Noname manuscript No.
(will be inserted by the editor)

Automatic Generation of Entity-Oriented Summaries for Reputation Management

Javier Rodríguez-Vidal · Jorge Carrillo-de-Albornoz · Enrique Amigó · Laura Plaza · Julio Gonzalo · Felisa Verdejo

Received: date / Accepted: date

Abstract Producing online reputation summaries for an entity (company, brand, etc.) is a focused summarization task with a distinctive feature: issues that may affect the reputation of the entity take priority in the summary. In this paper we (i) present a new test collection of manually created (abstractive and extractive) reputation reports which summarize tweet streams for 31 companies in the banking and automobile domains; (ii) propose a novel methodology to evaluate summaries in the context of online reputation monitoring, which profits from an analogy between reputation reports and the problem of diversity in search; and (iii) provide empirical evidence that producing reputation reports is different from a standard summarization problem, and incorporating priority signals is essential to address the task effectively.

Keywords Summarization · Search with diversity · Twitter · microblogs · Online Reputation Management

1 Introduction

Since the advent of Social Media, an essential part of PR (Public Relations) for organizations and individuals is Online Reputation Management, which consists of actively listening to online media, monitoring what is being said about an entity and deciding how to act upon it in order to preserve or improve the public reputation of the entity. Monitoring the massive stream of online content is the first task of online reputation experts. Given a client (e.g. a company), an online reputation expert must provide frequent (e.g. daily) re-

Javier Rodríguez-Vidal, Jorge Carrillo-de-Albornoz, Enrique Amigó, Laura Plaza, Julio Gonzalo, Felisa Verdejo
UNED IR & NLP Group. Calle Juan del Rosal, 16. 28040 Madrid (Spain) E-mail: jrodriguez@lsi.uned.es, E-mail: jcalbornoz@lsi.uned.es, E-mail: enrique@lsi.uned.es, E-mail: lplaza@lsi.uned.es, E-mail: julio@lsi.uned.es, E-mail: felisa@lsi.uned.es.

ports summarizing which are the issues that people are discussing in relation with the company, and which ones require PR actions.

A reputation report is a summary – produced by an online reputation expert – of the issues being discussed online which involve a given client: a company, organization, brand, individual... in general, an entity. In daily reputation reports, micro-blogs (and Twitter in particular) are of special relevance, as they anticipate issues that may later hit other media.

Typically, the reputation expert follows this procedure (with the assistance of more or less sophisticated software):

- Starts with a set of queries that cover all possible way of referring to the client.
- Takes the set of results and filter out irrelevant content.
- Identifies the different issues (topics) people are discussing and group tweets accordingly.
- Evaluates the priority of each issue, establishing at least three categories: reputation alerts (which demand immediate attention), important topics (that the company must be aware of), and unimportant content (refers to the entity, but do not have consequences from a reputational point of view).
- Produces a reputation summary (report) for the client summarizing the result of the analysis.

Crucially, the report must include any issue which may potentially affect the reputation of the client (reputation alerts) so that actions can be taken upon it. The summary, therefore, is guided by the relative priority of issues. This notion of priority differs from the signals that are usually considered in summarization algorithms, and it depends on many factors, including popularity (How many people are commenting on the issue?), polarity for reputation (Does it have positive or negative implications for the client?), novelty (Is it a new issue?), authority (Are opinion makers engaged in the conversation?), centrality (Is the client central to the conversation?), etc. This complex notion of priority makes the task of producing reputation-oriented summaries both a challenging (from the point of view of research) and practical (from the point of view of the market) scenario.

Our first contribution is **to create a dataset for the evaluation of reputation-oriented tweet stream summarization**, which includes manually created reports for banking and automotive companies. To the best of our knowledge, this is the first resource of its kind that is available for research. With this enabling new dataset, our main goal is to characterize the task of producing reputation reports, investigating whether it is actually a differentiated task (rather than a mere instance of the classic multi-document summarization task), and studying how it should be evaluated. In this context, we address two related research questions:

RQ1. Is Topic Priority different from Topic Centrality?

The most distinctive feature of reputation reports is that issues related with the entity are classified according to their priority from the perspective of reputation handling (the highest priority being a *reputation alert*, i.e., an issue that requires an immediate response from the entity). We want to investigate how the notion of priority translates to the task of producing extractive summaries, and how important it is to consider reputational signals of priority when building an appropriate summary.

RQ2. Is it possible to apply metrics from search with diversity to the evaluation of automatic summaries?

We propose a novel evaluation methodology that models the task of automatic summarization as producing a ranking of tweets that maximizes both coverage of topics and priority. This provides an analogy with the problem of search with diversity, where the search system must produce a rank that maximizes both relevance and coverage. We investigate the possibility of using Information Retrieval evaluation metrics within this perspective, and study how they differ from conventional summarization metrics. Note that there has been cross-fertilization of techniques between the fields of summarization and search with diversity (for instance, Maximal Marginal Relevance is a summarization technique that has been applied to search with diversity, and Learning to Rank techniques has been used for summarization), but no one has yet applied (to the best of our knowledge) search with diversity metrics to evaluate summarization systems. We compare this new methodology with the traditional summarization evaluation practices that compare manually created summaries (extractive and abstractive) with the automatically created summaries.

The rest of the article is organized as follows. We start by reviewing most relevant related work. Second, we present the *RepLab Summarization Dataset* for entity-oriented tweet stream summarization. Third, we formulate and discuss our novel evaluation methodology. Four, we describe our experimental framework and present and discuss the experimental results. Finally, we summarize the main lessons learned.

2 Related Work

In this section we first present works that have also addressed the problem of summarizing tweets. Next, we revise previous researches that have exploited the notion of priority in the summarizing process as opposed to traditional approaches that focused on centrality.

2.1 Multi-tweet summarization

2.1.1 Multi-tweet summarization

There is much recent work focusing on the task of multi-tweet summarization. Most publications rely on general-purpose techniques from traditional text summarization along with redundancy detection methods to avoid the repetition of contents in the summary (Inouye and Kalita, 2011; Takamura et al., 2011). Social network specific signals (such as user connectivity and activity (Liu et al., 2012) and time-based features (Alsaedi et al., 2016; De Maio et al., 2016; He et al., 2017)) have also been widely exploited.

Instead of ranking sentences as in traditional document summarization (Nguyen-Hoang et al., 2012) tweets are ranked to select the most salient ones for the summary. Two different types of approaches may be distinguished: feature-based and graph-based. Feature-based approaches represent tweets as sets of features, being the following the most frequently used: term frequency (Takamura et al., 2011), time delay (Takamura et al., 2011), user-based features (Duan et al., 2012) and readabilitybased features (Liu et al., 2012). Graph-based approaches usually adapt traditional summarization systems (such as LexRank (Erkan and Radev, 2004), DegExt (Litvak et al., 2013) and TextRank (Mihalcea and Tarau, 2004)) to take into consideration the particularities of Twitter posts (Inouye and Kalita, 2011; Liu et al., 2012; Sharifi et al., 2010). These approaches usually include both content-based and network-based information into the text graph.

The most popular algorithm for microblog summarization is presented in Sharifi et al. (2010). The authors propose a topic-oriented summarization system for Twitter posts, that automatically summarizes a collection of posts related to a same topic into a short, one-line summary. The system is based on the phrase reinforcement algorithm that iteratively constructs a graph for the set of post where the nodes are overlapping sequences of words (i.e. phrases) that occur both before and after the topic phrase. Nodes are labeled with the number of times each sequence of words occurs in the collection. Most overlapping phrases are selected to generate the summary.

Chakrabarti and Punera (2011) present a summarization system for tweets describing a same event that separates the problem into two subproblems: detecting segments of an event, and summarizing tweets in each segment. To segment an event, the hidden Markov model is used, while summarization of tweets within each segment is performed using a simple TF-IDF approach.

Inouye and Kalita (2011) present a comparison of different summarization methods on sets of tweets. Such methods include traditional summarization systems, such as LexRank, TextRank and SumBasic, as well as a modification of the traditional TF-IDF criteria specially designed to deal with Twitter's posts. The results showed that simple frequency based summarizers (TF-IDF and SumBasic) performed better than summarizers that incorporated more information or more complexity such as LexRank and TextRank, due to the nature of Twitter posts, which often have little structure and few words, and where syntax is often incorrect or missing.

Liu et al. (2012) propose a graph-based multi-tweet summarizer that leverage social network features, readability and user diversity for selecting representative tweets. As social network features, they consider the re-tweeted times and follower numbers of the Twitter account that produces the tweet. Diversity is introduced by preferring tweets from different user's accounts. However, the fact that one user may post from different account is not addressed. Finally, readability is assessed using traditional criteria such as the sentence length, the word length in syllables, the number of abnormal symbols and the number of out-of-vocabulary words. Similarly, Duan et al. (2012) develop a method that implements a graph-based ranking algorithm that takes into consideration both social influence of users and content quality of tweets.

More recently, Alsaedi et al. (2016) present a modification of the traditional centroid approach that includes the time dimension of tweets, so that tweets that have been centroid of the clusters for the longest time on average over a time-window are selected for the summary. Zhuang et al. (2016) create a model, called S-model, which takes advantage of two social contexts that are important for topic generation and dissemination: the impact of experts and majority users, as as well as the content diversity based on entropy measures.

Concerning the subject of the input tweets, most works have focused on sport and celebrity events (Inouye and Kalita, 2011; Sharifi et al., 2010). These events are massively reported in social networks, so that the number of tweets to summarize is huge. In this context, simple frequency-based summarizers perform well and even better than summarizers that use more complex information (Inouye and Kalita, 2011). However, the problem of summarizing tweets on a company's reputation has been, to the best of our knowledge, never tackled before and presents additional challenges derived from the less massive availability of data and the greater diversity of issues involved. The most closely related work is that of Louis and Newman (2012), which presents a method for summarizing collections of tweets related to a business. To this end, they first learn groups of related words from business news articles that describe relevant business concepts. Next, tweets related to each company are clustered using a method that combine the sentiment of a tweet and the entropy of word distribution in the cluster, so that clusters discussing common issues are ranked higher than clusters with diverse content. Finally, the clusters are ranked using information such as influential subtopic and sentiment.

2.1.2 Priority versus centrality-based summarization

Since the pioneering works in automatic summarization, centrality has been one of the most widely used criteria for content selection. Centrality refers to the idea of how much a fragment of text (usually a sentence) covers the main topic of the input text (a document or set of documents). Centrality of a sentence is often defined in terms of the centrality of the words that it contains. Given a cluster of sentences that represents a document topic, the sentences that contain more words from the centroid of the cluster are considered as central (i.e. most representative of the document topic).

A great number of summarization systems use centrality to identify relevant sentences for the summary, along with an algorithm to avoid redundancy (Erkan and Radev, 2004; Litvak and Last, 2008; Mihalcea and Tarau, 2004; Zhang et al., 2011). Concerning more recent approaches, Cho and Kim (2015) propose a social network-inspired method for the extraction of key sentences from a document. To this end, sentences are represented by their TF-IDF scores and connected according to the co-occurrence of keywords among them. Sentences are then scored based on their centrality in the co-occurrence network. Marujo et al. (2015) use a multi-document approach based on KP-centrality (i.e. the centrality of key phrases found within the text). KP are extracted from the documents using supervised machine learning on a bag of words model, and then are used to build a pseudo-passage that represents the central topic of each document (centroid). Most representative passages from each document are then extracted based on their closeness to the centroid, and then merged to build the multi-document summary. Sarkar et al. (2015) improve the computation of the similarity between sentences to produce a single summary from a set of related documents. They build a graph were nodes represent sentences and edges are added between nodes representing similar sentences. Centrality of sentences is then computed as the degree of the nodes, and next ranked according to such centrality and extracted to generate the summary.

However, the information need of users frequently goes far beyond centrality and should take into account other selection criteria such as diversity, novelty and priority. This is specially true in the reputational scenario. Although the importance of enhancing diversity and novelty in various NLP tasks has been widely studied (Clarke et al., 2008; Mei et al., 2010), reputational priority is a domain-dependent concept that has not been considered before. Other priority criteria have been previously considered in some domains and scenarios: Plaza and Carrillo-de Albornoz (2013), for instance, showed that it is possible to improve summarization of scientific articles by prioritizing sentences covering the methods and results of the experiments reported in the articles. Similarly, Meena and Gopalani (2015) used the location of the sentence in a general-domain text as the main indicator of its priority, along with the presence of named entities and proper nouns. In Fiszman et al. (2009), concepts related to treatments and disorders are given higher importance than other clinical concepts when producing automatic summaries of MEDLINE citations. In opinion summarization, positive and negative statements are given priority over neutral ones. Moreover, different aspects of the product/service (e.g., technical performance, customer service, etc) are ranked according to their importance to the user (Pang et al., 2008). This is sometimes referred as to aspect-based summarization, and has been recently tackled using convolutional neural networks (Wu et al., 2016). Priority is also tackled in query (or topic)-driven summarization, where terms from the user query are given more weight under the assumption that they reflect the user relevance criteria (Litvak and Vanetik, 2017; Nastase, 2008).

In the ORM scenario, priority refers to the importance of comments and opinions made by users for the company being analyzed. The priority detection problem in ORM was addressed at RepLab 2013 contest (Amigó et al., 2013a). The systems participating showed that priority depends on a set of relevance criteria including the centrality of the topic, the influence of users that discuss on the topic, and the sentiment of the comments (Cossu et al., 2014), to name a few. However, the results of RepLab 2013 prove that priority classification for ORM is still an open problem and that further investigation on relevant priority signals must be done.

3 The RepLab Summarization Dataset: A New Dataset for Reputation-oriented Tweet Stream Summarization

As part of the present work, and because no similar resource is available for research, we have developed the *RepLab Summarization Dataset*. To this end, we have started from the dataset released in the RepLab 2013 evaluation forum (Amigó et al., 2013a). The RepLab 2013 collection consists in a set of tweets manually annotated for the following subtasks:

– *Filtering*: Systems are asked to determine which tweets are related to an entity and which are not. Manual annotations are provided with two possible values: related/unrelated.
– *Polarity for reputation classification*: The goal is to decide if the tweet content has positive or negative implications for the company's reputation. Manual annotations are: positive/negative/neutral.
– *Topic detection*: Systems are asked to cluster related tweets about the entity by topic with the objective of grouping together tweets referring to the same subject, event or issue.
– *Priority assignment*: It involves detecting the relative priority of topics. Manual annotations have three possible values: Alert/mildly_important/ unimportant.

RepLab 2013 uses Twitter data in English and Spanish. The collection comprises tweets about 61 entities from four domains: automotive, banking, universities and music. For the development of the RepLab Summarization dataset, we only use tweets from the *automotive* and *banking* domains, because they consist of large companies, i.e. Santander, Barclays, Audi, BMW, etc. , which are the standard subject of reputation monitoring as it is done by experts: the annotation of *universities* and *music bands and artists* is more exploratory and does not follow widely adopted conventions as in the case of companies. Moreover, we only use those tweets that are manually annotated as related to the entity (i.e., we discard the non-related tweets). As a result, our subset of RepLab 2013 comprises 71,303 English and Spanish tweets distributed as shown in Table 1. Language detection is done using the "langdetect"[1] library, by taking those tweets for which the probability of belonging to a certain language, English or Spanish, was greater than 95%.

To develop our summarization dataset, we presented to an annotator the tweets grouped by topic (since these clusters are already manually annotated in the RepLab 2013 dataset). Only "Alert" and "Mildly important" topics are considered: we discard "Unimportant" topics, as we consider them irrelevant for summarization purposes. For each tweet in a topic, the following information is available: the `ID` or unique identifier of the tweet, the `date` when the

[1] `https://code.google.com/p/language-detection/`

	Automotive	Banking	Total
Entities	20	11	31
# Tweets (training)	15,123	7,774	22,897
# Tweets (test)	31,785	16,621	48,406
# Tweets (total)	46,908	24,395	71,303

Table 1: Subset of RepLab 2013 used in the *RepLab Summarization dataset*

tweet was written, the number of `followers` of the author of the tweet, the `reputational polarity` of the tweet, and the `text` of the tweet.

For each topic, we asked the annotator to generate:

- An *extractive summary*, selecting the tweet or tweets that best summarize the content of the topic. The annotator was allowed to make no selections if she considered that no tweet is representative of the topic. We asked the annotators to be very careful not to include redundant tweets in the selection. If two tweets are equivalent for summarization purposes, the annotator was instructed to select the tweet whose author has more followers and, in case of a tie, to pick the one that was created first. In practice, the number of tweets selected as a representative summary ranges from 0 to 3.
- An *abstractive summary*, writing a paragraph that summarizes the content of the topic, both in English and in Spanish (note that the RepLab dataset contains tweets in both languages).

As a result, for each entity in the dataset we obtained (i) an **extractive summary** that consists of the list of tweets that summarize each of the topics for that entity, ordered by priority; and (ii) two **abstractive summaries** (one in English and one in Spanish), which are the concatenation of the paragraphs that summarize each of the alerts and mildly important topics. In order to create the manual abstractive summaries, the annotator proceeded as follows: (i) he read both English and Spanish tweets; (ii) he wrote Spanish summaries and (iii) he translated Spanish summaries into English. The average number of words in a entity's abstract depends on the domain and the language. Spanish abstracts in the automotive domain have, on average, 391 words while in the banking domain the average number of words per abstract is 677. For English abstracts, average number of words is 323 for automotive and 553 for banking. Average sentence length is 4.4 words in Spanish abstracts and 3.7 in English ones. Figure 1 shows the manual summaries generated for a topic (cluster) from the RepLab Summarization dataset.

Listing 1: Example summaries for a RepLab topic referring to BMW

```
<cluster label="Google and BMW rated most attractive employers
    by European business , engineering students" priority="
   mildly\_important">
<tweet id="278778028023230464" date="Wed Dec 12 09:27:15 CET
   2012" followers="875973" polarity="positive"> Google , BMW
   rated most attractive employers by European business ,
       engineering students http://tnw.to/f0ZH1 by
       @robinwauters</tweet>
<tweet id="278781059162849280" date="Wed Dec 12 09:39:18 CET
       2012" followers="556" polarity="positive"> #Google, #BMW
       rated most attractive employers by #European business,
       engineering students http://j.mp/S9haKD</tweet>
<tweet id="279123524097028096" date="Thu Dec 13 08:20:08 CET
       2012" followers="814" polarity="positive"> Google and BMW
       are the Most Attractive Employers for Europeans via
       PRNewswire http://pop.to/14zxv</tweet>
<summary
abstract\_EN="Google and BMW are the Most Attractive Employers
        for Europeans"
abstract\_ES="Google y BMW son elegidos como los empleadores
       mas atractivos para los europeos"
extract="278778028023230464"
</cluster>
```

4 A New Metric for Reputation-oriented Tweet Stream Summarization

Standard summarization evaluation makes use of ROUGE metrics (Lin, 2004) to measure the content overlap between a peer (an automatic summary) and one or more reference (manual) summaries (also called “models”). The need of model summaries, is however, one of the main disadvantages of ROUGE, since the production of manual summaries for large evaluation collections is a very time-consuming task.

As an alternative evaluation approach for extractive summaries, we propose to evaluate summarization systems in a way that does not require model summaries. We interpret the extractive summary produced by the system as a ranked list of tweets, and evaluate the quality of the ranking in terms of relevance and redundancy with respect to the manually annotated topics and topic priorities in the RepLab dataset. The idea is to make an analogy between the task of producing a summary and the task of document retrieval with diversity. When considering diversity, the retrieval system must provide a ranked list of documents that maximizes both relevance (documents are relevant to the query) and diversity (redundance is minimized: documents reflect the different query intents, when the query is ambiguous, or the different facets in the results when the query is not ambiguous).

Producing an extractive summary is, in fact, a similar task: the set of selected sentences should maximize relevance (they should convey essential information from the documents) and diversity (sentences should minimize redundancy and maximize coverage of the different information nuggets in the documents). The case of reputation reports using Twitter as a source is even more clear, as relevance is modeled by the priority of each of the topics. An optimal report should maximize the priority of the information conveyed and the coverage of high-priority entity-related topics (which, in turn, minimizes redundancy).

Let us now consider the following user model for tweet summaries: the user starts reading the summary from the first tweet. At each step, the user goes on to the next tweet or stops reading the summary, either because she is satisfied with the knowledge acquired so far, or because she does not expect the summary to provide further useful information. User satisfaction can be modeled via two variables: (i) the probability of going ahead with the next tweet in the summary; (ii) the amount of information gained with every tweet. The amount of information provided by a tweet depends on the tweets that precede it in the summary: a tweet from a topic that has already appeared in the summary contributes less than a tweet from a topic that has not yet been covered by the preceding tweets. To compute the expected user satisfaction, the evaluation metric must also take into account that tweets deeper in the summary (i.e. in the rank) are less likely to be read, weighting the information gain of a tweet by the probability of reaching it. We can measure the expected user satisfaction by adapting Rank-Biased Precision (RBP) (Moffat and Zobel, 2008), an Information Retrieval evaluation measure which is defined as:

$$RBP = (1-p)\sum_{i=1}^{d} r_i * p^{i-1} \qquad (1)$$

where r_i is a known function of the relevance of document at position i, p is the probability of moving to the next document, and RBP is defined as utility/effort (expected utility rate), with utility being $\sum_{i=1}^{d} r_i * p^{i-1}$ and $1/(1-p)$ the expected number of documents seen, i.e. the effort.

We prefer RBP to other diversity-oriented evaluation metrics because it naturally fits our task, the penalty for redundancy can be incorporated without changing the formula (we only have to define r_i adequately), and because it has been shown to comply with more desired formal properties than all other IR measures in the literature (Amigó et al., 2013b), and can be naturally adapted to our task.

Indeed, the need to remove redundancy and the relevance of priority information can be incorporated via r_i. We propose to model r_i according to two possible scenarios. In the first scenario, incorporating more than one tweet from a single topic still contributes positively to the summary (but increasingly less than the first tweet from that topic). This is well captured by the reciprocal of the number of tweets already seen from a topic (although many other variants are possible):

$$r_i = \frac{1}{|\{k \in \{1 \dots i-1\} | \text{topic}(i) = \text{topic}(k)\}|} \qquad (2)$$

We will refer to RBP with this relevance formula as **RBP-SUM-R** (RBP applied to SUMmarization with a Reciprocal discount function for redundancy).

In the second scenario, each topic is exhaustively defined by one tweet, and therefore only the first tweet incorporated to the summary, for each topic, contributes to the informative value of the summary. Then the relevance formula is binary:

$$r_i = \begin{cases} 1 & \text{if } \forall k \in \{1..i-1\}\text{topic}(i) \neq \text{topic}(k) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We will refer to RBP with this relevance formula as **RBP-SUM-B** (RBP applied to SUMmarization with a Binary discount function for redundancy). With respect to the parameter p (probability of going ahead reading the summary after reading a tweet), we must aim at large values, which better reflect the purpose of the summary. For instance, a value of $p = 0.95$ means that the user has a 60% chance of reading beyond the first ten tweets, and a value of $p = 0.5$ decreases that probability to only 0.1%. Figure 1 shows how the probability of reading through the summary decays for different values of p. We will perform our experiments with the values $p = 0.9$ (which decays fast for a summarization task) and $p = 0.99$ (which has a slower but still representative decay).

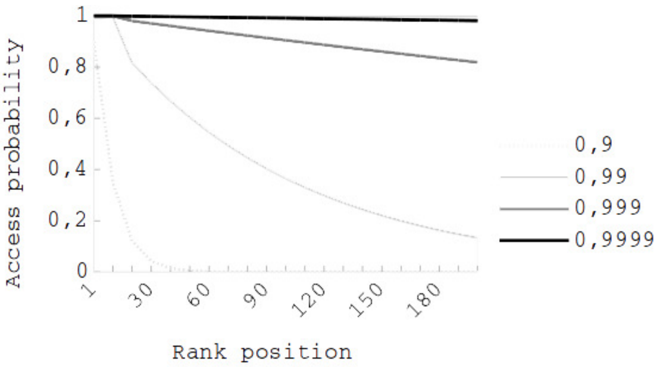

Fig. 1: Probability of reading through the summary for different p values

Our proposed metrics, RBP-SUM-R and RBP-SUM-B, have some **benefits** with respect to summarization metrics based on peer to model comparisons such as ROUGE:

- The first advantage is that they follow a user model: there is an underlying hypothesis of what users do and why. Conventional summarization metrics are agnostic with respect to what the users need from summaries, although we know that, for instance, the task of producing an indicative summary is quite different from the task of producing an informative summary.
- The second benefit is that our metric is interpretable: we can inspect what systems are doing right and what they are doing wrong and use the information from the metric to improve them, because we can inspect which topics are being properly summarized and which are not, how well systems remove redundancy and how well they select relevant information, etc.

– A third advantage is that the metric can accommodate different redundancy and priority weights, something which is not possible with ROUGE and most other (automatic) summarization metrics.
– Finally, our metric eliminates the need to create several model summaries for the same text stream, as they operate directly on the information that the summaries should have (i.e., priority topics) rather than on the several possible instantiations of that information in equally valid summaries. Note that we also eliminate the need to specify the compression rate both of the manual summaries and of the automatic summaries, which again makes producing datasets and evaluating systems more cost-effective.

Of course there are also some disadvantages with respect to the use of ROUGE-like metrics. The most important limitation is that RBP-SUM can only be applied to extractive summarization. Another issue is that we need to fix an appropriate value for p, which requires specific experimentation for each dataset.

5 Experimental design

In order to answer our second research question (*Is it possible to apply metrics from search with diversity to the evaluation of reputation-oriented automatic summaries?*), we will compare the results of our evaluation metric with those of the *de-facto* standard metric ROUGE-2, when evaluating a number of automatic summarizers on the RepLab Summarization Dataset. We selected ROUGE-2 variant due to its high correlation with human judges in many test collections. ROUGE-2 counts the number of bi-grams that are shared by the peer and reference summaries and computes a recall-related measure (Lin, 2004). These same experiments will allow us to investigate our first research question: *what is the relationship between centrality and priority?*; *May priority signals be effectively used to enhance summaries?*

For this purpose, we will compare five summarization approaches: two baselines and three contrastive systems, that will be evaluated using both ROUGE and RBP-SUM metrics. We build summaries at 5, 10, 20 and 30% compression rate, for all the approaches. The two baselines are described below:

LexRank. As a standard summarization baseline, we use LexRank (Erkan and Radev, 2004), one of the most popular centrality-based methods for multi-document summarization. LexRank is executed through the MEAD summarizer (Radev et al., 2004) (http://www.summarization.com/mead/) using these parameters: `-extract -s -p 10 -fcp delete`. It is worth noting that MEAD removes redundancy.

Followers. The number of followers is a basic indication of priority: things being said by people with more followers are more likely to spread over the social networks. As a baseline system based on priority, we simply rank the tweets by the number of followers of the tweet's author, and then apply a technique to remove redundancy. Redundancy is avoided using an iterative greedy algorithm: a tweet from the ranking is included in the summary only if it has a vocabulary overlap less than 0.02, in terms of the Jaccard measure, with each of the tweets already included in the summary. Once the process is finished, if the resulting compression rate is higher than desired, discarded tweets are reconsidered and included by recursively increasing the threshold in 0.05 similarity points until the desired compression rate is reached.

We now discuss the contrastive systems. The first contrastive system (**SSV**) is based on distant supervision methods, and uses a set of signals, derived from previous approaches used for detecting priority, to generate a ranking of tweets. The second contrastive system (**L2R**) is fully supervised, and uses Learning to Rank techniques to generate the ranking of tweets. The third approach (**SS/L2R**) is the combination of the two previous approaches.

Signal Selection & Voting (SSV). This contrastive system considers a number of signals of priority and content quality. Each signal (selected using the training set) provides a ranking of all tweets for a given test case (an entity), and we then combine the rankings using a voting procedure. The details of the algorithm are:

- Using the training part of the RepLab dataset, we compute two estimations of the quality of each signal: (i) the ratio between average values within priority values (if priority tweets receive higher values than unimportant tweets, the signal is useful), and (ii) the Pearson correlation between the signal values and the manual priority values. The signals (which are self-descriptive) and the indicators are displayed in Figures 2 and 3.

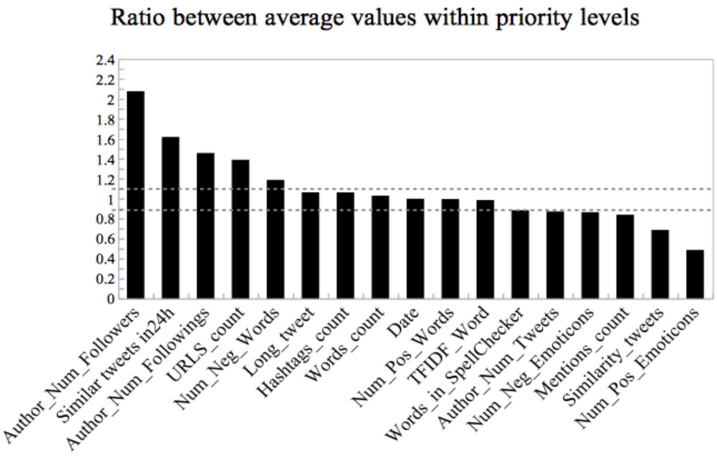

Fig. 2: Ratio between average values for priority vs unimportant topics

- We retain those signals with a Pearson correlation above 0.02 and with a ratio of averages above 10%. The resulting set of signals is: `URLS count` (number of URLs in the tweet), `24h similar tweets` (number of simi-

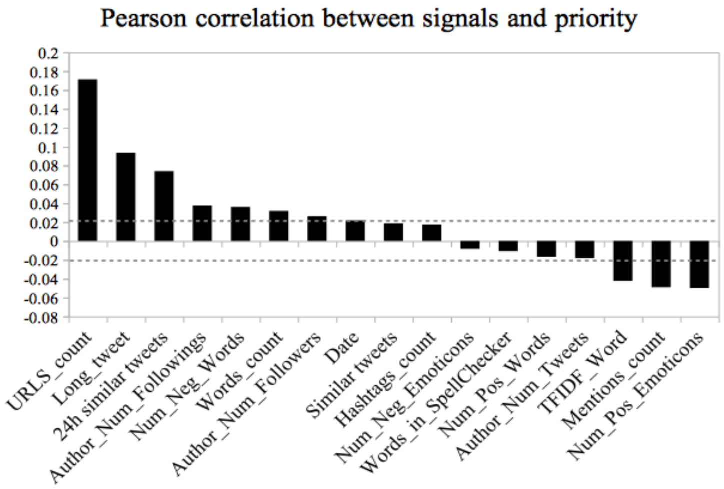

Fig. 3: Pearson correlation between signal values and manual priority

lar tweets produced in a time span of 24 hours), `Author num followers` (number of followers of the author), `Author num followees` (number of people followed by the author), `neg words` (number of words with negative sentiment), `Num pos emoticons` (number of emoticons associated with a positive sentiment), and `Mentions count` (number of Twitter users mentioned). Polarity features have been extracted using three affective lexicons: the General Inquirer (Stone et al., 1966), SentiSense (de Albornoz et al., 2012) and SentiStrength (Mike et al., 2010).

- Each of the selected signals produces a ranking of tweets. We combine them to produce a final ranking using Borda count (Van Erp and Schomaker, 2000), a standard voting scheme to combine rankings.
- We remove redundancy with the same iterative procedure used in the *Followers* baseline.

Learning to Rank (L2R). This contrastive system considers the same initial set of signals of priority and content quality than in the voting approach, but using a Learning to Rank approximation. In summary, a L2R approach makes use of a Machine Learning (ML) algorithm and an optimization function in order to generate several rankings with the aim of maximizing the optimization function. To implement this system we have used the RankLib[2] package. This software offers different ML algorithms and evaluation metrics for optimization. We have evaluated several ML algorithms and finally selected the Random Forest (Breiman, 2001) and the nDCG metric for optimization function due to its similarities with the evaluation of the proposed problem.

[2] https://sourceforge.net/p/lemur/wiki/RankLib/

Similarly, we remove redundancy with the same iterative procedure used in the *Followers* baseline.

Signal Selection and Learning to Rank (SS/L2R). Finally, we aim to test how the combination of both approaches behaves. To this aim, we use the set of signals selected in the SSV procedure, and then feed the L2R method with these signals. Again, the iterative procedure used in the *Followers* baseline is employed to remove redundancy.

An example of a reputation summary is shown in Appendix A.

6 Results and discussion

In this section we present the results of the experiments and discuss such results according to our two research questions. We also describe the results of a further experiment that investigates the role of different priority levels in the generation of the summaries and how can they be included into our evaluation metric.

6.1 RQ1: Is it a new problem? Is it topic priority different from topic centrality?

Our first Research Question was whether producing reputation reports is a standard summarization problem, or a new one, given that the notion of priority seems to play a significant and distinctive role in our reputational scenario.

To answer the question, we have evaluated all baseline and contrastive systems using the RepLab Summarization dataset. Figure 4 compares the results of the LexRank and Followers baselines with the three contrastive systems (SSV, L2R, SS/L2R) in terms of ROUGE-2. Firstly, ROUGE-2 is computed using the manual abstractive summaries as reference summaries (see Section 1).

Results clearly indicate that priority signals play a major role in the task, and therefore producing reputation reports is not a standard, centrality-oriented summarization problem. In terms of ROUGE, the combination of signals (SSV) is consistently better than both LexRank and the Followers baseline at all compression levels, with all differences being statistically significant. Remarkably, even the followers baseline, which uses a rough indication of priority as the only signal for summarization, also performs better than LexRank at all compression levels. This indicates that centrality alone is not adequate to characterize a good reputational summary.

Surprisingly, the direct supervision approach (L2R) does not outperform the followers baseline, and achieves only marginally better results than the LexRank baseline, specially for compression rates of 5% and 10%. This unexpected difference in the results between the SSV and the L2R may reside in the differences between the training and test sets in the RepLab 2013 dataset. Training and test tweets are separated by 6 months, which makes direct supervision challenging. It is possible that a more distant supervised technique

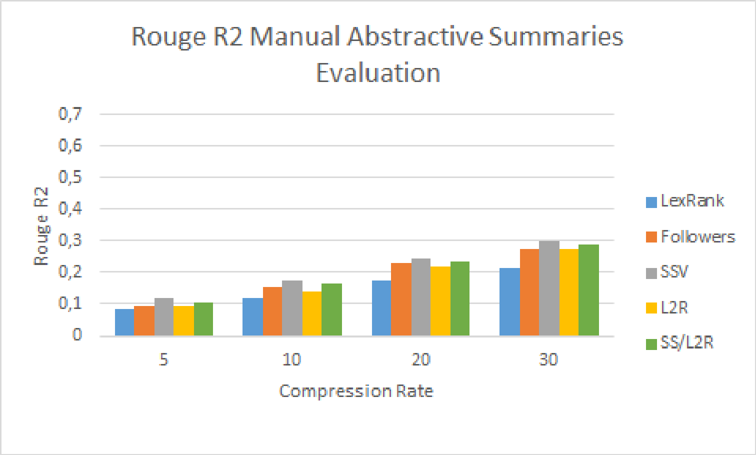

Fig. 4: Evaluation using Rouge-R2 with respect to Manual Abstractive Summaries.

(only signal selection is supervised), produces less over-fitting and obtains better results than a fully supervised approach. This is consistent with the results of the combined approach, SS/L2R, that achieves better results than both baselines, but is slightly lower than SSV.

Let us now do a similar study, but using the extractive manual summaries instead of the abstractive ones. Results are shown in Figure 5. Note that the extractive summaries were generated by manually selecting those tweets that are more relevant for a given topic (see Section 1). The vocabulary overlap of the automatic summaries with the extractive summaries is expected to be higher, and therefore the absolute evaluation figures are higher, in terms of ROUGE-2, than those achieved by comparing with the abstractive summaries (see Figure 4).

Consistently with the results with respect to abstractive summaries, the SSV and SS/L2R approaches clearly outperform the two baselines and the L2R system. Besides, as in the abstractive evaluation, the L2R approach does not improve the Followers baseline, and only outperforms the LexRank baselines for compression rates 20% and 30%, being the results even worse for compression rates 5% and 10%. Results are similar to those achieved by the abstractive evaluation, with the exception that absolute values are higher (as expected comparing to extractive summaries).

6.2 RQ2: Is it possible to apply metrics from search with diversity to the evaluation of automatic summaries?

Our second goal is to test our alternative methodology for evaluating reputational summaries, which leads to our RPB-SUM-R and RBP-SUM-B metrics

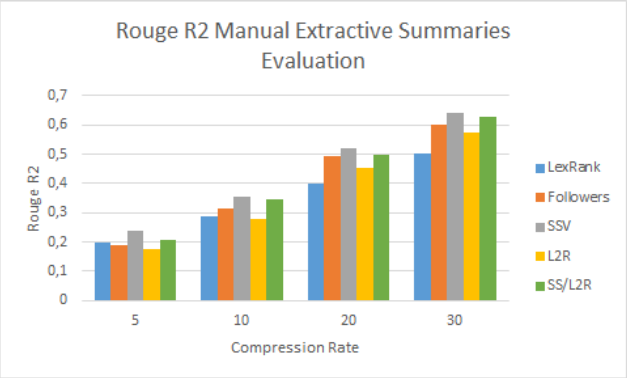

Fig. 5: Evaluation using Rouge-2 and Extractive Reference Summaries.

that use as ground truth a manual annotation of topics and their priorities, instead of actual summaries.

Figure 6 shows the results of RBP-SUM-R, which penalizes redundancy in the ranking with a reciprocal discount function. According to the metric, our three contrastive approaches outperform both baselines for both values of p. Consistently with the evaluation using ROUGE and manual summaries, SSV and SS/L2R are considerable better than the fully supervised L2R approach, and clearly outperform both baselines. The results of the L2R contrastive system are similar to those obtained by the Followers baseline, but still considerably better than the summarization baseline (LexRank).

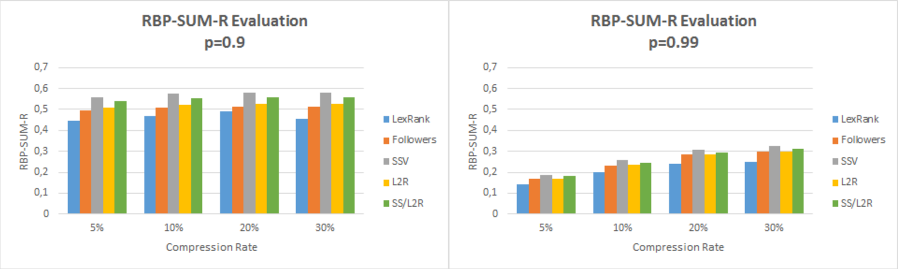

Fig. 6: Evaluation Results using RBP-SUM-R (reciprocal discount)

Figure 7 shows the results when considering redundancy with a binary discount function, RBP-SUM-B. The results are similar than those achieved by RBP-SUM-R, but with lower absolute values as the scoring function is stricter. As in the previous experiments, the SSV and SS/L2R approaches clearly outperform the two baselines and the L2R approaches. L2R achieves similar re-

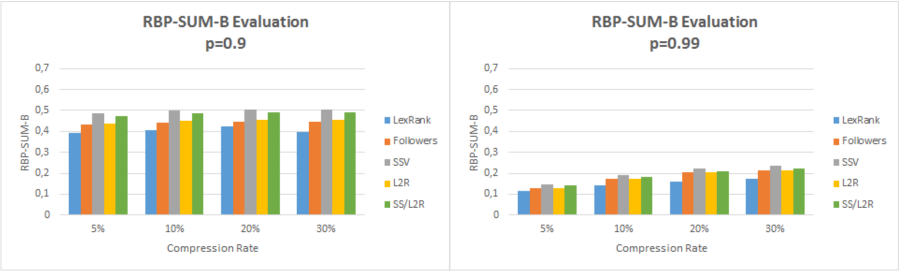

Fig. 7: Evaluation Results using RBP-SUM-B (binary discount)

sults than the Follower baseline, and substantially improves the LexRank baseline.

Overall, evaluation with RBP-SUM is consistent with the results obtained in the standard evaluation using ROUGE. The only difference is that this evaluation methodology, which penalizes redundancy more heavily (tweets from the same topic receive an explicit penalty), gives the followers baseline a higher score than LexRank at all compression levels (with both relevance scoring functions).

Relative differences are rather stable between both p values and between both relevance scoring functions. Naturally, absolute values are lower for RBP-SUM-B, as the scoring function is stricter. Although experimentation with users would be needed to appropriately set the most adequate p value and relevance scoring schema, evaluation results seem to be rather stable with respect to both choices.

6.3 Weighting different priority levels: detecting alerts vs mildly relevant information

As stated in the previous experiment, the reputational priority of analyzed conversations plays an important role in the generation and analysis of the final summary. In order to investigate the role of priority levels and how can they be considered in our evaluation metric, we have modified the parameter r_i of the RBP-SUM measures (see Section 4) to assign different weights to each priority level. In particular, we give a higher weight to reputation *Alerts* than to *Mildly Important* topics:

$$r_i = \frac{Pr_{lv}}{|\{k \in \{1 \ldots i-1\} | \text{topic}(i) = \text{topic}(k)\}|} \tag{4}$$

Where Pr_{lv} is 2 when the tweet belongs to an *Alert* topic and 1 when the tweet belongs to a mildly important topic. Regarding the equation 3 for the RBP-SUM-B metric, the priority levels are introduced as follows:

$$r_i = \begin{cases} 2 & \text{if } \forall k \in \{1..i-1\}\text{topic}(i) \neq \text{topic}(k) \wedge \text{topic}(i) \in \textit{Alerts} \\ 1 & \text{if } \forall k \in \{1..i-1\}\text{topic}(i) \neq \text{topic}(k) \wedge \text{topic}(i) \in \textit{Mildly Important} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

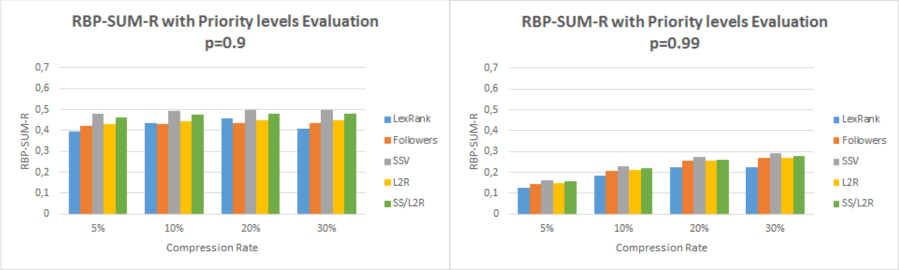

Fig. 8: Evaluation Results using RBP-SUM-R with Priority levels

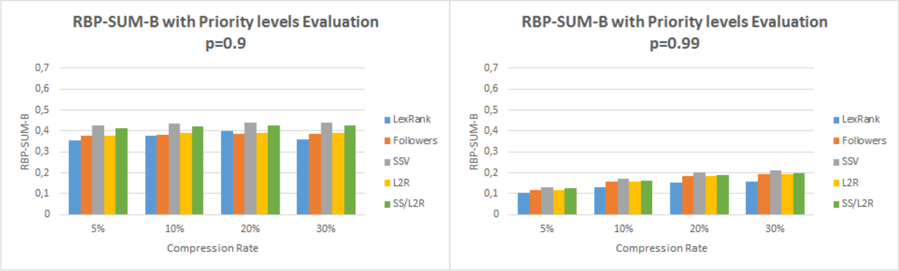

Fig. 9: Evaluation results using RBP-SUM-B with Priority levels

As can be seen in Figures 8 and 9, results are similar to those obtained without using different weights for alerts vs important topics. The main difference observed is with respect to the LexRank baseline, which now achieves better results compared with the Followers baseline than in the previous results. This indicates that centrality signals are better at identifying alerts than mildly important issues. In other words, alerts seem to have not only reputational priority, but also topic centrality.

Apart from that difference, the best results are once again achieved by the contrastive system SSV, followed by the combined approach SS/L2R.

7 Conclusions

In this paper we have hypothesized that the problem of generating reputation reports, which is both practical and challenging from a research perspective, is different from the general summarization task, because the notion of reputational priority is different from the traditional notion of importance or centrality. Our experiments support the hypothesis: even a naive baseline based on reputational priority, which simply takes the tweets written by the most followed authors and removes redundance, is significantly better than a standard summarization system based on centrality (LexRank). Our contrastive systems explore several priority and centrality signals, and they all outperform both LexRank and the followers baseline.

We have also seen that purely supervised methods (exemplified by a Learning to Rank algorithm) have difficulties, partly because reputation topics (at least in Twitter) evolve very quickly, and training on current data may be misleading for data to come. A distant supervision approach, that only uses training data to select valuable signals, performs much better than our fully supervised strategy. Another contributing factor may be that reputation alerts are typically new, unexpected issues, which are challenging to detect with purely supervised methods.

Once we know that producing reputational reports is a differentiated problem, an immediate question is how best to evaluate systems for this task. We have compared the most popular summarization evaluation method (ROUGE-like similarity measures that compare system outputs with reference manuals summaries) with a novel evaluation methodology that establishes an analogy with the problem of search with diversity, and adapts an IR evaluation metric to the task. We have called this new metric *RBP-SUM*. The results of the global, quantitative evaluation is highly correlated with using ROUGE with respect to manually produced reports. Overall, we advocate the use of RBP-SUM to evaluate reputation reports because of its many advantages: (i) it avoids the need of explicitly creating reference summaries, which is a costly process; the annotation of topics and priorities is sufficient; (ii) It is derived from a user model, and it allows an explicit modeling of the patience of the user when reading the summary, and of the relative contribution of information nuggets depending on where in the summary they appear and their degree of redundancy with respect to already seen text; (iii) It is interpretable: it is possible to analyze where the algorithm is failing (redundancy, failure to identify important topics, etc.) in order to make improvements; and (iv) it does not require systems to generate summaries at different levels of compression. The only clear limitation of RBP-SUM is that it can only be applied to extractive summarization techniques, as it operates on a rank of tweets (or sentences). Another minor limitation is that it requires to set the parameter p for each experimental setting.

Finally, other results from our experiments indicate that evaluation with respect to abstractive and extractive summaries both gives similar results. This is an interesting finding, since it allows to build abstractive and extractive summaries depending on what is more efficient depending on the actual scenario.

The generation of reputation reports, although critical in the field of Public Relations in general and Online Reputation Management in particular, is still not well understood from a research perspective. Ultimately, the main contribution of this paper is probably the test collection we have built, which comprises extractive and abstractive summaries in two languages for a large number of companies in two domains (banking and automobile). Our plan is to openly release the collection for research purposes, and we expect it to enable further experimentation on the topic.

Acknowledgements This research was partially supported by the Spanish Ministry of Science and Innovation (Vemodalen Project, TIN2015-71785-R) and UNED (project 2014V/-PUNED/001).

References

Alsaedi, N., Burnap, P., and Rana, O. F. (2016). Automatic summarization of real world events using twitter.

Amigó, E., De Albornoz, J. C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., De Rijke, M., and Spina, D. (2013a). Overview of replab 2013: Evaluating online reputation monitoring systems. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 333–352. Springer.

Amigó, E., Gonzalo, J., and Verdejo, F. (2013b). A general evaluation measure for document organization tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 643–652. ACM.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chakrabarti, D. and Punera, K. (2011). Event summarization using tweets. *ICWSM*, 11:66–73.

Cho, S. G. and Kim, S. B. (2015). Summarization of documents by finding key sentences based on social network analysis. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 285–292. Springer.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM.

Cossu, J.-V., Bigot, B., Bonnefoy, L., and Senay, G. (2014). Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on twitter. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 154–159. Springer.

de Albornoz, J. C., Plaza, L., and Gervás, P. (2012). Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *LREC*, pages 3562–3567.

De Maio, C., Fenza, G., Loia, V., and Parente, M. (2016). Time aware knowledge extraction for microblog summarization on twitter. *Information Fusion*, 28:60–74.

Duan, Y., Chen, Z., Wei, F., Zhou, M., and Shum, H.-Y. (2012). Twitter topic summarization by ranking tweets using social influence and content quality. *Proceedings of COLING 2012*, pages 763–780.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., and Rindflesch, T. C. (2009). Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of biomedical informatics*, 42(5):801–813.

He, R., Liu, Y., Yu, G., Tang, J., Hu, Q., and Dang, J. (2017). Twitter summarization with social-temporal context. *World Wide Web*, 20(2):267–290.

Inouye, D. and Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Litvak, M. and Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 17–24. Association for Computational Linguistics.

Litvak, M., Last, M., and Kandel, A. (2013). Degext: a language-independent keyphrase extractor. *Journal of Ambient Intelligence and Humanized Computing*, 4(3):377–387.

Litvak, M. and Vanetik, N. (2017). Query-based summarization using mdl principle. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 22–31.

Liu, X., Li, Y., Wei, F., and Zhou, M. (2012). Graph-based multi-tweet summarization using social signals. *Proceedings of COLING 2012*, pages 1699–1714.

Louis, A. and Newman, T. (2012). Summarization of business-related tweets: A concept-based approach. *Proceedings of COLING 2012: Posters*, pages 765–774.

Marujo, L., Ribeiro, R., de Matos, D. M., Neto, J. P., Gershman, A., and Carbonell, J. (2015). Extending a single-document summarizer to multi-document: a hierarchical approach. *arXiv preprint arXiv:1507.02907*.

Meena, Y. K. and Gopalani, D. (2015). Feature priority based sentence filtering method for extractive automatic text summarization. *Procedia Computer Science*, 48:728–734.

Mei, Q., Guo, J., and Radev, D. (2010). Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018. Acm.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Mike, T., Kevan, B., Georgios, P., and Di, C. (2010). Sentiment in short strength detection informal text. *JASIST*, 61(12):2544–2558.

Moffat, A. and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2.

Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 763–772. Association for Computational Linguistics.

Nguyen-Hoang, T.-A., Nguyen, K., and Tran, Q.-V. (2012). Tsgvi: a graph-based summarization system for vietnamese documents. *Journal of Ambient Intelligence and Humanized Computing*, 3(4):305–313.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Plaza, L. and Carrillo-de Albornoz, J. (2013). Evaluating the use of different positional strategies for sentence selection in biomedical literature summarization. *BMC bioinformatics*, 14(1):71.

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., et al. (2004). Mead-a platform for multidocument multilingual text summarization.

Sarkar, K., Saraf, K., and Ghosh, A. (2015). Improving graph based multidocument text summarization using an enhanced sentence similarity measure. In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pages 359–365. IEEE.

Sharifi, B., Hutton, M.-A., and Kalita, J. (2010). Summarizing microblogs automatically. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 685–688. Association for Computational Linguistics.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Takamura, H., Yokono, H., and Okumura, M. (2011). Summarizing a document stream. In *European conference on information retrieval*, pages 177–188. Springer.

Van Erp, M. and Schomaker, L. (2000). Variants of the borda count method for combining ranked classifier hypotheses. In *IN THE SEVENTH INTERNATIONAL WORKSHOP ON FRONTIERS IN HANDWRIT-*

ING RECOGNITION. 2000. AMSTERDAM LEARNING METHODOLOGY INSPIRED BY HUMANS INTELLIGENCE BO ZHANG, DAYONG DING, AND LING ZHANG. Citeseer.

Wu, H., Gu, Y., Sun, S., and Gu, X. (2016). Aspect-based opinion summarization with convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 3157–3163. IEEE.

Zhang, H., Fiszman, M., Shin, D., Miller, C. M., Rosemblat, G., and Rindflesch, T. C. (2011). Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of biomedical informatics*, 44(5):830–838.

Zhuang, H., Rahman, R., Hu, X., Guo, T., Hui, P., and Aberer, K. (2016). Data summarization with social contexts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 397–406. ACM.

Appendix A Example of a Reputation Summary

```
"Santander may sell U.S. car finance arm to raise cash http://bit.
   ly/WCi6Za "
"Santander planea absorber Banesto http://www.telecinco.es/
   informativos/economia/Santander-absorber-Banesto-CNMV-
   cotizacion\_0\_1526175033.html    "
"Sernac ofició al Banco Santander por nueva falla http://bit.ly/
   RfDthz "
"Inditex, Mercadona y Santander lideran el ranking de mejores
   empresas para trabajar en España #empleo #trabajo http://ow.ly
   /fo7Rh "
"Elmo: 6 de diciembre - 5.00 Santander aumenta las alarmas sobre
   Salfacorp: duda que pueda cumplir sus compromisos de http://
   goo.gl/bwn65 "
"Santander cerrará 700 oficinas tras la integración de las filiaes
    Banesto y Banif. http://bit.ly/U6ZCy7  #economia #finanzas #
   bolsa #forex"
"Banco Santander despide a 1.200 empleados de Brasil por el
   pinchazo ... http://bit.ly/Unyo3l "
"El #SERNAC pidió antecedentes al Banco Santander por nuevo fallo
   en sus sistemas http://ow.ly/fViPT "
" Financieros?: compras en CaixaBank y Santander, ventas en
   Mapfre y Popular http://bit.ly/Tx780W  #finanzas #economia"
"Concurso FotoTalentos 13 Fundación Banco Santander y Universia
   http://ow.ly/g1YEA "
"Santander y la burbuja: ""Algunas comunas de Santiago presentan
   alzas que no ... - Diario inmobilia... http://bit.ly/XjLdAI  #
   inmobiliaria"
"Negative outlook for Santander UK says S\&P: Santander UK has
   been taken off CreditWatch negative by Standard and... http://
   bit.ly/T5kdUT "
"Ingresa unos 11,9 millones Emilio Botín vuelve a cobrar todo el
   dividendo de Santander en efectivo http://www.cincodias.com/ "
"Anuncia Banco Santander en España cierre de 700 sucursales http
   ://mile.io/YbODpB "
```

```
"Santander plans to invest in Spain's bad bank http://dlvr.it/2
    VSJ4K  #forex"
"Santander y Aegon se alían para potenciar el negocio de
    bancaseguros en España | http://Diarioelaguijon.com  http://
    www.diarioelaguijon.com/noticia/12280/ECONOMIA-Y-EMPRESAS/
    Santander-y-Aegon-se-alian-para-potenciar-el-negocio-de-
    bancaseguros-en-Espana.html    "
"Segunda convocatoria del programa Becas Santander. http://buzz.mw
    /-SJp\_y "
"Get a Car - Enter your zip code to find dealers near you that
    offer financing with one of Santander programs. http://bit.ly/
    pZGfh0 "
"Santander considers absorbing Banesto - http://FT.com  -
    Financial Times http://tinyurl.com/d2ked9s "
"El Santander cerrará 700 sucursales al integrar Banesto en su
    estructura http://ow.ly/g9V8J  Banesto se dispara en Bolsa"
"VIDEO Un grupo de jóvenes arremete contra una sucursal del
    Santander y revienta el escaparate con una valla http://www.
    youtube.com/watch?v=x2QevygFits     #14N"
"Conveyancing Top solicitor pulls off Santander mortgage fraud -
    Bridging and Commerical: Top solicitor pulls off... http://bit
    .ly/W8WfYV "
"#Colombia Santander tiene un programa de tecnología para mujeres
    empresarias http://bit.ly/Wqorr5 "
"Santander says to close 700 bank branches after Banesto buyout:
    MADRID, Dec 17 (Reuters) - Spain's largest bank ... http://bit
    .ly/SDNW76 "
"#Spain's #Santander studying how to absorb #Banesto: http://bit.
    ly/Zci9x1  | #MADRID #Banco"
"Mirad gráfico al final del post y entenderéis como uno puede
    convertirse en banquero casi gratis #Santander # Banesto http
    ://www.gurusblog.com/archives/banco-santander-absorber-banesto
    /17/12/2012/    "
"La absorción de Banesto por parte del Santander pone fin a 110 añ
    os de historia de la entidad: http://www.telecinco.es/
    informativos/economia/absorcion-Banesto-Santander-historia-
    entidad\_0\_1526175166.html    "
"Santander México es reconocido como Banco del Año - http://bit.ly
    /SsTE9p "
"Santander invertirá 660 millones y Caixabank, 470 millones en la
    primera fase del banco malo: Santander y CaixaB... http://bit.
    ly/Scq4Hp "
```