Using an Emotion-based Model and Sentiment Analysis Techniques to Classify Polarity for Reputation

Jorge Carrillo de Albornoz*, Irina Chugur, and Enrique Amigó

Natural Language Processing and Information Retrieval Group, UNED
Juan del Rosal, 16 (Ciudad Universitaria), 28040 Madrid, Spain
{jcalbornoz,irina,enrique}@lsi.uned.es

Abstract. *Online Reputation Management is a novel and active area in Computational Linguistics. Closely related to opinion mining and sentiment analysis, it incorporates new features to traditional tasks like polarity detection. In this paper, we study the feasibility of applying complex sentiment analysis methods to classifying polarity for reputation. We adapt an existing emotional concept-based system for sentiment analysis to determine polarity of tweets with reputational information about companies. The original system has been extended to work with texts in English and in Spanish, and to include a module for filtering tweets according to their relevance to each company. The resulting UNED system for profiling task participated in the first RepLab campaign. The experimental results prove that sentiment analysis techniques are a good starting point for creating systems for automatic detection of polarity for reputation.*

Keywords: Online Reputation Management, Polarity for Reputation, Sentiment Analysis, Emotions, Word Sense Disambiguation

1 Introduction

Part of eMarketing, Online Reputation Management (ORM) has already become an essential component of corporate communication for public figures and large companies [14]. Being absolutely vital to maintain the good name and preserve the "reputational capital", ORM comprises activities that aim at building, protecting and repairing the image of people, organizations, products, or services.

In order to study a brand image, reputation management consultancies perform two main tasks: *monitoring* and *profiling*. As its name suggests, the former consists in a constant (daily) monitoring of online media, seeking and analysing information related to the entity, with the objective of detecting any topic that might damage its image. In contrast, profiling refers to a single or periodic (e.g., monthly) revision of a company's reputation as it distils from news, opinions and comments expressed in social media or online press. Unlike monitoring, which is

* This research was supported by the European Unions (FP7-ICT-2011-7 - Language technologies - nr 288024 (LiMoSINe).)

essentially a real-time problem, profiling is a static study of opinions and polar facts concerning a certain entity and extracted for a given period. Normally, this information is contrasted with what has been said in the same period of time about the company's potential competitors, and with the opinions about the entity in earlier periods of time. These practical scenarios have been adopted as tasks for RepLab 2012, the first evaluation campaign for ORM systems[1]. In this paper, we will focus exclusively on the profiling task.

Although for reputation analysts profiling implies a complex of subtasks such as identifying the dimension of the entity's activity affected by a given content, detecting opinion targets and determining the type of opinion holder[2], the basis of adequate profiling is undoubtedly an effective named entity disambiguation and detection of polarity for reputation. The system described in the present paper centres precisely on the problem of classifying tweets into related and unrelated with respect to a given company (filtering) and on determining the polarity of tweets based on the emotion concepts they contain.

Some tasks considered in profiling are similar to the research problems in opinion mining and sentiment analysis that comprise *subjectivity detection* [23, 19, 16], *polarity classification* [18, 21, 10, 24] and *intensity classification* [10, 4, 25], among others. Although opinion mining has made significant advances in the last years, most of the work has been focused on products. However, mining and interpreting opinions about companies and individuals generally is a harder and less understood problem, since unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modelling of these entities.

Identifying polarity of a given content, ORM system should assess if it has negative, positive or neutral effect on the company's image. Again, this problem is related to sentiment analysis and opinion mining, but significantly differs from the mainstream research in these areas. First of all, what is analysed are not only opinions, subjective content, but also facts, and more to the point, **polar facts**, i.e. objective information that might have negative or positive implications for the company's reputation. This means that ORM systems have to be able to detect polarity also in non-opinionated texts. On the second place, focus or perspective plays sometimes a decisive role, since the same information may be negative from the point of view of clients and positive from the point of view of investors. We will refer to this complex notion of polarity with the term **polarity for reputation**.

In this paper, we study the feasibility of using complex sentiment analysis approaches to classifying polarity for reputation. To this end, as exposed in Section 2, we adapt an existing emotional concept-based system for sentiment analysis to classify tweets with reputation information about companies. The system is also extended to filter tweets according to their relevance to each

[1] `http://www.limosine-project.eu/events/replab2012`

[2] Both companies' dimensions and the types of opinion holder are standard parameters of RepTrak System `http://www.reputationinstitute.com/thought-leadership/the-reptrak-system`.

company and works with texts both in English and in Spanish. It is worthy to mention that one of the applied approaches combines our runs with the algorithm provided by Barcelona Media (see Section 3). Our experimental results, described in detail in Section 4, demonstrate that sentiment analysis techniques are a good starting point to process and classify online reputation information. Finally, after a brief discussion of the obtained results (see Section 5), we outline some lines for future work in Section 6.

2 An emotional concept-based approach for polarity for reputation

As mentioned above, our main concern is to analyze the applicability of sentiment analysis techniques for classifying the polarity of reputation information. To this aim, we have adapted the approach presented in [6] for polarity and intensity classification of opinionated texts. The main idea of this method is to extract the WordNet concepts in a sentence that entail an emotional meaning, assign them an emotion within a set of categories from an affective lexicon, and use this information as the input of a machine learning algorithm. The strengths of this approach, in contrast to other more simple strategies, are: (1) use of WordNet and a word sense disambiguation algorithm, which allows the system to work with concepts rather than terms, (2) use of emotions instead of terms as classification attributes, and (3) processing of negations and intensifiers to invert, increase or decrease the intensity of the expressed emotions. This system has been shown to outperform previous systems designed for the same task. For filtering, we have implemented a simple approach based on a vote mechanism and contextual information.

2.1 Filtering

In order to determine if a text is related to a given entity we have implemented a simple vote mechanism that calculates a score depending on how many entity context words are found in the input text. The entity context is obtained from the entity website and from the entity entry in Wikipedia, as provided by RepLab. An input text is determined as related or not to a given entity when the final score is upper a certain threshold. Different thresholds have been evaluated, changing from the simple presence of the search query to mentions of the complete entity name or to the presence of more entity context words. Our main objective for this approach is to determine if a simple method of word presence is able to correctly determine the relatedness of the text to a given entity. It is also important to highlight the complexity of the task, since even for humans it is often difficult to decide if a text is ambiguous or not with the respect to a given entity due to the lack of context in the input text. This problem is more evident in microblogging systems such as Twitter, where the text of the post is limited to 140 characters.

As a first step, the system pre-processes each input text splitting it in sentences and isolating the tokens of each sentence using the GATE architecture [9]

for English and the FreeLing library for Spanish [5]. In the same way, the entity contexts, i.e. the entity website and the Wikipedia entry, are also pre-processed. Besides, in this pre-processing step all stop words and symbols included in the input text and the entity contexts are removed to reduce noise. Finally, the search query and the complete entity name from the RepLab 2012 data set are retrieved and preprocessed. The score for an input is calculated using four rules:

- **Rule I:** If a text contains the complete entity name, the highest score, *0.75*, is added to the input score. This decision is based on the idea that a text that includes the complete name of a company rarely will not refer to the entity, as usually, complete names of the companies are the most meaningful and distinctive (e.g., Banco Santander, S.A., Bank of America Corporation, etc.).
- **Rule II:** However, the most frequent way of referring to companies is by means of short names, which are frequently used as queries, such as "Santander" for Banco Santander, S.A. or "Bank of America" for Bank of America Corporation. That is why, when the input text contains an identical to the search query sequence of tokens, the system adds to the total score 2/3 of the maximum score, *0.5*. The reason for using a lower value is that we have found the search queries to be highly ambiguous. For example, "Santander" could be interpreted as a region of Spain or as a bank, depending on the context. Note that in this case we use token matching rather than string matching.
- **Rule III:** Due to the limited length of Twitter posts, in many cases the string of the search query is not tokenized, but included into a hashtag or written omitting blanks (e.g., *#bankofamerica* or *BankofAmerica* instead of "Bank of America"), so different tokens cannot be correctly identified by GATE and FreeLing. To solve this, we have included a further rule that checks if the input string contains the search query after removing blanks, and assigns 1/3 of the maximum score, *0.25*.
- **Rule IV:** Finally, we assume that an input text that contains words from the entity context is more probably related to the entity than other, the more words in common the higher probability. However, as the website and the Wikipedia entry usually include not only domain specific terms, but also many generic words, for each token in the input text that matches a token in the entity context the score of the input is increased only *0.25*.

The score of each input text is then compared to the threshold to determine if the text is related to the entity or not, filtering all the input texts that are not related to the entity.

2.2 Polarity for reputation

The original method presented in [6] has been modified to improve the scope detection approach for negation and intensifiers to deal with the effect of subordinate sentences and special punctuation marks. Besides, the list and weights of

the intensifiers have been adjusted to the most frequent uses in English. Moreover, the presented approach uses the SentiSense affective lexicon [7], that was specifically designed for opinionated texts. Sentisense attaches an emotional category from a set of 14 emotions to WordNet concepts. It also includes the antonym relationship between emotional categories, which allows to capture the effect of some linguistic modifiers such as negation. We also have adapted the system to work with Spanish texts, as the original system was conceived only for English. The method comprises four steps that are described below:

Pre-processing: POS Tagging and Concept Identification The objective of the first step is to translate each text to its conceptual representation in order to work at the concept level in the next steps and avoid word ambiguity. To this aim, the input text is split into sentences and the tokens are tagged with their POS using GATE for English texts and FreeLing for Spanish texts. At this stage, the syntax tree of each sentence is also retrieved using the Stanford Parser [15] for the English texts and the FreeLing library for the Spanish texts. With this information, the system next maps each token to its appropriate WordNet concept using the Lesk word sense disambiguation algorithm as implemented in the WordNet::SenseRelate package [20] for English, and the UKB algorithm [1] as included in the FreeLing library for Spanish. Besides, to enrich the emotion identification step, the hypernyms of each concept are also extracted from WordNet.

Emotion Identification Once the concepts are identified, the next step maps each WordNet synset to its corresponding emotional category in the SentiSense affective lexicon, if any. The emotional categories of the hypernyms are also retrieved. We hypothesize that the hypernyms of a concept entail the same emotions than the concept itself, but decreasing the intensity of the emotion as we move up in the hierarchy. So, when no entry is found in the SentiSense lexicon for a given concept, the system retrieves the emotional category associated to its nearest hypernym, if any. However, only a certain level of hypernymy is accepted, since an excessive generalization introduces some noise in the emotion identification. This parameter has been empirically set to 3. In order to accomplish this step for Spanish texts we have automatically translated the SentiSense lexicon to the Spanish language. To do this, we have automatically updated the synsets in SentiSense to their WordNet 3.0 version using the WordNet mappings. In particular, for nouns and verbs we use the mappings provided by the WordNet team [26] and for adjectives and adverbs, the UPC mappings [22]. In this automatic process we have only found 15 labeled synsets without a direct mapping, which were removed in the new SentiSense version. Finally, in order to translate the SentiSense English version to Spanish we use the Multilingual Central Repository (MRC) [12]. The MCR is an open source database that integrates WordNet versions for five different languages: English, Spanish, Catalan, Basque and Galician. The Inter-Lingual-Index (ILI) allows the automatic translation of synsets from one language to another.

Post-processing: Negation and Intensifiers In this step, the system has to detect and solve the effect of negations and intensifiers over the emotions discovered in the previous step. This process is important, since these linguistic modifiers can change the polarity and intensity of the emotional meaning of the text. It is apparent that the text *#Barclays bank may not achieve 13% on equity target by 2013* entails different polarity than the text *#Barclays bank may achieve 13% on equity target by 2013*, and reputation systems must be aware of this fact.

To this end, our system first identifies the presence of modifiers using a list of common negation and intensification tokens. In such a list, each intensifier is assigned a value that represents its weight or strength. The scope of each modifier is determined using the syntax tree of the sentence in which the modifier arises. We assume as scope all descendant leaf nodes of the common ancestor between the modifier and the word immediately after it, and to the right of the modifier. However, this process may introduce errors in special cases, such as subordinate sentences or those containing punctuation marks. In order to avoid this, our method includes a set of rules to delimit the scope in such cases. These rules are based on specific tokens that usually mark the beginning of a different clause (e.g., *because*, *until*, *why*, *which*, etc.). Since some of these delimiters are ambiguous, their POS is used to disambiguate them. Once the modifiers and their scope are identified, the system solves their effect over the emotions that they affect in the text. The effect of negation is addressed by substituting the emotions assigned to the concepts by their antonyms. In the case of the intensifiers, the concepts that fall into the scope of an intensifier are tagged with the corresponding percentage weight in order to increase or diminish the intensity of the emotions assigned to the concepts.

In particular, for English texts we use the original list of negation signals from [6] and an adapted list of intensifiers with the most frequent uses in English. The percentage of each intensifier has been set empirically. In order to determine the scope for English texts, we use the syntax tree as generated by the Stanford Parser. The same process is replicated for the Spanish texts, and a list of common negation tokens in Spanish (such as *no*, *nunca*, *nada*, *nadie*, etc.) and common intensifiers (*más*, *menos*, *bastante*, *un poco*, etc.) were developed. In order to determine the scope of each modifier, the syntax tree as generated by the FreeLing library is used.

Classification In the last step, all the information generated in the previous steps is used to translate each text into a Vector of Emotional Intensities (VEI), which will be the input to a machine learning algorithm. The VEI is a vector of 14 positions, each of them representing one of the emotional categories of the SentiSense affective lexicon. The values of the vector are generated as follows:

- For each concept, C_i, labeled with an emotional category, E_j, the weight of the concept for that emotional category, $weight(C_i; E_j)$, is set to 1.0.

– If no emotional category was found for the concept, and it was assigned the category of its first labeled hypernym, $hyper_i$, then the weight of the concept is computed as:

$$weight(C_i; E_j) = 1/(depth(hyper_i) + 1) \quad (1)$$

– If the concept is affected by a negation and the antonym emotional category, E_anton_j, was used to label the concept, then the weight of the concept is multiplied by $\alpha = 0.6$. This value has been empirically determined in previous studies. It is worth mentioning that the experiments have shown that α values below 0.5 decrease performance sharply, while it drops gradually for values above 0.6.
– If the concept is affected by an intensifier, then the weight of the concept is increased/decreased by the intensifier percentage, as shown in Equation 2.

$$weight(C_i; E_j) = weight(C_i; E_j) * (100 + intensifier\_percentage)/100 \quad (2)$$

– Finally, the position in the VEI of the emotional category assigned to the concept is incremented by the weight previously calculated.

3 Heterogeneity based ranking approach for polarity for reputation

Our last approach consists in combining our runs (`Uned_1..4`) with the runs provided by Barcelona Media (`BMedia_1..5`) [8]. Given that the RepLab 2012 trial data set is not big enough for training purposes we opted for a voting method. In [27] several voting algorithms for combining classifiers are described. They are focused on the multiplicity of classifiers that support a certain decision. However they do not consider the diversity of classifiers, which is a strong evidence of accuracy when combining classifiers [17]. There are works in which classifiers are selected to be combined while trying to maximize their diversity [11].

We propose a voting algorithm that directly considers the diversity of classifiers instead of the amount of classifiers that corroborate a certain decision. As far as we know, this perspective has not been applied before due to the lack of a diversity measure to be applied over classifier sets rather than pairwise measures. Our approach is inspired in the Heterogeneity Based Ranking [3].

We define the Heterogeneity of a set of classifiers $\mathcal{F} = \{f_1..f_n\}$ as the probability over decision cases $\mathcal{C}$ that there exists at least two classifiers contradicting each other.

$$H(\mathcal{F}) = P_{\mathcal{C}}(\exists f_i, f_j \in \mathcal{F}/f_i(c) \neq f_j(c))$$

The approach basically consists in selecting the label (e.g. positive, neutral) that maximizes the heterogeneity of classifiers corroborating this decision.

4 Evaluation

The data set used for evaluation consists of a training set of 1800 tweets crawled for six companies (300 per company) and a test set of 6243 tweets for 31 companies. However, since Twitter's Terms of Service do not allow redistribution of tweets, some of them have been removed from the release version of the data set. So, the training set that has been finally used contains 1662 tweets manually labeled by experts, 1287 for English and 375 for Spanish. For the filtering task, the training set comprises 1504 tweets related to any of the companies and 158 that were not related to any company. In polarity for reputation, the distribution of tweets between classes is: positive=885, neutral=550, and negative=81. The test set contains tweets manually labeled by experts, 3521 of them are in English and 2722 in Spanish. Besides, this set contains 4354 related tweets, and 1889 unrelated tweets, while the distribution of tweets in polarity for reputation classes is 1625, 1488 and 1241 tweets for positive, neutral and negative, respectively. For the profiling task (i.e., combining systems of filtering and polarity), the performance is evaluated as accuracy on a four-class classification problem: irrelevant, relevant negative, relevant neutral and relevant positive. For the individual evaluation of the filtering and polarity tasks, performance is evaluated in terms of reliability (R) and sensitivity (S)[3]. Accuracy (% of correctly annotated cases) is also included for comparison purposes. Overall scores (R, S, F(R,S), accuracy) are computed as the average of individual scores per entity, assigning the same weight to all entities.

As the system for polarity for reputation is a supervised method, we have tested different machine learning algorithms with the training set in order select the best classifier for the task. To determine the best machine learning algorithm for the task, 20 classifiers currently implemented in Weka [13] were compared. We only show the results of the best performance classifiers: a logistic regression model (Logistic) and a classifier using decision trees (RandomForest). The best outcomes for the two algorithms were reported when using their default parameters in Weka. For the filtering task, we have evaluated over the training set different thresholds. In particular, we have evaluated the values *0.25* (just the presence of the query search or the presence of an entity context word), *0.5* (the presence of the query search plus some entity context word, or the exact match of tokens with the query search), *0.75* (the presence of the company name, or different combinations of the query search and entity context words) and *1.0* (multiple combinations of the query search or the company name and the entity context word). We have found that, upper this last threshold, the performance of the system decreases sharply. For RepLab 2012, we have selected two thresholds that produce better results, i.e. *0.5* and *0.75*.

Based on these considerations, five systems have been presented to RepLab 2012: `Uned_1` (Logistic + threshold=*0.5*), `Uned_2` (RandomForest + threshold=*0.5*), `Uned_3` (Logistic + threshold=*0.75*), `Uned_4` (RandomForest + threshold=*0.75*), and `Uned_5` or `Uned-BMedia`, which is the system described in the sec-

[3] See guidelines at `http://www.limosine-project.eu/events/replab2012`

tion 3 that combines the outputs of the algorithms `Uned_1..4` and `BMedia_1..5` using the heterogeneity-based ranking approach.

Table 1 shows the results of our systems for the filtering task when evaluated over the test set. As can be seen, the best performance in terms of F(R,S) is obtained by the two approaches that use the *0.75* threshold (`Uned_3` and `Uned_4`), followed by the combined version `Uned-BMedia` and the two approaches that use the *0.5* threshold (`Uned_1` and `Uned_2`). In terms of accuracy, the best result is obtained by the combined version `Uned-BMedia`, only *2.0* percentage points more than the two approaches that use the *0.75* threshold and *3.1* percentage points more than the *0.5* threshold approaches. Comparing these results with the *All relevant* baseline, it is evident that the performance of our approaches is quite acceptable but may be easily improved. It is important to recall the difficulty of the task even for humans. In fact, the best results obtained in the challenge by our systems, `Uned_3` and `Uned_4`, are placed 13th and 14th, respectively, out of 33 participants [2].

Table 1. Results for the filtering task

Systems	**Acc.**	**R**	**S**	**F(R,S)**
Uned_1	0,693414595	0,161145627	0,151777683	0,08899625
Uned_2	0,693414595	0,161145627	0,151777683	0,08899625
Uned_3	0,705452028	0,172086104	**0,254595211**	**0,134324737**
Uned_4	0,705452028	0,172086104	**0,254595211**	**0,134324737**
Uned-BMedia	**0,724939924**	**0,177089017**	0,212287163	0,114018152
All relevant	0,709480928	0	0	0

Table 2 summarizes the results obtained by our systems in terms of accuracy, reliability and sensitivity over the test set. The best performance in terms of F(R,S) is achieved by the combined version `Uned-BMedia`, while the systems `Uned_2` and `Uned_4` that use the RandomForest classifier are only *4.0* percentage points lower. The systems `Uned_1` and `Uned_3` that use the Logistic classifier achieve *8.0* percentage points less than the `Uned-BMedia`, which is an important difference in performance. In contrast, in terms of accuracy the best performance is obtained by the RandomForest classifier (`Uned_2` and `Uned_4`), followed by the combined version `Uned-BMedia` and the Logistic classifier. As can be seen in the table, all the systems improved over the *All positives* baseline. It is worth noticing that the results achieved for the polarity for reputation task are much better than the obtained for filtering task. This is evidenced by the fact that the combined system `Uned-BMedia` obtained the 2nd position within the 35 participating systems in term of F(R,S), and the two approaches that use the RandomForest classifier obtained the 5th and 6th position, respectively. However, in terms of accuracy the results reported by our systems are even better, the two approaches that use the RandomForest classifier have achieved the 1st and 2nd

position in the ranking, while the combined version `Uned-BMedia` has achieved the 4th position.

Table 2. Results for the polarity for reputation task

Systems	**Acc.**	**R**	**S.**	**F(R,S)**
Uned_1	0,442352774	0,315276995	0,243799781	0,262266787
Uned_2	**0,486569957**	0,325450278	0,314695031	0,307839944
Uned_3	0,442352774	0,315276995	0,243799781	0,262266787
Uned_4	**0,486569957**	0,325450278	0,314695031	0,307839944
Uned-BMedia	0,449501547	**0,340229898**	**0,374731432**	**0,341946295**
All positives	0,438481115	0	0	0

The results of our systems for the profiling task, i.e. combining filtering and polarity for reputation, are shown on Table 3. The best result is achieved by the system `Uned_4`. However, the difference is not so marked with respect to the `Uned_3` and the combined version `Uned-BMedia`, *0.6* and *1.9* percentage points of accuracy, respectively. The difference becomes higher comparing to the systems `Uned_2` and `Uned_1`, *5.6* and *7.2* percentage points of accuracy. Moreover, all systems considerably improve the performance over the *All relevant and positives* baseline. As in the polarity for reputation task, the results achieved by our systems compare favorably to those of other participants. In particular, three of our systems are among six best systems out of 28 systems that participated in the task (the 3rd, the 5th and the 6th for `Uned_4`, `Uned_3` and `Uned-BMedia`, respectively), which proves that the proposed approaches are a good starting point for a more complex profiling system.

Table 3. Results for the profiling task

Systems	**Acc.**
Uned_1	0,319710973
Uned_2	0,335352452
Uned_3	0,385183273
Uned_4	**0,391658071**
Uned-BMedia	0,372795736
All relevant and positives	0,274057566

5 Discussion

Analysing and discussing the results obtained by the systems proposed for RepLab 2012, first of all and regarding the filtering task, we have to say that our

system correctly classified 3352 out of 4354 related tweets (76.9%) and 1019 out of 1831 non-related tweets (55.6%). These results suggest that the *0.75* threshold in the vote mechanism is a good choice for discriminating between related and unrelated tweets. Most of the related tweets were above this threshold, and nearly half of the unrelated tweets were below it. However, a number of unrelated tweets obtained a score far above *0.75*, which seems to suggest that the filtering method should take into account not only positive terms but also negative ones, i.e. terms that decrease the score. We also have analyzed which are the most frequent combination of rules for assigning scores, as well as their adequacy. We have found that rule II (i.e. the one that considers the presence in the input of the individual tokens within the search query) and rule III (i.e. the one that looks for the search query without blanks), are frequently launched together producing satisfactory results. The combination of rule III and rule IV (i.e. the presence of entity context words) is the second most frequent one. However, it is also the one that introduces more noise. Finally, the presence of the complete entity name (rule I) is the less launched rule, but the one that produces the best performance, as expected.

Regarding the polarity for reputation task, our best systems in terms of accuracy (`Uned_2` and `Uned_4`) correctly classified 822 tweets as positive, 810 tweets as neutral and only 179 tweets as negative. These results show the good performance of these systems at identifying the positive and neutral classes. In contrast, the systems only classify 15% of negative tweets correctly. This difference may be due to the fact that the number of negative instances in the training set is not big enough for the classifiers to correctly learn this class.

Even if the systems achieve a good performance in polarity for reputation, it is worth mentioning that these results are lower than those reported by the same systems when evaluated with other datasets containing product reviews. In order to determine the reason of this, we analyzed the datasets and found that an important number of tweets are not labeled with any emotion. In particular, the coverage of the vocabulary in SentiSense for the training set is 12.5% for English and 12.6% for Spanish, while the coverage in the test set is 11.3% and 11.2% for English and Spanish, respectively. This was expected, since SentiSense is specially designed for processing product reviews. Therefore, taking this low coverage into account, we expect that expanding SentiSense with reputation-related vocabulary will allow us to significantly improve the classification results.

However, as we suspected, another important source of classification errors is that many of the tweets labeled with polarity in the dataset (positive and negative) do not contain any emotional meaning per se. These are factual expressions that, from a reputation perspective, entail a polarity, and therefore must be classified as negative or positive. For example, the tweet *I bough a Lufthansa ticket to Berlin yesterday* does not express any clear emotion, but is considered as positive for reputation purposes. To this aim, reputation management knowledge must be used to allow the system to correctly interpret these expressions.

The importance of the opinion holder and, specially for microblogging, the importance of the external links in the input text are other two important find-

ings of our analysis. On the one hand, we find some examples of tweets that, from our point of view, should be classified as neutral, but were classified by the experts as polar due to the relevance or popularity of the the tweet's author. So, we can conclude that correctly determining the opinion holder is important to weight the final polarity for reputation. On the other hand, we find that many of the tweets that have external links obtain their polarity from the linked documents. This suggests that studying polarity for reputation in social media, especially in such microblogging services as Twitter, needs more context than the posts themselves.

Finally, the results obtained in the profiling task show the good adequacy of complex sentiment analysis approaches as a starting point to analyze texts according to the reputation of a given entity. Even if the performance of the combined system, filtering + polarity for reputation, is quite acceptable, there is still room for improvement. It is important to notice that the evaluation of this task includes, first, filtering relevant tweets, and next, the polarity classification of these relevant tweets, so that the error of the filtering step is propagated to the polarity for reputation classification.

6 Conclusion and Future Work

In this paper, we have presented a system to address the task of classifying polarity for reputation, based on sentiment analysis techniques. We also face the problem of determining whether a text is related or not to a given entity (e.g., company or brand). The results showed that, even if the task is more complex than classifying polarity in opinionated texts, the use of complex sentiment analysis approaches seems a good basis for classifying polarity for reputation. However, it is still necessary to incorporate a high level of expert knowledge to correctly analyze the reputation of a company.

As future work, we plan to develop a system for separating objective from subjective statements. This separation will allow us to study objective texts from a reputation perspective, while subjective opinions will be analyzed using sentiment analysis approaches. Our analysis has also revealed the importance of detecting the opinion holder, which may influence the polarity of the text. To address this, we plan to study existing approaches in the area and evaluate them for the polarity for reputation task. Finally, we would like to extend the SentiSense affective lexicon to cover more domain specific vocabulary.

References

1. Agirre, E., Soroa, A.: Personalizing PageRank for Word Sense Disambiguation. In: proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics. (2009)
2. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., Rijke, M.: Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In: proceedings of CLEF 2012 Labs and Workshop Notebook Papers. (2012)

3. Amigó, E., Gimenez, J., Gonzalo, J., Verdejo, F.: UNED: Improving Text Similarity Measures without Human Assessments. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 454–460. (2012)
4. Brooke, J.: A Semantic approach to automated text sentiment analysis. Unpublished doctoral dissertation, Simon Fraser University, Canada. (2009)
5. Carreras, X., Chao, I., Padr, LL., Padr, M.: FreeLing: An Open-Source Suite of Language Analyzers. In: proceedings of the 4th International Conference on Language Resources and Evaluation. (2004)
6. Carrillo de Albornoz, J., Plaza, L., Gervas, P.: A Hybrid Approach to Emotional Sentence Polarity and Intensity Classification. In: proceedings of the 14th Conference on Computational Natural Language Learning, pp. 153-161. (2010)
7. Carrillo de Albornoz, J., Plaza, L., Gervs, P.: SentiSense: An easily scalable concept-based affective lexicon for Sentiment Analysis. In: proceedings of the 8th International Conference on Language Resources and Evaluation. (2012)
8. Chenlo, J.M., Atserias, J., Rodriguez, C., Blanco, R..: FBM-Yahoo! at RepLab 2012. In: proceedings CLEF 2012 Labs and Workshop Notebook Paper. (2012)
9. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities, 36, pp. 223-254. (2002)
10. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 193–200. (2006)
11. Giacinto, G., Roli, F.: An Approach to the Automatic Design of Multiple Classifier Systems. Pattern Recognition Letters, 22, pp. 25–33. (2001)
12. Gonzalez-Agirre, A., Laparra, E., Rigau, G.: Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In: proceedings of the Sixth International Global WordNet Conference. (2012)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl., 11, pp. 10–18. (2009)
14. Hoffman, T.: Online reputation management is hot? But is it ethical? In: Computerworld, 44, February (2008)
15. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430. (2003)
16. Kim, S-M., Hovy, E.: Determining the sentiment of opinions. In: proceedings of the 20th Conference on Computational Linguistic, pp.1367–1373. (2004)
17. Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., Duin, R. P. W.: Is Independence Good For Combining Classifiers?. In: proceedings of the 15th International Conference on Pattern Recognition, pp. 168–171. (2000)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using Machine Learning techniques. In: proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86. (2002)
19. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 271–278. (2004)
20. Patwardhan, S., Banerjee, S., Pedersen, T. : SenseRelate::TargetWord - A generalized framework for word sense disambiguation. In: proceedings of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions, pp. 73-76. (2005)

21. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424.(2002)
22. Universidad Politécnica de Cataluña mappings. http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=57.
23. Wiebe, J., Bruce, R., O'Hara, T.: Development and use of a gold-standard data set for subjectivity classification. In: proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp 246–253. (1999)
24. Wiegand, M., Klakow, D.: Bootstrapping supervised machine-learning polarity classifiers with rule-based classification. In: proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 59–66. (2010)
25. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35, pp. 399–433. (2009)
26. WordNet mappings. http://wordnet.princeton.edu/wordnet/download/.
27. Xu, L., Krzyzak, A., Suen, C.Y.: Methods for combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 22, pp. 418–435. (1992)