Sentiment Propagation for Predicting Reputation Polarity

Anastasia Giachanou[1], Julio Gonzalo[2], Ida Mele[1], and Fabio Crestani[1]

[1] Faculty of Informatics, Università della Svizzera italiana (USI), Switzerland
[2] UNED NLP & IR Group, Madrid, Spain
{anastasia.giachanou,ida.mele,fabio.crestani}@usi.ch
julio.gonzalo@lsi.uned.es

Abstract. One of the core tasks of Online Reputation Monitoring is to determine whether a text mentioning the entity of interest has positive or negative implications for its reputation. A challenging aspect of the task is that many texts are polar facts, i.e. they do not convey sentiment but they do have reputational implications (e.g. *A Samsung smartphone exploded during flight* has negative implications for the reputation of Samsung). In this paper we explore the hypothesis that, in order to determine the reputation polarity of factual information, we can propagate sentiment from sentiment-bearing texts to factual texts that discuss the same issue. We test two approaches that implement such hypothesis: the first one is to directly propagate sentiment to similar texts, and the second one is to augment the polarity lexicon. Our results (i) confirm our propagation hypothesis, with improvements of up to 43% in weakly supervised settings and up to 59% with fully supervised methods; and (ii) indicate that building domain-specific polarity lexicons is a cost-effective strategy.

Keywords: Reputation Polarity, Sentiment Propagation

1 Introduction

One of the core tasks in online reputation management is to monitor what is posted online about an entity (a company, celebrity, etc.) and react in case there is an alert of a possible damage on the entity's reputation. Analysts have first to filter the stream of data and find the content that is relevant for the entity of interest. Then, they have to determine if a relevant post is likely to have positive, neutral or negative implications on the entity's reputation.

Reputation polarity is not a trivial task, and it is much more challenging than sentiment analysis. A key problem is that there is a significant amount of tweets with positive or negative reputation polarity which do not explicitly express a sentiment. These tweets are known as *polar facts*. For example, the tweet *Chrysler recalls 919,000 Jeeps to fix air bags* does not convey any sentiment but it has negative impact on the reputation of *Chrysler*.

To address this challenge, we hypothesize that tweets that are about a specific event should tend to have the same reputation polarity. In this way, if there are

many tweets about a specific event, then some of those tweets will explicitly express some sentiment towards the event. Table 1 shows some example tweets relevant to the entity *HSBC* and which are about the same topic. Table 1 also shows the actual (manually annotated) reputation polarity of each tweet, and the sentiment polarity as assigned by a state-of-the-art lexicon based approach. Note that there are some tweets (i.e. *t3* for the topic *accusations*) that do not contain any sentiment word (*sentiment by lexicon* is *neutral*) but they have a negative impact on the entity's reputation, whereas other tweets in the same topic (i.e. *t1, t2*) have an explicit sentiment indicator. Propagation of sentiment across texts discussing the same issue might then be a way of annotating reputation polarity.

We consider two ways of propagating sentiment to sentiment-neutral texts: (i) direct propagation to texts with similar content; (ii) augmenting the lexicon with terms that indicate reputation polarity even if they do not convey sentiment polarity. Hence, we focus on two related research questions:

- *Can we use training material to detect terms with reputational polarity and use them to augment a general sentiment lexicon?* One of the state-of-art approaches in sentiment analysis is the lexicon based approach. However, the general lexicons are not effective for reputation polarity. Hence, we propose to augment general lexicons at different levels of granularity with terms extracted from training data to build reputation lexicons. An associated question is *what is the right level of generalization for a reputation lexicon.* We will explore three alternatives: (i) building a general purpose lexicon with all available training material; (ii) building domain-specific lexicons with training material for entities in a given domain (e.g. banking, automotive); (iii) building entity-specific lexicons with separated training material for each entity. In principle, the more specific a lexicon is, the most accurate results will give, but at a substantial cost, because we need more training examples. We want to investigate whether there is an optimal level of specificity that provides competitive results at a moderate cost.
- *Can we propagate sentiment to texts that are similar in terms of content to improve reputation polarity?* In order to answer this question we will consider two propagation alternatives: (i) first perform text clustering to detect topics, and then propagate sentiment within each topic; (ii) directly propagate sentiment from a sentiment-bearing text to other texts that are pairwise similar. In addition, we will also experiment with the use of a polar fact filter to avoid overpropagation to polarity-wise neutral texts.

2 Related Work

Although reputation polarity is substantial different to sentiment analysis, the two tasks have some similarities. To this end, past work on reputation polarity evolved from sentiment analysis. Previous work on opinion retrieval and sentiment analysis can be roughly divided into two categories: lexicon based and classification based approaches. The lexicon based approaches estimate the sentiment of a document using a list of opinion words [25, 24] known as opinion

Table 1. Examples of annotated tweets in the RepLab 2013 training dataset.

Oracle Topic	id	Tweet	Reputation Polarity	Sentiment by Lexicon
accusations	t1	When I wake up I want to find these trending: Barclays, HSBC, executive arrests, fraud & Tory party. NOT Justin Bieber	Negative	Negative
accusations	t2	THE CORPORATE POLITICIANS: 20 years of failure for Britain as they skimmed the system. #cnn, #times, #cnbc, #hsbc	Negative	Negative
accusations	t3	@PoliticalPryers he's ceo of one of the banks involved . He high but not the top! By this time next week RBS, Llyods, HSBC will get same	Negative	Neutral

lexicons. The presence of any opinionated word in a document is an indicator of sentiment. In its most typical scenario, lexicon based approach is unsupervised since it does not require any training data. More sophisticated approaches incorporate additional sentiment indicators such as proximity between query and opinion terms [7] or topic-based stylistic variations [9].

The classification based approaches use sets of features to build a classifier that can predict the sentiment polarity of a document [19]. The features range from simple n-grams to semantic features and from syntactic to medium's specific features. A number of researchers analyzed the impact of different features on Twitter sentiment analysis and established feature selection criteria [17, 1, 13]. The classification based approaches can be further divided into semi-supervised and supervised approaches. The major difference between the two categories is that the semi-supervised approaches combine labeled and unlabeled data. A comprehensive review on opinion retrieval and sentiment analysis can be found in a survey by Pang and Lee [18] whereas a comprehensive survey focused on Twitter sentiment analysis can be found by Giachanou et al. [8].

A number of proposed approaches for reputation polarity treated the task with methods similar to sentiment analysis' methods. Classifiers trained on sentiment and textual features showed to be very effective on RepLab evaluation campaign [3, 2]. The best result on RepLab 2013 was achieved by Hangya and Farkas [10] who trained a Maximum Entropy classifier using sentiment lexicon, bigrams, number of negation words and character repetitions. Castellanos et al. [4] addressed the reputation polarity problem with an information retrieval based approach and found the most relevant class using the tweet's content as a query. Other approaches considered sentiment classifiers and lexicons [15, 22].

Peetz et al. [20] assumed that understanding how the tweet is perceived is an important indicator for estimating the reputation polarity of a tweet. To this end, they proposed a supervised approach that also considered reception features such as tweet's replies and retweets. Their results showed that reception features were effective and their best result was obtained on entity dependent data.

Different form the previous work, we explore the hypothesis that texts that are about the same event should share the same reputation polarity. To this end, we consider propagating sentiment using topically similar tweets. In addition,

we are the first to consider a polar fact filter that is able to differentiate neutral tweets from polar facts.

3 Proposed Approach

Our starting point is a standard lexicon based approach for sentiment analysis. This approach detects the sentiment of a document by using a general list of words annotated with their sentiment polarity (*positive* or *negative*). The presence of any opinionated word in a document indicates the document's polarity. Hence, this approach generates a sentiment score for the document based on the number of opinionated terms it contains.

Let $polarity(d)$ be the reputation polarity of a document d, where $polarity(d)$ can take one of the values $\{-1, 0, 1\}$ referring to a positive, neutral and negative polarity respectively. Also, let S_d denote the sentiment score of a document d based on the sentiment scores of its terms, calculated as: $S_d = \sum_{t \in d} opinion(t)$, where $opinion(t)$ is the opinion score of the term based on an opinion lexicon. Then, according to the lexicon based approach the reputation polarity of a document is determined as follows:

$$polarity(d) = \begin{cases} 1, & \text{if } S_d > 0 \\ -1, & \text{if } S_d < 0 \\ 0, & otherwise \end{cases}$$

Here we should note that the sentiment score S_d depends on the number of opinionated words that appear in the document and for this reason the score is an integer value. One of the advantages of this method is that it does not require any training data. We use this method as our baseline.

In this paper we use the lexicon based approach as a starting point to find the sentiment of tweets and then we explore two different approaches to improve the reputation polarity. First, we extract terms that are closely related to positive or negative sentiment and use these words to augment a sentiment lexicon. Second, we propagate sentiment to factual tweets to determine their reputation polarity using the sentiment of tweets that are similar in terms of content.

3.1 Lexicon Expansion

One limitation of the lexicon based approaches is the word mismatch between the tweet and the general opinion lexicons. Tweets contain a lot of idiomatic words as with the case of the "elongated" words (e.g. *goooooood*). This problem is more evident for the reputation polarity task where there are a lot of tweets that do not contain any sentiment word but have an impact on the entity's reputation.

To address the problem of the word mismatch, we explore the effectiveness of lexicon augmentation. To learn new positive/negative words we use the training data provided in the collection. The positive/negative lexicons are expanded

with the terms of the positive/negative tweets of the training set. We augment the lexicons on three different levels of granularity: *domain/entity independent*[3], *domain dependent* and *entity dependent*. After augmenting the lexicons, we use the lexicon based approach that uses the number of occurrences of opinionated terms to predict the reputation polarity of a document. This approach that we refer to it as *simple lexicon augmentation* considers only the presence of words as an indicator of reputation polarity.

In addition, we also investigate a fully supervised way to learn the words that indicate reputation polarity. This approach is based on the Pointwise Mutual Information (PMI) method originally proposed by Church and Hanks [6]. According to this approach, every term t is assigned a PMI score for each of the three reputation polarity classes: positive, neutral and negative. The sentiment score for a term t is calculated using the training data as follows:

$$PMI(d, positive) = \sum_{t \in d} PMI(t, positive)$$

$$PMI(t, positive) = \log_2 \frac{c(t, positive) * N}{c(t) * c(positive)}$$

where $c(t, positive)$ is the frequency of the term t in the positive tweets, N is the total number of words in the corpus, $c(t)$ is the frequency of the term in the corpus and $c(positive)$ is the number of terms in the positive tweets. The PMI of the terms for the negative and neutral classes is calculated in a similar way. Then these scores can be used to predict the polarity of the test documents. We assume that the polarity of a document is the one with the highest PMI score.

3.2 Polar Fact Filter

A limitation of propagation methods is that they may overestimate the number of tweets with reputation polarity (i.e. the sentiment polarity is potentially propagated to polar facts and to reputation-neutral tweets). A possible supervised solution is to first detect polar facts, building a classifier (*polar fact filter*) that takes a single tweet as an input and decides if the tweet is a polar fact or not. To this end, we address the task of identifying polar facts as a binary classification problem and do not differentiate between positive and negative tweets. We train a linear kernel Support Vector Machine (SVM) classifier to discriminate between polar facts and neutral tweets. SVM [5] is a state-of-art learning algorithm that has been effectively applied on text categorization tasks.

First, we separate the polar facts and the neutral tweets into two classes, $y_i \in \{-1, 1\}$, where N is the number of the labeled training data. The training examples are $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N), \mathbf{x} \in R^k$ where k is the number of features.

For the classification, we explored a number of different features that have proved to be effective for sentiment classification [12]. The features can be grouped in three classes as follows:

[3] In the rest of the paper we refer to this setting as *independent* for brevity

- n-grams: n-grams with $n \in [1, 4]$, character grams
- stylistic: number of capitalised words, number of elongated words, number of emoticons, number of exclamation and question marks
- lexicons: manual and automatic lexicons

We explore the effectiveness of the polar fact filter on three different training settings: *independent*, *domain dependent* and *entity dependent*.

3.3 Sentiment Propagation

As already mentioned, we assume that similar tweets in terms of content (topic) should tend to have the same polarity for reputation. Hence, we propose to propagate sentiment to tweets that are annotated as polar facts using the sentiment of similar tweets. We explore two different propagation approaches: *clustering* and *tweet to tweet similarity*. Also, we explore two different ways to propagate sentiment. The first method is based on the *maximum* sentiment of the similar tweets whereas the second is based on *tweet's similarity* to each of the reputation polarity classes.

To better describe our approach we introduce some notation. Let $D = \{d_1, ..., d_M\}$ be some tweets we want to predict their reputation polarity using a set of other tweets $D' = \{d'_1, ..., d'_N\}$ for which we already know their polarity. Also, let $D^+ = \{d_1^+, d_2^+, ..., d_K^+\}$, $D^\bullet = \{d_1{}^\bullet, d_2{}^\bullet, ..., d_V{}^\bullet\}$ and $D^- = \{d_1^-, d_2^-, ..., d_L^-\}$ be three different sets of tweets that are annotated as positive, neutral and negative respectively and $D' = D^+ \cup D^\bullet \cup D^-$.

To annotate a tweet d that belongs to D we count the number of tweets in D' that belong to each of the reputation polarity classes *positive*, *neutral* and *negative* denoted as $|D^+|$, $|D^\bullet|$ and $|D^-|$ respectively. The polarity of a document d is calculated as follows:

$$polarity(d) = \begin{cases} 1, & \text{if } |D^+| = max\{freq(d)\} \\ -1, & \text{if } |D^-| = max\{freq(d)\} \\ 0, & otherwise \end{cases}$$

where $max\{freq(d)\} = \max |D^+|, |D^\bullet|, |D^-|$. Here we should note that we propose to use the polar fact filter to differentiate between the tweets in D and in $\mathrm{D}^\bullet$ and that $D \cap D' = \emptyset$.

The second approach to propagate sentiment is based on the tweet's similarity to each of the polarity classes. To annotate a tweet d that belongs to D, we first calculate the similarity to each of the three classes. For the positive class we calculate the similarity as follows:

$$sim^+(d) = \sum_{d_i \in D^+} sim(d, d_i^+)$$

The next step is to calculate the average similarity to the positive class as $avgSim^+(d) = sim^+(d)/|D^+|$ where $|D^+|$ is the number of positive tweets. We follow a similar way to calculate the similarities and the average similarity of the

neutral and negative classes. Next, we calculate the maximum average among the three classes as

$$max\{avgSim(d)\} = \max avgSim^{+}(d), avgSim^{\bullet}(d), avgSim^{-}(d)$$

and finally we determine the polarity of the tweet d as:

$$polarity(d) = \begin{cases} 1, & \text{if } avgSim^{+}(d) = max\{avgSim(d)\} \\ -1, & \text{if } avgSim^{-}(d) = max\{avgSim(d)\} \\ 0, & otherwise \end{cases}$$

To determine D' (the set of tweets for which we already know the sentiment), we explore two different approaches: clustering and tweets' similarity. For clustering the tweets we used the approach that obtained the best result in Spina et al. [23]. This approach first trains a classifier to predict if two tweets belong to the same topic using term, semantic, metadata and temporal features and then uses a hierarchical agglomerative clustering algorithm to identify the clusters. The tweets' clusters are publicly available[4]. For the tweet to tweet similarity, we consider cosine similarity over a bag of terms representation.

4 Experimental Setup

Dataset. For this study, we use the RepLab 2013 [2] data set, which is the largest available test collection for the task of monitoring the reputation of entities (companies, organizations, celebrities, etc.) on Twitter. The RepLab 2013 collection contains 142,527 manually annotated tweets in English and Spanish. The tweets are about 61 different entities that belong to 4 domains: *automotive*, *banking*, *universities* and *music*.

Experimental Settings. We use publicly available word lexicons in English [16] and in Spanish [21] to identify the words that indicate positive or negative sentiment. We use information from tweets' metadata to identify the language of the tweet. We use the same tokenizer for English and Spanish tweets. For the results that are reported we considered the tweets that are relevant to an entity (tweets manually annotated as *related*) from the test set.

Polar Fact Filter. To build the polar fact filter we use a linear SVM classifier. As training data, we use the tweets in the training set which are annotated as neutral by the simple lexicon based approach. We explore a wide range of features such as n-grams, character grams, number of capitalised words, number of elongated words, number of emoticons, number of exclamation and question marks, automatic and manual lexicons. With respect to the lexicons explored for the polar fact filter, we consider Liu's lexicon [11], NRC emotion lexicon [14],

[4] https://github.com/damiano/learning-similarity-functions-ORM

MPQA lexicon [26] and Hashtag Sentiment Lexicon [12]. We explore three different levels of granularity for training the classifier: *independent*, *domain dependent* and *entity dependent*.

Evaluation. We present evaluation scores for our methods on all the three polarity classes, *positive*, *neutral* and *negative*, according to the instructions given at RepLab 2013. We report *F-score* for the proposed methods and the polar fact classification. We use the McNemar test to evaluate the statistical significance of differences, which is more appropriate for comparisons of nominal data.

5 Results and Discussion

In this section, we present the results of our proposed methodology on the reputation polarity task. First, we discuss the effectiveness of augmenting the lexicon at different levels of granularity, we continue with the performance of the polar fact filter and finally we present the results of sentiment propagation.

5.1 Lexicon Expansion

In order to address the first research question, we compare the results of augmenting the lexicon at different levels of granularity with the lexicon based approach (*baseline*). Results are displayed in Table 2. The main outcome is that augmenting the lexicon is effective at all levels of granularity, with improvements ranging from +17% in the general expansion to +25% if a specific lexicon is created for each individual entity. All improvements are statistically significant with respect to the baseline. Unsurprisingly, entity-specific lexicons give the best result, but note that the difference between domain and entity specific lexicons is thin (only 1%). This is an interesting observation, because it indicates that training data can be generalized for entities within a domain, and that is more cost-effective than having to annotate training data for every entity in a domain.

Table 2. Performance results of the lexicon based approach before and after augmenting the lexicon using independent, domain dependent and entity dependent data. A star(∗) indicates statistically significant improvement over the lexicon based approach.

Method	F-measure
Lexicon Based	0.368
Lexicon Augmentation - Independent	0.431∗ (+17%)
Lexicon Augmentation - Domain Dependent	0.455∗ (+24%)
Lexicon Augmentation - Entity Dependent	**0.460∗ (+25%)**

Alternatively, we also explore the effectiveness of PMI for predicting the reputation polarity. Similar to the simple lexicon augmentation approach, we use three different settings to learn the PMI scores: *independent* referring to

all the training data, *domain dependent* referring to the setting where we learn PMI scores for each domain and *entity dependent* where we learn PMI scores for each entity. Table 3 displays the results. The conclusions are the same as for the previous method (the expansion substantially improves performance, entity-dependent expansion is the best but domain-dependent expansion is very close). The general performance of this method (which is fully supervised) is superior, and in fact entity-dependent PMI results are 5.6% better than the best results published to date on this dataset [20].

Table 3. Performance results of the supervised method based on PMI, when trained on independent, domain dependent and entity dependent data. A star(∗) indicates statistically significant improvement over the lexicon based approach.

	F-score
Lexicon Based	0.368
PMI - Independent	0.547∗ (+49%)
PMI - Domain Dependent	0.572∗ (+55%)
PMI - Entity Dependent	**0.586∗ (+59%)**

5.2 Polar Fact Filter

Table 4 presents the effectiveness of the polar fact filter when it is trained on different set of features and when it is trained on an *independent*, *domain dependent* or *entity dependent* setting. Similarly to the previous reported results, the best performance is obtained when the classifier is trained on the *entity dependent* setting. One interesting observation is that the best performance is obtained when the classifier is trained on *n-grams* and *character* grams using entity dependent data. This result was expected since this classifier aims to differentiate between polar fact tweets and neutral tweets and neither of them contain sentiment words.

However, the results indicate that sentiment lexicons are effective features for the polar fact filter when we use independent and domain dependent data. Note that for the polar fact filter we used 4 different lexicons that have been found to be effective for sentiment analysis [12] and which contain more information compared to the general lexicons. The results indicate that in case of independent and domain dependent data, sentiment lexicons can still provide useful information for reputation polarity. The model with the best performance (trained on *n-grams, character grams/entity-dependent*) is used in the rest of the experiments to detect the tweets that are polar facts and that have to be annotated with reputation polarity.

5.3 Sentiment Propagation

For the second research question, we explore the effectiveness of propagating sentiment with the aim to improve reputation polarity. We compare the results

Table 4. Performance results (F-measure) of the polar fact filter classification when trained on independent, domain dependent and entity dependent data.

	Independent	Domain Dependent	Entity Dependent
n-grams	0.633	0.654	**0.692**
n-grams, stylistic	0.635	0.655	0.691
n-grams, stylistic, lexicons	**0.654**	**0.660**	0.668

of propagating sentiment using an automatic clustering and a cosine similarity approach. Table 5 presents the results of propagating sentiment to tweets that have been annotated as polar facts. The results indicate that sentiment can be propagated topically to annotate tweets with reputation polarity: in all cases, the improvement is above 20% with respect to the no propagation baseline, and for the best experimental setting (propagating to similar tweets using the max approach), the improvement is +43%. This directly confirms the hypothesis that tweets that share a similar (factual) content tend to share the same reputation polarity.

Table 5. Performance results (F-measure) of sentiment propagation approaches.

	Max	Similarity to Class
No propagation	0.368	0.368
Cluster propagation	0.472 (+28%)	0.457 (+24%)
similar tweets propagation	**0.526 (+43%)**	0.495 (+35%)

Finally, Table 6 compares the best results published until now for reputation polarity on the RepLab 2013 dataset (SVM trained on message and reception features and on an entity-dependent scenario) [20] with our best supervised and weakly-supervised approaches in terms of F-measure. The supervised approach based on PMI outperforms [20] with a 5.6% relative improvement in terms of F-measure (0.586 vs 0.553). This indicates that it is not necessary to use many features to get competitive results in reputation polarity. Unsurprisingly, we also see that fully supervised approaches outperform weakly supervised ones. Our best weakly supervised approach (propagation to similar tweets using max combination), however, is only 5% worse than [20] (0.526 vs 0.553). This small difference indicates that weakly supervised annotation of reputational polarity is feasible, which is a promising result as such methods are less dependent on the availability of training data.

6 Conclusions and Future Work

The results of our experiments strongly support our initial hypothesis: sentiment signals can be used to annotate reputational polarity, starting with sentiment-

Table 6. Comparison with state-of-the-art results.

System	F-Measure
Peetz et al. 2016 (Best published result)	0.553
Supervised - PMI & Entity Dependent	0.586
Weakly Supervised - Propagation (Tweets' similarity & Max)	0.526

bearing texts and propagating sentiment to sentiment-neutral similar texts. We have explored two approaches: augmenting the sentiment lexicon via propagation, and directly propagating sentiment to topically similar tweets.

Augmenting the sentiment lexicon in a weakly-supervised way improves results up to 25% if we generate a specific lexicon for each entity of interest. But, remarkably, generating domain-specific lexicons (which requires less training material) gives very similar results (24% improvement over the original sentiment lexicon). The conclusion is that sentiment lexicons can be augmented to create reputation polarity lexicons, and that the domain level is a cost-effective level of granularity for doing so. If we use a fully supervised approach to learn reputation polarity words (based on PMI scores), performance is 5.6% better than the best published result on the dataset so far [20]. This indicates that learning PMI values to predict reputation polarity is very effective.

Direct propagation of sentiment is also effective. In all conditions, the improvement is above 20% with respect to the no propagation baseline, and for the best setting (propagating to similar tweets using the max approach), the improvement is +43%. This is also a weakly supervised approach, because both the initial sentiment annotation and the propagation are unsupervised; the only supervised mechanism is the polar fact filter that prevents propagation to truly neutral tweets. Results, however, are only 5% worse than [20] (0.526 vs 0.553), which is a fully supervised approach. This small difference indicates that weakly supervised annotation of reputational polarity is feasible, which is a promising result as such methods are less dependent on the availability of training data.

Future work includes carefully analyzing the augmented vocabularies. We need to identify the percentage of erroneous additions, how frequently the new terms are sentiment-bearing terms that were absent from the initial vocabulary simply for lack of coverage, and non sentiment-bearing terms which specifically indicate factual polarity. We also plan to analyze different ways of propagating sentiment, and to explore the effectiveness of additional features (e.g. semantic, temporal) on finding the tweets that can be used for sentiment propagation.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: LSM '11. pp. 30–38. ACL (2011)
2. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: CLEF 2013. pp. 333–352. Springer (2013)

3. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M.: Overview of replab 2012: Evaluating online reputation management systems. In: CLEF 2012 (2012)
4. Castellanos, A., Cigarrán, J., García-Serrano, A.: Modelling techniques for twitter contents: A step beyond classification based approaches. In: CLEF 2013 (2013)
5. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. Transactions on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011)
6. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
7. Gerani, S., Carman, M., Crestani, F.: Aggregation methods for proximity-based opinion retrieval. ACM Transactions on Information Systems (TOIS) 30(4), 1–36 (2012)
8. Giachanou, A., Crestani, F.: Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR) 49(2), 28 (2016)
9. Giachanou, A., Harvey, M., Crestani, F.: Topic-specific stylistic variations for opinion retrieval on twitter. In: ECIR '16. pp. 466–478. Springer (2016)
10. Hangya, V., Farkas, R.: Filtering and polarity detection for reputation management on tweets. In: CLEF 2013 (2013)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD '04. pp. 168–177. ACM (2004)
12. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research 50(1), 723–762 (2014)
13. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: ICWSM '11. pp. 538–541. AAAI Press (2011)
14. Mohammad, S.M., Turney, P.D.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: CAAGET '10. pp. 26–34. ACL (2010)
15. Mosquera, A., Fernández, J., M. Gómez, J., Martínez-Barco, P., Moreda, P.: Dlsi-volvam at replab 2013: Polarity classification on twitter data. In: CLEF 2013 (2013)
16. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis of microblogs. In: ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. pp. 93–98 (2011)
17. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREC '10. pp. 1320–1326. ELRA (2010)
18. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
19. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: EMNLP '02. pp. 79–86. ACL (2002)
20. Peetz, M.H., de Rijke, M., Kaptein, R.: Estimating reputation polarity on microblog posts. Information Processing Management 52(2), 193–216 (2016)
21. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in spanish. In: LREC '12. ELRA (2012)
22. Saias, J.: In search of reputation assessment: Experiences with polarity classification in replab 2013. In: CLEF 2013 (2013)
23. Spina, D., Gonzalo, J., Amigó, E.: Learning similarity functions for topic detection in online reputation monitoring. In: SIGIR '14. pp. 527–536. ACM (2014)
24. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307 (2011)
25. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ACL '02. pp. 417–424. ACL (2002)
26. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT '05. pp. 347–354. ACL (2005)