UAMCLyR at RepLab 2013: Profiling Task*

Notebook for RepLab at CLEF 2013

Esaú Villatoro-Tello[1], Carlos Rodríguez-Lucatero[1],
Christian Sánchez-Sánchez[1], and A. Pastor López-Monroy[2]

[1] Departamento de Tecnologías de la Información,
Universidad Autónoma Metropolitana, Unidad Cuajimalpa,
Ave. Vasco de Quiroga Num. 4871 Col Santa Fe, México D.F.
{evillatoro,crodriguez,csanchez}@correo.cua.uam.mx
[2] Department of Computer Science,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
pastor@ccc.inaoep.mx

Abstract. This paper describes the participation of the Language and Reasoning Group of UAM at RepLab 2013 Profiling evaluation lab. We adopted Distributional Term Representations (DTR) for facing the following problems: *i) filtering* tweets that are related to an entity, and *ii)* identifying positive or negative implications for the entity's reputation, *i.e., polarity for reputation.* Distributional Term Representations help to overcome, to some extent, the small-length/high-sparsity issues. DTRs are a way to represent terms by means of contextual information, given by term co-occurrence statistics. In order to evaluate our approach, we compared the proposed approach against the traditional Bag-of-Words representation. Obtained results indicate that by means of DTRs it is possible to increase the reliability score of a *profiling* system.

Keywords: Bag of words, Distributional term representations,Term co-occurrence representation, Term selection, Supervised text classification

1 Introduction

From its inception in 2006, Twitter has become in one of the most important platform for microblog posts. Recent statistics reveal that there are more that 200 million users that write more than 400 million posts every day[3], talking about a great diversity of topics. As a consequence, several entities such as companies, celebrities, politicians, etc., are very interested in using this type of platform for increasing or even improving their presence among Twitter users, aiming at obtaining good reputation values. As an important effort for providing effective solutions to the above problem, RepLab[4] proposes a competitive evaluation exercise for Online Reputation Management (ORM) systems. As one of the main tasks evaluated in RepLab is the *Profiling* task. This particular task

* This work was partially supported by CONACyT México Project Grant CB-2010/153315, and SEP-PROMEP Project Grant UAM-C-CA-31/10847.

[3] *http://blog.twitter.com/2013/03/celebrating-twitter7.html*

[4] *http://www.limosine-project.eu/events/replab2013*

consists of mining the reputation of a company from online media. Adequate *profiling* systems must be able to retrieve several posts from several online sources, and annotating them according to their relevancy, *i.e.*, to preserve online documents related to the company and to identify all positive or negative implications for the company contained in such documents [1].

As mention in [1], systems that face the *profiling task* must annotate two different types of information: *i)* **Filtering**: This means that an automatic system must be able to decide whether a given tweet is related to a particular company or not. Basically it represents a two class problem since systems must tag a tweet as "related" or "not related"; and, *ii)* **Polarity for Reputation**: The idea of this particular subtask is to identify if a given tweet contains positive or negative implications for the company's reputation. This problem represent a three class problem since an automatic system have to assigns a "positive", "negative" or "neutral" tag for each tweet related to a particular company.

Our proposed approach for facing both *filtering* and *polarity* problems is based on distributional term representations (DTRs) [3], which are a way to represent terms by means of contextual information, given by term-co-occurrence statistics. Accordingly, this paper presents the details of the participation of the Language and Reasoning group from UAM-C to the CLEF 2013 RepLab profiling task (*i.e.*, *filtering* and *polarity* for reputation). The main objectives of our experiments were:

1. To test if a richer document representation based on term co-occurrences can be successfully applied to *filtering* and *polarity* subtasks.
2. To estimate how useful our previously developed methods for sentiment analysis on Twitter can be adopted for detecting positive and negative implications of tweets in the context of the RepLab exercise.
3. To evaluate to what extent supervised techniques are able to solve both *filtering* and *polarity* problems.

The rest of this paper is organized as follows. The next section describes all the steps considered in the pre-processing stage. Section 3 describe the proposed representation strategy. Section 4 describes the experimental setup we followed, as well as our results obtained for both *filtering* and *polarity* subtasks. Finally, Section 5 presents the conclusions derived from this work and outlines future work directions.

2 Tweets pre-processing

It is worth mentioning that for performing all our experiments we collected two different versions of the collection of tweets which are described below:

Main: For this configuration we crawled only the main tweet from each given tweet id. In other words, all other tweets contained in the original tweet id (e.g., answers or comments generated by the original tweet) are ignored.

All: For this configuration, we crawled both the main tweet and all answers or comments generated by the original tweet from each given tweet id.

When retrieving the **All** version of the tweets collection, our intuitive idea was to evaluate the impact of all conversational elements of a tweet when deciding its polarity as well as its relevancy. Notice that this crawling procedure was replicated when retrieving test tweets.

As pre-processing steps we applied the following procedures to each tweet in the two versions of the tweets collection (*i.e.*, Main and All):

1. All tweets are transform to lowercase.
2. All users mentions (*i.e.*, *@user*) are replaced by the tag: AT-USER.
3. Every outgoing link is replaced by the tag: OUTGOING-LINK, hence, for performed experiments we did not use the information contained in these links, however we believe they can be useful when trying to detect if a tweet is related or not to a company.
4. All hashtags (*i.e.*, *#hashtagX*) are replaced by the tag: HASHTAG.
5. All punctuation mark as well as emoticons are deleted.
6. We apply the Porter stemming [2].
7. All stopwords are deleted.

3 Tweets representation

Distributional term representations (DTRs) are tools for term representation that rely on term occurrence and co-occurrence statistics [3]. Intuitively, the meaning of a term is determined by the context in which it occurs. Where the context is given in terms of other terms in the vocabulary. In this paper we consider one popular DTR, namely term-co-occurrence representation. This DTR has been mainly used in term classification and term clustering tasks, and very recently for short-text categorization [4], where their potential benefits for term expansion are shown.

The term co-occurrence representation (TCOR) is based on co-occurrence statistics. The underlying idea is that the semantics of a term t_j can be revealed by other terms it co-occur with across the document collection. Here, each term $t_j \in T$ is represented by a vector of weights $w_j = \langle w_{1,j}, \ldots, w_{|T|,j} \rangle$, where $0 \leq w_{k,j} \leq 1$ represents the contribution of term t_k to semantic description of t_j:

$$w_{k,t} = tff(t_k, t_j) \cdot log \frac{|T|}{T_k} \tag{1}$$

where T_k is the number of different terms in the dictionary T that co-occur with t_j in at least one document and

$$tff(t_k, t_j) = \begin{cases} 1 + \log(\#(t_k, t_j)) & if(\#(t_k, t_j) > 0) \\ 0 & otherwise \end{cases} \tag{2}$$

where $\#(t_k, t_j)$ denotes the number of documents in which term t_j co-occurs with the term t_k. The intuition behind this weighting scheme is that the more t_k and t_j co-occur the more important t_k is for describing term t_j; the more terms co-occur with t_k the less important is to define the semantics of t_j. At the end, the vector of weights is normalized to have unit 2-norm: $||w_j||_2 = 1$.

Finally, let w_{t_j} denote the DTR of term t_j in the vocabulary, where w_{t_j} is the TCOR representation. The representation of a document d_i based on this DTR is obtained as follows:

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j} \tag{3}$$

where α_j is a scalar that weights the contribution of term $t_j \in d_i$ into the document representation. Thus, the representation of a document is given by the (weighted) aggregation of the contextual representations of terms appearing in the document. That is, the document representation is a summary of the contextual information present in the terms that appear in the document.

Under TCOR, a document d_i is represented by $d_i^{dtr} \in \mathbb{R}^{|T|}$, a vector of the same dimensionality as the vocabulary. The values of d_i^{dtr} indicate the association between terms in the vocabulary and those terms that occur in d_i. Notice that scalar α_{t_j} aims to weight the importance that term t_j has for describing document d_i. Many options are available for defining α_{t_j}, in this work we considered the following weights: *Boolean* (BOOL), *Term-Frequency* (TF), and *Relative Frequency* (TF-IDF).

Notice that using this type of representations can lead to problems of high dimensionality, since the number of terms (features) usually accomplish that $T \rightarrow \infty$. This fact may lead to problems of *over-fitting* when training a classifier. A technique that has been used as a feature selection strategy is by means of preserving terms near to the transition point pt_T [5,6]. The pt_T represents a frequency value that divides vocabulary terms T in two sets, those of low frequency and those of high frequency.

In a previous work [6], we have shown that by means of preserving high frequency terms in conjunction with a subset of low frequency terms, it is possible to solve (to some extent) the problem of assigning polarity values to twitter posts, especially for a three class problem (*i.e.*, positive, negative and neutral). Accordingly, we defined a subset of experiments for the polarity subtask employing this strategy as features selection technique.

4 Experimental Results

For the RepLab 2013 edition participant teams were given a large dataset (61 entities) from four domains: automotive, banking, universities and music/artists. For trial dataset, approximately 700 tweets were provided for each entity. Contrary to the RepLab 2012 edition, RepLab 2013 organizers provided as test dataset tweets from the same 61 entities that where used as trial dataset. For these, approximately 1700 tweets were crawled.

Given this situation, *i.e.*, same entities for training and for testing, we decided to adopt a supervised strategy for solving the problem of *filtering* and *polarity*. We report our results for the test dataset in terms of Reliability, Sensibility and their harmonic mean[7].

As we mentioned in Section 1, our goals were to test if by means of employing a richer documents representation (see Section 3) it would be possible to solve both sub tasks involved in the *profiling* problem. Consequently, we defined as our baseline

method the traditional Bag-of-Words (BOW) representation. Finally, it is worth mentioning that we used, for all our experiments; as our main classifier the Weka's[5] Support Vector Machine implementation considering a linear kernel configuration.

4.1 Filtering results

Table 1 describe the configuration assigned to each experiment for performed experiments in terms of type of *representation* (BOW or TCOR), *weighting* scheme (BOOL, TF or TF-IDF) and type of *tweets* collection used (Main or All). Notice that each column, from 2nd to 7th, represent one experiment definition, *i.e.*, one run (6 runs were submitted in total).

Table 1. Configuration for submitted experiments: Filtering subtask.

Configuration/Run ID	**Run 01**	**Run 02**	**Run 03**	**Run 04**	**Run 05**	**Run 06**
Representation	BOW	BOW	TCOR	BOW	BOW	TCOR
Weighting	BOOL	TF	BOOL	BOOL	TF	BOOL
Tweets	Main	Main	Main	All	All	All

Table 2 show obtained results for *filtering* subtask. Last two rows indicate: *i)* the *baseline* performance as defined in[8], and *ii)* the average performance of all participant teams in the RepLab 2013 edition.

Table 2. Filtering subtask results

Run ID	**Reliability (R)**	**Sensitivity (S)**	**F (R, S)**	**Accuracy**
UAMCLyR filtering 01	0.6311	0.3960	0.3759	**0.9132**
UAMCLyR filtering 02	0.5731	0.3132	0.2918	0.9007
UAMCLyR filtering 03	**0.6964**	0.3038	0.3220	0.9041
UAMCLyR filtering 04	0.5554	**0.4015**	**0.3787**	0.9110
UAMCLyR filtering 05	0.5688	0.3075	0.2858	0.8996
UAMCLyR filtering 06	0.6292	0.2828	0.2637	0.8906
BASELINE	*0.4902*	*0.3199*	*0.3255*	*0.8714*
Average	*0.4663*	*0.2951*	*0.2596*	*0.7628*

Notice that by means of using a BOW representation with a boolean weighting scheme (run 01, and run 04) allows to obtain the higher accuracy values. This might be an indicator that only by the presence of some words it is possible to decide whether a tweet is related to a company or not.

Additionally, it is important to note that our DTR representation (run 03 and run 06) were able to achieve a better performance than the traditional BOW in terms of

[5] *http://www.cs.waikato.ac.nz/ml/weka/index.html*

reliability measure without considerably decreasing the accuracy. Somehow, this results are an indicator of a better precision, which under a real scenario, it might be more important than the *sensibility*.

4.2 Polarity for reputation results

Table 3 describe the configuration assigned to each performed experiment for the *polarity* subtasks, and Table 4 show obtained results for our performed experiments in the *polarity* subtask.

Table 3. Configuration for submitted experiments: Polarity for reputation subtask.

Configuration/Run ID	**Run 01**	**Run 02**	**Run 03**	**Run 04**	**Run 05**	**Run 06**
Representation	BOW	TCOR	BOW	TCOR	BOW	TCOR
Weighting	TF-IDF	TF-IDF	TF	TF	BOOL	BOOL
Tweets	Main	Main	All	All	All(tp_T)	All(tp_T)

Notice that our bets results in terms of *reliability* and *accuracy* were obtained by means of using a TCOR representation with a TF-IDF weighting scheme using only the Main version of tweets (*i.e.*, run 02). This represent an interesting result, since indicates that the polarity of a tweet can be determined by considering the context in which the tweet's terms occurs. In general, DTR experiments (run 02, 04 and 06) obtain better *reliability* performance.

Table 4. Polarity subtask results

Run ID	**Reliability (R)**	**Sensitivity (S)**	**F (R, S)**	**Accuracy**
UAMCLyR polarity 01	0.3461	0.2695	0.2922	0.5827
UAMCLyR polarity 02	**0.3802**	0.2651	0.2946	**0.6177**
UAMCLyR polarity 03	0.3480	0.2660	0.2891	0.5846
UAMCLyR polarity 04	0.3696	0.1933	0.2251	0.5836
UAMCLyR polarity 05	0.3291	**0.2864**	**0.3008**	0.5778
UAMCLyR polarity 06	0.3440	0.1855	0.2157	0.5370
BASELINE	*0.3151*	*0.2899*	*0.2973*	*0.5840*
Average	*0.4833*	*0.2087*	*0.2267*	*0.5007*

It is also important to remark that performed experiments applying a feature selection strategy by means of the tp_T (run 05 an 06) are able to obtain acceptable results in terms of *sensitivity* and *F(R,S)*. We think that performing additional experiments under similar circumstances but using the "Main" version of the tweets collection will allow to obtain better results.

5 Conclusions and Future work

In this paper, we have described the experiments performed by the Language and Reasoning group from UAM-C in the context of the RepLab 2013 evaluation exercise. Our proposed system was designed for addressing the problem of *filtering* tweets (*i.e.*, determining whether a tweet is related or not to a given entity name) as well as for classifying *polarity* for reputation, *i.e.*, identifying positive or negative implications contained in the tweet.

Our proposed system is based on the use of DTRs as form of representation for tweets texts. This type of representations assume that the meaning of a term is determined by the context in which it occurs. Where the context is given in terms of other terms in the vocabulary. Obtained results showed that DTR representation allows to obtain a better performance in terms of the *reliability* measure, indicating to some extent that this type of representations allow better precision values both in *filtering* and *polarity* subtasks.

Additionally, we also observed that applying the transition point (tp_T) as feature selection strategy allowed our system to obtain good results in terms of the *sensibility* measure. We believe that this strategy might be useful when employing the "Main" version of the tweets collection.

As future work we plan to develop a system that considers information contained on the entity's web page, as well as considering all the emoticons and hashtags contained in tweets texts. Additionally, we plan to evaluate some other DTR representations, since obtained results motivate us to keep working on this direction.

References

1. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., and Rijke, M. (2012) Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In *Working Notes for the CLEF 2012 Evaluation Labs and Workshop.* Rome, Italy.
2. Porter , M. F. (1997) An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc. pp. 313-316.
3. Lavelli, A. and Sebastiani, F. and Zanoli, R. (2004) Distributional Term Representations: An Experimental Comparison. In *Italian Workshop on Advanced Database Systems.*
4. Cabrera, J. M., Escalante, H. J., Montes-y-Gómez, M. (2013) Distributional term representations for short text categorization. In *14th International Conference on Intelligent Text Processing and Computational Linguistics, CI-CICLING 2013.* Samos, Greece.
5. Reyes-Aguirre, B., Moyotl-Hernández, E., y Jiménez-Salazar, H. (2003) Reducción de términos índice usando el punto de transición. En *Avances en Ciencias de la Computación.* pp. 127-130.
6. Leon Martagón, G., Villatoro-Tello, E., Jiménez-Salazar, H., and Sánchez-Sánchez, C. (2013) Análisis de Polaridad en Twitter. In *Journal of Research in Computer Science.* Vol. 62, pp. 69-78.
7. Amigó, E. and Gonzalo, J. and Verdejo, F. (2013) A General Evaluation Measure for Document Organization Tasks. In *Proceedings SIGIR 2013.* Dublin, Ireland.
8. Amigó, E. and Carrillo de Albornoz, J. and Chugur, I. and Corujo, A. and Gonzalo, J. and Martín, T. and Meij, E. and de Rijke, M. and Spina, D. (2013) Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In *Proceedings of the Fourth International Conference of the CLEF initiative, CLEF 2013.* Springer LNCS, Valencia, Spain.