (2) 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 were intended to avoid a large set of systems that blindly apply Machine Learning machinery, and to push participants into creative solutions to the problem.

In its second year, **RepLab 2013** focused on the daily tasks of an online reputation management expert. The collection comprises tweets mentioning 61 different entities from four domains: automotive, banking, universities and music. The domain selection was intended to offer a variety of scenarios for reputation studies. To this aim, we included (1) entities whose reputation largely relies on their products (automotive), (2) entities for which transparency and ethical side of their activity are the most decisive reputation factors (banking); (3) entities for which their reputation depends on a very broad and intangible set of products (universities) and, finally, (4) entities for which their reputation depends almost equally on their products and personal qualities (music bands and artists).

Crawling was performed from 1 June, 2012 up to 31 Dec, 2012, using each entity's canonical name as query. For each entity, at least 2200 tweets were collected: the first 700 were reserved for the training set and the last 1500 for the test collection. This distribution was set in this way to obtain a temporal separation (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. These data sets were manually labelled by thirteen annotators who were trained, guided and

constantly monitored by experts from Llorente & Cuenca. 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.

The RepLab 2013 dataset is the largest of the three produced for the RepLab campaigns, and consists of more than 142,000 labelled tweets in English and Spanish, containing more than 500,000 manual labels overall. The total annotation workload was of 21 person-month. The dataset is divided in 45,679 tweets for the training set, and 96,848 tweets for the test set.

Finally, **RepLab 2014** comprises two different datasets: the *Reputation Dimensions* Dataset and the *Author Profiling* Dataset. The first one provides additional annotations to the RepLab 2013 tweet dataset, with over 48,000 manually labelled English and Spanish tweets related to 31 entities from the automotive and banking domains. The training set is composed of 15,562 Twitter posts and 32,446 tweets are reserved for the test set. Both data sets were manually labelled by annotators trained and supervised by experts in ORM from the online division of Llorente & Cuenca.

The tweets were classified according to the RepTrak dimensions[4]: *Performance*, *Product and Services*, *Leadership*, *Citizenship*, *Governance*, *WorkPlace*, and *Innovation*. In case a tweet cannot be categorised into any of these dimensions, it was labelled as "Undefined". As in the RepLab 2013 dataset, the reputation dimensions corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets temporally situated between the training (earlier tweets) and test material (the latest tweets) in the timeline.

The Author Profiling data collection contains over 7000 Twitter profiles (all with at least 1000 followers) from the automotive and banking domains, together with an additional set of miscellaneous profiles (the idea of this extra set is to evaluate if approaches designed for a specific domain are suitable for a broader multi-domain scenario). Each profile contains (1) its screen name; (2) its profile URL, and (3) the most recent 600 tweets published by the author at crawling time.

The collection was split into training and test sets: 2500 profiles in the training set and 4991 profiles in the test set. Reputation experts from Llorente & Cuenca provided manual annotations for two subtasks: Author Categorisation and opinion makers identification. For the first task, author profiles are categorized according to the following options: *company* (i.e., corporate accounts of companies), *professional*, *celebrity*, *employee*, *stockholder*, *journalist*, *investor*, *sportsman*, *public*

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

institution, and *non-governmental organisation (NGO)*. For the second task, reputation experts manually identified opinion makers (i.e., authors with reputational influence) and annotated them as "Influencer". The profiles that were not considered opinion makers were labelled as "Non-Influencer". Profiles that could not be clearly classified into one of these categories were labelled as "Undecidable".

Note that the current amount of available tweets may be lower, as some posts may have been deleted or made private by the authors: in order to respect Twitter's terms of service, we did not provide the contents of the tweets, but only tweet ids and screen names.

2.3 Participation

Overall, the RepLab evaluation campaigns attracted a remarkable number of research teams. A total of 132 groups registered for one or more tasks (39 in 2012, 44 in 2013 and 49 in 2014). Out of them, 42 groups (from 15 countries) were able to submit runs. Broadly speaking, the main focus of interest was the filtering task, which attracted a total of 23 participants (9 in 2012 and 14 in 2013), followed by the polarity for reputation task, with 21 teams submitting runs (10 in 2012 and 11 in 2013).

The topic detection and topic priority tasks attracted less participation, with eight teams submitting runs (3 in 2012 and 5 in 2013). In 2014, eight groups participated in the Reputation Dimensions task and five groups submitted their results to the Author Profiling challenge (all of them attempted the opinion maker identification subtask, and all but one the author categorization subtask).

3 Tasks and Results

Typically, an online reputation analyst periodically performs the following tasks (with the assistance of more or less sophisticated software):

- Starts with a set of queries that cover all possible ways of referring to the client.
- Takes the set of results and filters out irrelevant content.
- Identifies the different issues (topics) in relation with the client, and groups tweets accordingly.
- Evaluates the reputational priority of each issue, establishing at least three categories: reputation alerts (which demand immediate attention), relevant topics (that the company must be aware of), and unimportant content (refers to the entity, but does not have consequences from a reputational point of view).
- Produces a reputation report for the client, summarizing the results of the analysis.

Figure 1 describes the main steps carried out during the annotation process for reputation monitoring. The process starts by selecting one of the entities assigned to the expert. In the system, each entity has a list of tweets that the expert has to annotate manually. The expert processes tweets sequentially: first, she decides whether the tweet does refer to the entity of interest or not. If the tweet is unrelated to the entity, the annotation process for the tweet finishes and the expert continues with the next tweet in the list. Otherwise, the polarity and topic annotations follow. Polarity annotation consists in deciding whether the tweet may affect positively or negatively the reputation of the entity.

Topic annotation consists of identifying the aspects and events related to the entity that the tweet refers to. If the tweet refers to an already identified topic, the tweet is assigned to it. Otherwise, the expert defines a new topic. A topic receives a label that summarizes what the topic is about, and it is also classified in a priority scale (Alert, Medium or Low). When the tweet is assigned to a topic, the annotation of the current tweet is finished.

In this process, reputational experts take into account several aspects of the tweet in order to determine the different labels described above. Some of them include the novelty of the topic (already known issues tend to be less relevant), centrality (whether the company is the main focus of the content), its potential impact, the company dimensions affected by the text, and the profile of the author (her influence and her role). The first three features focus on the tweet itself, and aim to better understand it as a whole. On the other hand, the reputation dimensions contribute to a better understanding of the topic of a tweet or group of tweets, whilst author profiling provides important information for priority ranking of tweets, as certain characteristics of the author can make a tweet (or a group of tweets) an alert, requiring special attention of reputation experts. The types of opinion holders and the company dimensions are standard annotations (RepTrack guidelines[5]), while the influence of the author must be interpreted by the expert for each specific domain.

The next subsections describe the different text understading tasks that are involved in this labelling process.

3.1 Named Entity Disambiguation

Reputation monitoring is strongly recall-oriented (nothing relevant to the company should be missed), and therefore queries are usually short and ambiguous, and may generate a lot of noise (consider Blackberry, Orange and Apple, just to mention a few companies whose names are also words for fruits). 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:

[5] https://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework.

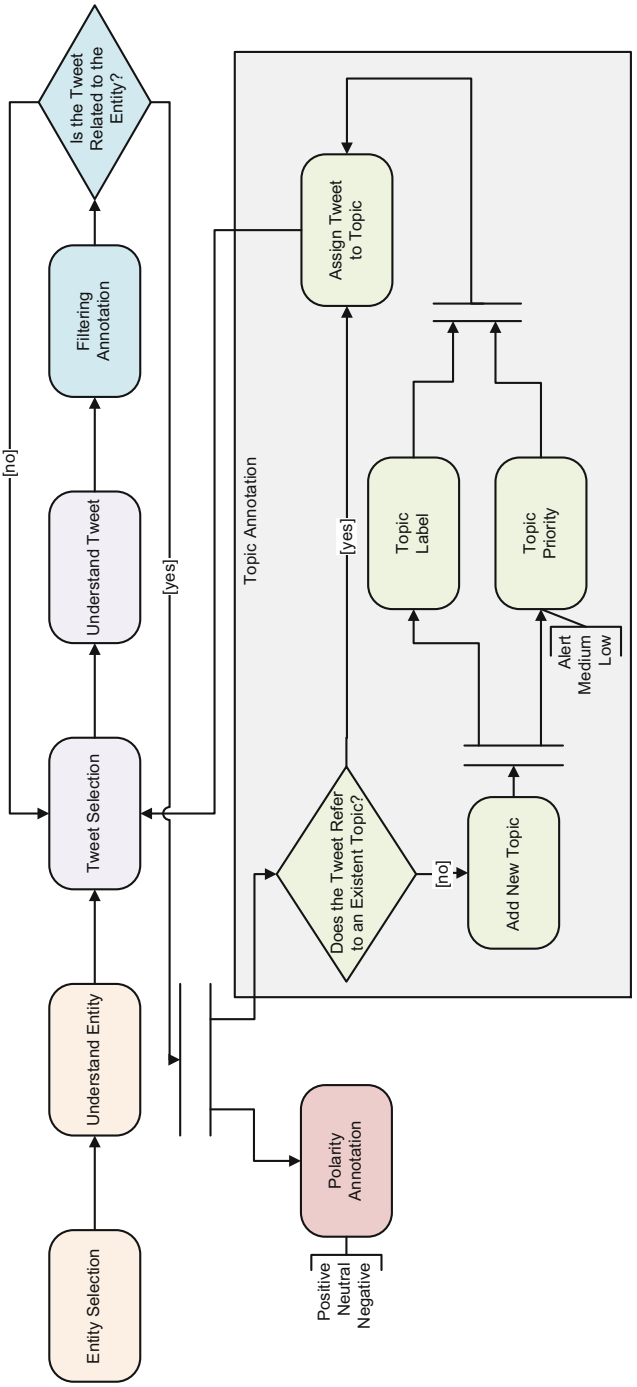

Fig. 1 Workflow of the online reputation monitoring annotation process

the best fully automatic system had a performance of 75% accuracy, which is not impressive considering that a random baseline gets 50%.

Systems were 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 were provided with two possible values: related/unrelated. As explained above, Reliability and Sensitivity were used for this task; for a filtering task, they correspond to the products of precision in both classes and the product or recall scores, respectively. Systems were ranked by the harmonic mean of their Reliability and Sensitivity (F(R,S)), and Accuracy was also reported, although classes are imbalanced to different degrees depending on the company.

Looking at the top performing systems for RepLab 2012 in terms of F(R, S) (0,26) (Villena-Román et al. 2012) and accuracy (0,81 for a baseline of 0,71) (Kaptein 2012), it seems that there is still a wide margin to improve system performance. Note that the Replab setting in this first edition was, 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. In the RepLab 2013 edition, training and test sets referred to the same company, which led to better system performance. Best systems achieved F(R,S) of 0,49 (Filgueiras and Amir 2013) and accuracy of 0,93 (for a baseline of 0,87) (Hangya and Farkas 2013), making filtering as a real candidate for a fully automatic task.

3.2 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. First, when analyzing polarity for reputation, both facts and opinions may have reputational polarity. For instance, "Barclays plans additional job cuts in the next 2 years" is a fact with negative implications for reputation. Therefore, systems were not explicitly asked to classify tweets as factual vs. opinionated: the goal was 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.

While only a small percentage (around 15%) of generic tweets have sentiment polarity, tweets talking about companies and celebrities are highly polar from the point of view of their reputational implications. According to the reputational

experts, tweets in our collections have positive or negative polarity in 67% of the cases in the 2012 RepLab collection and 73% in the RepLab 2013 collection.

Regarding the results, again, the task was much more challenging in 2012, with the best systems achieving 0,40 F(R,S) (Villena-Román et al. 2012) and 0,49 accuracy (Carrillo-de-Albornoz et al. 2012), respectively. According to F (R, S), 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" in the filtering task, while for the polarity task, as it has three relatively balanced classes, gives lower results for the baselines. In the 2013 scenario, the results of the best participants (Hangya and Farkas 2013) considerably outperform the best 2012 results in terms of accuracy (0,69), but not in terms of F(R,S) (0,38). This probably indicates that in 2013 systems were learning about the majority class, but were not generalizing adequately.

3.3 Topic Detection

The ability of distinguishing the different issues people are talking about, 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 tasks.

Systems are asked to cluster related tweets about the entity by topics, with the goal of identifying subjects/events/conversations and their relative size. Topic detection is, therefore, a clustering task that was evaluated according to R&S, which for the clustering problem corresponds to Bcubed precision and Recall (Amigó et al. 2009).

In terms of clustering, the three participant groups in 2012 (Martın et al. 2012; Qureshi et al. 2012; Balahur and Tanev 2012) achieved a similar performance (F(R,S) between 0,38 and 0,40), below the baseline algorithm provided by the organizers (Hierarchical Agglomerative Clustering) with thresholds 0, 10, 20. This was an indication that systems were not yet substantially contributing to solve the problem. Note that the topics are of a rather small size when compared to other clustering problems, and standard methods that require more data, such as LDA, turned out not to be effective in this context. 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 had an "oracle" effect, i.e., it is using the optimal threshold setting for the test corpus. The best results in 2013 (0,33 and 0,29 F(R,S), achieved by Spina et al. (2013) and Berrocal et al. (2013), respectively), are remarkably lower than those achieved in 2012, even taking into account the availability of training data. In any case, it seemed obvious that the topic detection problem is a complex one.

3.4 Topic Ranking and Alert Detection

Early detection of issues that may have a snowball effect is crucial for reputation management. 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, a topic may have 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. 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 higher 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 user (for example, in terms of the number of followers or the expertise) has better chances of receiving 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 wanted to investigate with this test collection.

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). Reliability represents the ratio of correct priority relationships per tweet, while Sensitivity represents the ratio of captured relationships per tweet. Results are quite similar to those achieved in the topic detection tasks, 0,27 F(R,S) for the best participant in 2012 (Martın et al. 2012) and 0,34 for the best participants in 2013 (Cossu et al. 2013).

3.5 Reputational Dimension Classification

One of the main goals when monitoring a company in Social Media is to assess the company's positioning with respect to different aspects of its activity and with respect to its peer companies. This involves a comparative analysis of the content related to that company, aiming at finding out what image the company projects in dimensions such as commercial, financial, social, labour or sectoral, and how the company's image compares to that of other companies within the same sector.

The aim of the Reputational Dimension classification in RepLab 2014 was to assign tweets to one of the seven standard reputation dimensions of the RepTrak

Table 1 RepTrak dimensions. Definitions and examples of tweets

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

Framework[6] developed by the Reputation Institute. These dimensions reflect the affective and cognitive perceptions of a company by different stakeholder groups. The task can be viewed as a complement to topic detection, as it provides a broad classification of the aspects of the company under public scrutiny. Table 1 shows the definition of each reputation dimension, supported by an example of a labelled tweet:

The system ranking for the Reputation Dimensions task was reported in terms of Accuracy. Note that tweets manually tagged as "Undefined" were excluded from the evaluation, and tweets tagged by systems as "Undefined" were considered as non-processed. The results achieved by the best team, 73% accuracy (McDonald et al. 2014), clearly outperform the proposed baseline (62% accuracy). Note that classifying every tweet in the most frequent class (majority class baseline) would

[6] https://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework.

get an accuracy of 56%. Most runs are above this threshold and provide, therefore, some useful information beyond a non-informative run.

3.6 Author Classification

The type of author may be of great interest when analysing the reputation of a company, as it may be a clear indicator of relevance. As an example, the influence of some profiles such as celebrities is of special interest for reputational experts, regardless of the domain expertise of the celebrity. The fact that the tweet author is an employee of the company, a journalist, an activist, etc., may have implications in the interpretation of the content and also in predicting its potential impact on the reputation of the entity.

The *Author Classification* task in RepLab 2014 was to classify Twitter profiles by type of author: Company (i.e., corporate accounts of the company itself), Professional (in the economic domain of the company), Celebrity, Employee, Stockholder, Investor, Journalist, Sportsman, Public Institution, and Non-Governmental Organisation (NGO). The system's output was expected to be a list of profile identifiers with the assigned categories, one per profile.

Accuracy values were computed separately for each domain (automotive, banking and miscellaneous). Average accuracy of the banking and automotive domains was used to rank systems. Interestingly, there is a high correlation between system scores in the automotive and banking domains (0,97 Pearson coefficient). The most relevant aspect of these results is that, in terms of accuracy, assigning the majority class (which is non informative) outperforms all runs (46%) except the best system (47%) (Cossu et al. 2014b). The question, then, is how much information are the systems able to produce. In order to answer this question we computed the Macro Average Accuracy (MAAC), which assigns the same (low) score to any non informative classifier. The results shows that most systems are able to improve the majority class baseline according to MAAC. This means that systems are able to abstract informative features of classes even if they make less accurate decisions than the majority class baseline.

3.7 Opinion Makers Identification

The capacity of influence of an author in the public opinion is a key element when aiming to determine the importance of topics about a company, and is the only key to fire an alert regardless of the content of the tweet. Some obvious aspects that determine the influence of an author in Twitter (from a reputation analysis perspective) are be the number of followers, number of comments on a domain or the type of author.

Using as input the same set of Twitter profiles as in the task above, systems had to find out which authors had more reputational influence (who the influencers or opinion makers are) and which profiles are less influential or have no influence at all. For a given domain (e.g., automotive or banking), systems were asked to rank profiles according to their probability of being an opinion maker in the domain, optionally including the corresponding weights. Note that, because the number of opinion makers is expected to be low, we modeled the task as a search problem (hence the system output is a ranked list) rather than as a classification problem.

The results for the Author Ranking task were ranked according to their average MAP using TREC_EVAL software. Unfortunately, some participants returned their results in the gold standard format (binary classification as influencers or non influencers) instead of using the prescribed ranking format. Instead of discarding those submissions, we mapped them into the official format by separating profiles marked as influencers at the top and non-influencers at the bottom of the results list, otherwise keeping the original list order.

The *followers baseline* simply ranks the authors by descending number of followers. It is clearly outperformed by most runs, indicating that additional signals provide useful information. The exception is the miscellaneous domain, where probably additional requirements over the number of followers, such as expertise in a given area, do not clearly apply. The system with the best results achieved a 0,57 MAP (McDonald et al. 2014), closely followed by Vilares et al. (2014) with a 0,56 MAP. The correlation between MAP values achieved by the systems in the automotive and banking domains seems to be low, suggesting that the performance of systems is highly biased by the domain. For future work, it is probably necessary to consider multiple domains to extract robust conclusions. On the other hand, runs from three participants exceeded 0.5 MAP, using very different approaches; Therefore, the results of the competition do not clearly point to one particular technique.

3.8 Full Monitoring Task

In 2013, the RepLab *full task* was a combination of all other tasks, and consisted of searching the stream of tweets for potential mentions the entity, filtering those that do refer to the entity, clustering relevant tweets by topic, and ranking topics based on their probability to be 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).

The use of Reliability and Sensitivity allowed us to apply the same evaluation criterion to all subtasks and therefore, to combine all of them in a single quality measure. It was possible to apply R&S directly over the full set of relationships (priority, filtering and clustering), but then the most frequent binary relationships would dominate the evaluation results (in our case, priority relationships would be predominant). Therefore, we finally decided to use a weighted harmonic mean (F

measure) of the six Reliability and Sensitivity measures corresponding to the three subtasks embedded in the full task. Due to the complex nature of this task, the results achieved by most participants were considerably low, with the best system reporting 0,19 F(R,S) (Spina et al. 2013).

This evaluation, however, is highly sensitive to the relative importance of measures in the combining function. For this reason, we also computed the Unanimous Improvement Ration (UIR) between each pair of runs. Here we considered as an unanimous improvement of system A over system B those test cases (entities) for which A improves B in all the six measures (R and S for each of the tasks). It only includes those run pairs for which UIR is bigger than 0.2. Differences in UIR turned out to be small, which indicates that the different performance of systems may not be due to intrinsic system differences, but to whether they are more optimized for reliability or sensitivity, and how this compares with the actual balance in the test data.

4 Post-competition Progress Using RepLab Datasets

RepLab evaluation campaigns have been, to the best of our knowledge, the most comprehensive effort to advance the understanding and automation of the online reputation management process. The availability of RepLab datasets, and the definition of the different tasks involved in the ORM process has encouraged researchers to investigate novel algorithms and methods for assisting reputational analysts in their daily work.

After the conclusion of the different RepLab editions, a good number of research teams have dealt with the problem of online ORM. Up to January 2018, RepLab overviews have received over 230 citations, and some of these citations come from studies using RepLab datasets.

In the **filtering** task, post RepLab research introduced active learning techniques to improve accuracy (Spina et al. 2015). These techniques emulate the real work of reputational analysts, interacting with the user for updating the classification model. Other recent works have employed Wikipedia to disambiguate the company's names in tweets (Qureshi et al. 2015). Others have generalized the problem of microblog filtering to consider topics of broad and dynamic nature (Magdy and Elsayed 2016).

The **reputational polarity** task has also attracted the attention of the research community after RepLab. As already mentioned, polarity for reputation strongly relies on the detection of polar facts, which is still an open problem. The most recent work known to us that has addressed the detection of polar facts in a reputational context is that of Giachanou et al. (2017), which determines the polarity of factual information by propagating the sentiment from sentiment-bearing text to factual texts that discuss the same issue. Giachanou et al. (2017) reported large improvements (over 50%) with respect to the use of sentiment analysis approaches. Previously, Peetz et al. (2016) explored the role of sender-based features (e.g., location, followers and user language), message-based features (e.g., hashtags, links

and punctuation marks) and reception-based features (e.g., sentiment strengths and scores from different lexicons). Before that, Gârbacea et al. (2014) outperformed state of the art methods using a simple supervised approach that considers three types of features: surface features (e.g., number of positive and negative words, emoticons, etc.), sentiment features (e.g., SentiWordNet scores of terms) and textual features (e.g., unigrams and bigrams). Overall, work on the RepLab dataset has clearly shown that sentiment analysis is only a starting point to deal with reputational polarity, but a lot more information is needed to provide usable results.

Post RepLab experiments in the **topic detection** task considerably improved the results of the competition. Spina et al. (2014) investigated whether it was possible to learn a generalized similarity function from the training data (to be fed in the clustering algorithm), and whether semantic signals could improve the topic detection process, with positive results in both cases. Their best system achieved a performance near inter-annotator agreement levels. They also found that the main source of disagreement was in the so-called organizational topics, while event-like topics, the ones most interesting from the point of view of reputation monitoring, were easier to handle by systems. Other approaches have employed transfer learning and LDA techniques by contextualizing a target collection of tweets with a large set of unlabeled "background" tweets (Martín-Wanton et al. 2013). In Panem et al. (2014), two unsupervised approaches are presented, the first based on keyword extraction and keyphrase identification, and the second based on a conceptual representation using Wikipedia.

The **priority** task has only attracted limited attention from the research community after RepLab. Cossu et al. (2014a) presented the only work that, to the best of our knowledge, has addressed the problem after the RepLab campaigns. They combine different clustering for topic detection with different priority classification methods, and conclude that actual methods are not yet mature enough to reach better performances than any priority assignment system taken alone.

With respect to **author profiling**, post-RepLab research has focused on the study of Twitter features that are relevant to characterize influential profiles (Cossu et al. 2014b), including features related to the user activity, the network topology, stylistic aspects, tweets characteristics, and profile fields. Mabrouk et al. (2018) proposed a simple model based on tf*idf and feature vector reduction. Mahalakshmi et al. (2017) propose to find the influential users in a community using a combination of the user position in networks that emerge from Twitter relations, and the textual quality of her tweets. Nebot et al. (2018) experimented with deep neural networks and word embeddings obtaining competitive results; and recently, Rodriguez et al. (2019b) investigated the different roles of authority signals (those that point out that the user is an influencer) and domain signals (those that indicate that the user is associated with the economic domain of interest) in detecting domain-specific opinion makers, and found out that both can be handled effectively with language models of influencers in the domain. Both in Nebot et al. (2018) and Rodriguez et al. (2019b), one of the salient conclusions is that text contains enough information to address the task, and additional non-textual signals, which in principle seem very

relevant for the problem, such as the number of followers, do not improve the use of textual information.

As for the task of **classification into reputational dimensions**, Qureshi et al. (2017) obtain Wikipedia dominant categories to generate "associativeness" with respect to the various reputation dimensions, and then are used in a random forest classifier, showing significant improvement over the baseline accuracy. McDonald et al. (2015) present a tweet enrichment approach that expands tweets with additional discriminative terms from a contemporary Web corpus, and that outperforms effective baselines including the top performing submission to RepLab 2014.

Work on RepLab data goes beyond the tasks defined in the evaluation exercise. A new and strongly related task has emerged post-RepLab: the automatic generation of reputational reports using the output of the tasks investigated in RepLab. Carrillo-de Albornoz et al. (2016) investigated the problem with two goals: determining if it is substantially different from a standard summarization task, and finding out appropriate evaluation metrics. Their experiments showed that producing reputation reports differs from standard summarization in the key role played by the reputational priority of information nuggets, which must be handled by systems together with centrality (the standard signal in summarization). In Rodriguez et al. (2019a), a test collection for the task of producing reputation reports is created, with extractive and abstractive summaries manually created for each of the alerts and important topics identified in each of the RepLab 2013 entities.

Finally, it is worth mentioning that the websites of the competitions have been accessed over 8000 times by more than 5000 different users, and that the datasets and results of the different systems are available for the research community in the EvALL (Amigó et al. 2017) framework (http://evall.uned.es/).

5 Discussion

Over a period of 3 years, RepLab was a CLEF Lab where computer scientists and online reputation experts worked together to identify and formalize the Natural Language Processing challenges in the area of online reputation monitoring. Two main results emerged from RepLab: a community of researchers engaged in the problem, and an extensive Twitter test collection comprising more than half a million expert annotations covering many relevant tasks in the field of online reputation: named entity resolution, topic detection and tracking, reputational alerts identification, reputational polarity, author profiling, opinion makers identification and reputational dimension classification. It has probably been the CLEF lab with the largest set of expert annotations provided to participants in a single year, and one of the labs where a user community has been more actively engaged in an evaluation initiative. Four years after completion of the lab, RepLab data is still being used by the research community.

A characteristic of the problems studied in RepLab is the size of the data to be handled per client: in a typical case, online information about a company is too much

to be processed manually, but too little to apply the simple statistics that are perfectly fit for massive trending topics. Companies are, so to speak, in the "long tail" of social media information, except for a handful of prominent multinational corporations such as Coca Cola, Apple, etc. Another key feature of dealing with reputation monitoring is that straightforward Machine Learning is usually not enough; the focus of reputation monitoring is on early discovery of the unexpected (an issue that arises about the entity that was not foreseen). And, from that point of view, a Machine Learning algorithm has to be able to generalize in a very clever way to distinguish a new reputational issue based on what has been seen and tagged before. Often, Machine Learning methods extract statistics from data that do not generalize well on new material; for instance, they can learn that "ecologist" is a term that usually correlates with something bad for the reputation of oil companies; if the unseen data unexpectedly contains some positive actions of oil companies in the environment, the algorithm will fail to analyze that content properly.

The close work with reputation experts did not stop at RepLab; in the framework of the Limosine project which funded the evaluation campaigns, the consortium built and tested annotation assistants with the help of the experts. There, we discovered that the main scientific findings in RepLab did not necessarily correlate to the techniques needed to optimize the work of the experts. For instance, in Spina et al. (2014) we discovered that semantic signals (such as entity linking of tweet terms with Wikipedia entries) could improve topic detection in a statistically significant way. In practice, however, it was preferable to deploy a system able to re-train very fast when the experts corrected an automatic detection; fast adaptive learning was far more important than the level of sophistication of the signals used for the initial automatic annotation.

Acknowledgement This research was partially supported by the Spanish Ministry of Science and Innovation (Vemodalen Project, TIN2015-71785-R).

References

Alsop RJ (2006) The 18 immutable laws of corporate reputation: creating, protecting and repairing your most valuable asset. Kogan Page, London

Amigó E, Gonzalo J, Artiles J, Verdejo F (2009) A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf Retr 12(4):461–486

Amigó E, Gonzalo J, Artiles J, Verdejo F (2011) Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. J Artif Intell Res 42(1):689–718

Amigó E, Gonzalo J, Verdejo F (2013) A general evaluation measure for document organization tasks. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, SIGIR '13, pp 643–652. https://doi.org/10.1145/2484028.2484081

Amigó E, Carrillo-de Albornoz J, Almagro-Cádiz M, Gonzalo J, Rodríguez-Vidal J, Verdejo F (2017) Evall: open access evaluation for information access systems. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information

retrieval. ACM, New York, NY, SIGIR '17, pp 1301–1304. https://doi.org/10.1145/3077136.3084145
Balahur A, Tanev H (2012) Detecting entity-related events and sentiments from tweets using multilingual resources. In: CLEF (Online Working Notes/Labs/Workshop)
Berrocal A, Luis J, Figuerola CG, Zazo Rodríguez ÁF (2013) Reina at replab2013 topic detection task: community detection. In: CLEF (Working Notes)
Carrillo-de-Albornoz J, Chugur I, Amigó E (2012) Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In: CLEF (Online Working Notes/Labs/Workshop)
Carrillo-de Albornoz J, Amigó E, Plaza L, Gonzalo J (2016) Tweet stream summarization for online reputation management. In: Ferro N, Crestani F, Moens MF, Mothe J, Silvestri F, Di Nunzio GM, Hauff C, Silvello G (eds) Advances in information retrieval. Springer International Publishing, Cham, pp 378–389
Chong WN, Tan G (2010) Obtaining intangible and tangible benefits from corporate social responsibility. Int Rev Bus Res Pap 6(4):360
Cossu JV, Bigot B, Bonnefoy L, Morchid M, Bost X, Senay G, Dufour R, Bouvier V, Torres-Moreno JM, El-Bèze M (2013) Lia at replab 2013. In: CLEF (Working Notes)
Cossu JV, Bigot B, Bonnefoy L, Senay G (2014a) Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on twitter. In: Métais E, Roche M, Teisseire M (eds) Natural language processing and information systems. Springer International Publishing, Cham, pp 154–159
Cossu JV, Janod K, Ferreira E, Gaillard J, El-Bèze M (2014b) Lia at replab 2014: 10 methods for 3 tasks. In: 4th international conference of the CLEF initiative
Doorley J, Garcia HF (2011) Reputation management: the key to successful public relations and corporate communication. Routledge, New York
Filgueiras J, Amir S (2013) Popstar at replab 2013: polarity for reputation classification. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
Gârbacea C, Tsagkias M, de Rijke M (2014) Detecting the reputation polarity of microblog posts. In: Proceedings of the twenty-first european conference on artificial intelligence. IOS Press, Amsterdam, ECAI'14, pp 339–344. https://doi.org/10.3233/978-1-61499-419-0-339
Giachanou A, Gonzalo J, Mele I, Crestani F (2017) Sentiment propagation for predicting reputation polarity. In: Jose JM, Hauff C, Altıngovde IS, Song D, Albakour D, Watt S, Tait J (eds) Advances in information retrieval. Springer International Publishing, Cham, pp 226–238
Glance N, Hurst M, Nigam K, Siegler M, Stockton R, Tomokiyo T (2005) Deriving marketing intelligence from online discussion. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. ACM, New York, NY, KDD '05, pp 419–428. https://doi.org/10.1145/1081870.1081919
Hangya V, Farkas R (2013) Filtering and polarity detection for reputation management on tweets. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
Hoffman T (2008) Online reputation management is hot – but is it ethical. Computerworld (44). https://www.computerworld.com/article/2537007/networking/online-reputation-management-is-hot----but-is-it-ethical-.html
Jansen B, Zhang M, Sobel K, Chowdury A (2009a) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169–2188
Jansen BJ, Zhang M, Sobel K, Chowdury A (2009b) Twitter power: tweets as electronic word of mouth. J Assoc Inf Sci Technol 60(11):2169–2188
Kaptein R (2012) Learning to analyze relevancy and polarity of tweets. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
Kreps DM, Wilson R (1982) Reputation and imperfect information. J Econ Theory 27(2):253–279
Krishnamurthy B, Gill P, Arlitt M (2008) A few chirps about Twitter. In: Proceedings of the first workshop on online social networks (WOSP'08), pp 19–24
Li YM, Li TY (2013) Deriving market intelligence from microblogs. Decis Support Syst 55(1):206–217. https://doi.org/10.1016/j.dss.2013.01.023. http://www.sciencedirect.com/science/article/pii/S0167923613000511

Mabrouk O, Hlaoua L, Nazih Omri M (2018) Profile categorization system based on features reduction. In: International symposium on artificial intelligence and mathematics, ISAIM 2018, Fort Lauderdale, Florida. http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Mabrouk_etal.pdf

Magdy W, Elsayed T (2016) Unsupervised adaptive microblog filtering for broad dynamic topics. Inf Process Manag 52(4):513–528

Mahalakshmi GS, Koquilamballe K, Sendhilkumar S (2017) Influential detection in twitter using tweet quality analysis. In: 2017 second international conference on recent trends and challenges in computational models (ICRTCCM), pp 315–319. https://doi.org/10.1109/ICRTCCM.2017.62

Martın T, Spina D, Amigó E, Gonzalo J (2012) Uned at replab 2012: monitoring task. In: CLEF 2012 Working Notes, CLEF

Martín-Wanton T, Gonzalo J, Amigó E (2013) An unsupervised transfer learning approach to discover topics for online reputation management. In: Proceedings of the 22nd ACM international conference on conference on information & knowledge management, ACM, New York, NY, CIKM '13, pp 1565–1568. https://doi.org/10.1145/2505515.2507845

McDonald G, Deveaud R, McCreadie R, Gollins T, Macdonald C, Ounis I (2014) University of glasgow terrier team/project abacá at replab 2014: reputation dimensions task. In: CLEF (Working Notes), pp 1500–1504

McDonald G, Deveaud R, Mccreadie R, Macdonald C, Ounis I (2015) Tweet enrichment for effective dimensions classification in online reputation management. In: 9th international AAAI conference on web and social media, pp 654–657

Nebot V, Rangel F, Berlanga R, Rosso P (2018) Identifying and classifying influencers in twitter only with textual information. In: Proceedings of the NLDB 2018

Panem S, Bansal R, Gupta M, Varma V (2014) Entity tracking in real-time using sub-topic detection on twitter. In: de Rijke M, Kenter T, de Vries AP, Zhai C, de Jong F, Radinsky K, Hofmann K (eds) Advances in information retrieval. Springer International Publishing, Cham, pp 528–533

Peetz MH, de Rijke M, Kaptein R (2016) Estimating reputation polarity on microblog posts. Inf Process Manag 52(2):193–216. https://doi.org/10.1016/j.ipm.2015.07.003. http://www.sciencedirect.com/science/article/pii/S0306457315000874

Qureshi M, Younus A, O'Riordan C, Pasi G (2015) Company name disambiguation in tweets: a two-step filtering approach. In: Information retrieval technology, vol 9460

Qureshi MA, O'Riordan C, Pasi G (2012) Concept term expansion approach for monitoring reputation of companies on twitter. In: CLEF (Online Working Notes/Labs/Workshop)

Qureshi MA, Younus A, O'Riordan C, Pasi G (2017) A wikipedia-based semantic relatedness framework for effective dimensions classification in online reputation management. J Ambient Intell Humaniz Comput 9:1403

Rodriguez J, Carrillo-de Albornoz J, Plaza L, Amigó E, Gonzalo J (2019a) Automatic generation of entity-oriented summaries for reputation management. J Ambient Intell Humaniz Comput:1–15. https://link.springer.com/article/10.1007/s12652-019-01255-9

Rodriguez J, Gonzalo J, Plaza L, Anaya H (2019b) Automatic detection of influencers in social networks: authority versus domain signals. J Assoc Inf Sci Technol 70:7

Spina D, Carrillo-de-Albornoz J, Martín-Wanton T, Amigó E, Gonzalo J, Giner F (2013) Uned online reputation monitoring team at replab 2013. In: CLEF (Working Notes)

Spina D, Gonzalo J, Amigó E (2014) Learning similarity functions for topic detection in online reputation monitoring. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. ACM, New York, NY, SIGIR '14, pp 527–536. https://doi.org/10.1145/2600428.2609621

Spina D, Peetz MH, de Rijke M (2015) Active learning for entity filtering in microblog streams. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, SIGIR '15, pp 975–978. https://doi.org/10.1145/2766462.2767839

Vilares D, Hermo M, Alonso MA, Gómez-Rodríguez C, Vilares J (2014) Lys at clef replab 2014: creating the state of the art in author influence ranking and reputation classification on twitter. In: CLEF (Working Notes), pp 1468–1478

Villena-Román J, Lana-Serrano S, Moreno C, García-Morera J, Cristóbal JCG (2012) Daedalus at replab 2012: polarity classification and filtering on twitter data. In: CLEF (Online Working Notes/Labs/Workshop), vol 60