Lexical Context for Profiling Reputation of Corporate Entities
Jean-Valère Cossu, Liana Ermakova

To cite this version:

HAL Id: hal-01537189
https://hal.archives-ouvertes.fr/hal-01537189
Submitted on 12 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Keywords: Online Reputation Monitoring, Topic Categorization, Contextualization, Query Expansion, Natural Language Processing, Information Retrieval, Tweet, Classification

Abstract: Opinion and trend mining on micro-blogs like Twitter recently attracted research interest in several fields including Information Retrieval (IR) and Natural Language Processing (NLP). However, the performance of existing approaches is limited by the quality of available training material. Moreover, explaining automatic systems’ suggestions for decision support is a difficult task thanks to this lack of data. One of the promising solutions of this issue is the enrichment of textual content using large micro-blog archives or external document collections, e.g. Wikipedia. Despite some advantages in Reputation Dimension Classification (RDC) task pushed by RepLab, it remains a research challenge. In this paper we introduce a supervised classification method for RDC based on a threshold intersection graph. We analyzed the impact of various micro-blogs extension methods on RDC performance. We demonstrated that simple statistical NLP methods that do not require any external resources can be easily optimized to outperform the state-of-the-art approaches in RDC task. Then, the conducted experiments proved that the micro-blog enrichment by effective expansion techniques can improve classification quality.

1 INTRODUCTION

The rise of online social media and the increasing amount of user-generated content led an emerging trend in Opinion Mining: Online Reputation Monitoring (ORM) helping keep track of online reputation of a personage or an organization. ORM helps respond to comments in a timely manner, and improve/change products, services or/and customer experience based on the obtained feedback. Reputation affects decisions, especially in business and politics. Tackling this issue then necessitates new management tools and strategies. We consider here, Reputation analysis tools which may be viewed as a part of an ORM platform. We define Reputation analysis as the process of tracking, investigating and reporting entity’s actions and opinions about those actions expresses by other entities.

Web represents a large source of opinions about various kind of entities (Malaga, 2001). Also, Twitter is currently considered as one of the largest online communities with an increasing number of users and data becoming of first interest for researches in the fields of Social Network Analysis, IR and NLP (Anwar Hridoy et al., 2015). Created in March 2006, Twitter rapidly gained worldwide popularity and in March 2016 had more than 310 million monthly active users (Twitter, 2016). On the day of the 2016 U.S. presidential election, Twitter proved to be the largest source of breaking news, with 40 million tweets sent that day (Miller, 2016). However, conversely to Yelp or TripAdvisor, tweets do not provide explicit ratings and, therefore, cannot be directly used in a reputation survey. Then, although significant advances have been made with RepLab (Amigó et al., 2013; Amigó et al., 2014), analyzing reputation of companies and individuals requires a complex modeling of these entities (e.g. companies, artists) and it is still a significant research challenge.

The reputation analysis as defined in RepLab in-
cludes 5 sub-tasks:
1. Filtering;
2. Reputation Polarity Detection;
3. Topic Detection;
4. Priority Ranking;
5. Reputation Dimensions Classification.

In this research we mainly focus on the latter task, namely Reputation Dimension Classification (RDC). RDC is a classification task aiming at categorizing tweets according to their reputational dimensions, among those defined by experts for instance esteem in Products and Services or trust in Financial Governance and Performance. These aspects represent a particular interest for stakeholders in the companies. We propose a statistical classification method based on lexical context relations. We tested the proposed approach on the RepLab 2014 data-set that provides a framework to evaluate ORM systems on Twitter.

We look towards improvements we can expect using (i) typical parameter optimization, e.g. weights, vs. (ii) tweet expansion techniques, e.g. Query Expansion (QE) or contextualization. Tweet Contextualization (TC) was introduced at INEX 2011-2014. This challenge, partially based on RepLab data, aims at automatically providing context information for a tweet using external text collection such as Wikipedia to allow (a human) better understanding of tweets meaning. The INEX organizers provide a framework to evaluate these contexts which represents an interesting material to investigate how much additional information could improve high performing classifiers. In contrast to Query Expansion that mainly provides keywords, Contextualization takes the form of a small readable summary by selecting appropriate passages in the restricted context of topic related to the tweet which is much more difficult than other generic tweet-based summarization tasks. This then raises the three following questions:

- Does query expansion really improve a well adapted classifier?
- Is there a difference in terms of RDC classification performances between query expansion and contextualization?

In this paper we try to answer these questions.

The rest of the paper is organized as follows: Section 2 gives an overview of related work. Sections 3 presents the proposed approach. The descriptions of the used data set, evaluation measures and competitors are provided in Section 4. A thorough discussion of our results is provided in Section 5. Finally, Section 6 draws the conclusions of our work and opens several perspectives.

2 Related Work

Much attention was paid to Opinion Mining over the last decade. However, several times researchers limited the role of Opinion Mining tools to a module in a more general framework providing feedback concerning the sentiment of the population with respect to several concepts. Moreover, previous researches also exploited the use of supervised methods for Topic Categorization of short social chat messages which also required heavy human annotations. However, due to the lack of applicable performance metrics and exploitable gold-standard labels it is hard to report the system performance, nor generalize approaches used within this framework. Especially with large-scale micro-blogging sets of messages, like tweet-collection where annotation is not always available and when it is not possible to consider specific lexicon such as Wikipedia context. The INEX organizers provide a framework to evaluate these contexts which represents an interesting material to investigate how much additional information could improve high performing classifiers. In contrast to Query Expansion that mainly provides keywords, Contextualization takes the form of a small readable summary by selecting appropriate passages in the restricted context of topic related to the tweet which is much more difficult than other generic tweet-based summarization tasks. This then raises the three following questions:

- Does query expansion really improve a well adapted classifier?
- Is there a way to automatically evaluate the quality of these additional information?

http://www.limosine-project.eu/events/replab2013
http://inex.mmci.uni-saarland.de/tracks/qa/

A system which parameters had been tuned to obtain a better performance for a particular task. Learn-to-rank is currently a popular approach (Deveaud et al., 2016).

http://www.popstar.pt

- Is there a difference in terms of RDC classification performances between query expansion and contextualization?
tion issue used links between content and meta-data combining both supervised (Naive Bayes and Sequential Minimal Optimization Support Vector Machines) and unsupervised algorithms (K-star) with terms selection (Sánchez-Sánchez et al., 2013).

In 2014, RepLab (Amigó et al., 2014) focused on the RDC. The task can be viewed as a complement to Topic Detection, as it provides a broad classification of the aspects of the company under public scrutiny. These dimensions reflect affective and cognitive perceptions of a company by several stakeholders and contribute to a better understanding of a tweet’s topic or group of tweets. Participants employed different features. A first group proposed psychometric and linguistic information (Vilares et al., 2014) while a second group mainly considered information beyond the tweet textual content in their approaches (Karisani et al., 2015). We can additionally mention Query Expansion with pseudo-relevant document (McDonald et al., 2015) or Semantic Expansion (Rahimi et al., 2014) to improve a weak classifier using enrichment from large web corpora. Focusing on Wikipedia (Qureshi, 2015) proposed a Wikified representation of the tweet content. This trend also attracted several research teams (Amir et al., 2014; Ling et al., 2015) in other domains.

However, according to the RepLab organizers based on the results there is no correspondence between performances, approaches and features used. Furthermore, none of these approaches were intended to be used by computer science non-specialist. This motivated us to investigate (i) the best possible performance of tweet-content based classifier and if we can expect further improvement by expanding the tweet with additional resources, (ii) to provide humanely comprehensible resources for decision support.

3 Approach

The purpose of this section is to improve understanding our approach, particularly regarding how we deal with textual contents. We start by providing the general idea. Further, we detail our method by describing the pre-processing procedure, introduce term weighting and expansion techniques.

To tackle the RDC, we propose a supervised classification method based on a threshold intersection graph computed over the discriminant bag-of-words (BoW) representation of each tweet. In this graph vertices represent tweets. Two vertices are linked by an edge if the corresponding tweets share at least one word (lexical relation). Edges are weighted here using a cosine similarity. Classes are learned from a labeled training set. A class is viewed as an aggregation of BoW of the tweets belonging to it. Then, we estimate the similarity of a given unlabeled tweet by comparing it to each class represented as BoW and rank tweets by decreasing cosine value. An unlabeled tweet is assigned a class if its similarity to this class is greater than a predefined threshold. Multiple classes can be assigned to a single tweet but in our case we only focus the class which obtained the best similarity. Overall, the classifier entirely relies on the graph of lexical similarities between tweets having common words.

3.1 Term Weighting

The features used by our proposals are words. To compute the similarity between tweets (edge weights) we consider n-gram (n=3) and skip-grams (bi-grams with gaps of length = 1). They compose the tweet discriminant BoW representation. BoW is built after the following pre-processing:

- words are lower-cased;
- stop-words links and punctuation are removed;
- the author name is added.

To weight terms we use the TF-IDF measure (Sparck Jones, 1972) combined with the Gini purity criterion (Torres-Moreno et al., 2013), as several works reported improvements using this association such as (Cossu et al., 2015). The Gini purity criterion $G(i)$ of a word $i$ is defined as follows:

$$G(i) = \sum_{c \in C} (p(c|i))^2 = \sum_{c \in C} \left( \frac{DF_c(i)}{DF(i)} \right)^2$$

where $C$ is the set of document classes and $DF_c(i)$ is the class-wise document frequency, i.e. the number of documents belonging to class $c$ and containing word $i$, in the training set. $G(i)$ indicates how much a term $i$ is spread over different classes. It ranges from $1/|C|$ when a given word $i$ is well spread in all classes, to 1 when the word only appears in a single class which means that this word has a strong discriminant power.

This factor is used to weight the contribution $\omega_{i,c}$ of each term $i$ in a class $c$ as (2):

$$\omega_{i,c} = DF_{(i),c}^\alpha \times \log\left(\frac{N}{DF_{(i),c}}\right)^\beta \times G_{(i)}^{\gamma}$$

We could have considered for instance Jaccard or any else similarity measure.

\textsuperscript{3}non-consecutive bi-grams

\textsuperscript{4}using short stop-lists from Oracle’s website (http://docs.oracle.com) for English and Spanish
3.2 Tweets Expansion

3.2.1 Word2Vec Model for Lexical Context

Literature often reports that the performances of statistical NLP approaches are limited by the data available to them. More precisely, these methods rely on the lexical proximity of training and test documents. In order to reduce the impact of the information loss carried by Out Of Vocabulary words (OOV) we project OOV into the known vocabulary in a Continuous distributed words representation (Bengio et al., 2003) (considered as a generalization engine). We used a Word2Vec (Mikolov et al., 2013) model which is learned by a Skip-gram neural-network. This network tries to maximize the following log probability (Mikolov et al., 2013):

\[
\frac{1}{N} \sum_{n=1}^{N} \sum_{i \neq j < c} \log \left( \frac{exp(i^n_j, w_i)}{\sum_{w=1}^{N} exp(i^n_j, w)} \right)
\]

where \( N \) is the number of tweets in the training corpus, \( w_0..w_N \) is the sequence of training words, \( c \) is the size of the training context.

Word2vec models were proved being able to capture syntactic and semantic relationship between words (Mikolov et al., 2013). They allow to measure similarity with simple geometric operations like sum and angle metrics. We trained a 600 dimension 10 context window multilingual (English+Spanish) Skip-gram model over RepLab’s background messages (Amigo et al., 2013) as (Amir et al., 2014) did supplemented by a large amount of easily available corpora.\(^\text{11}\) This trained model is then used as a generalization engine by other classifier, i.e. it finds for each OOV in the test sample the closest word in the Continuos distributed words representation which exists in the training vocabulary and has a sufficient purity (0.5) as defined in Formula (1).

3.2.2 Pseudo-Relevance Based Query Expansion

The principle of QE is to add new query terms to the initial query in order to enhance the users’ need formulation. Candidate terms for expansion are either extracted from external resources such as WordNet or from the documents themselves; based on their links with the initial query terms. In the latter types of methods, the most popular one is the pseudo-relevance feedback (PRF) (Buckley, 1995). The initial method was to add terms from relevant documents (Rocchio, 1971); since this information is not easily available, Buckley suggested to consider the first retrieved document as relevant and select candidate terms from these documents. PRF is now a common practice and used in many expansion methods (Carpineto and Romano, 2012).

We used several QE methods based PRF:

- Proximity Relevance Model (PRM) (Ermakova et al., 2016).
- Word-based Proximity Relevance Model (PRM-W) (Ermakova et al., 2016).

Terrier is an open-source state-of-the-art platform for highly effective information retrieval based on DFR models. The DFR models rely on the assumption that informative words are relatively more frequent in relevant documents than in others. The best-scored terms from the top-ranked documents are extracted. Terms are ranked using one of the DFR weighting models. We used the default parameter-free settings, but Rocchio’s QE mechanism can be also applied. We used the following DFR QE models:

- Kullback-Leibler (KL);
- Chi-square (CS);
- Bose-Einstein 1 (Bo1);
- Bose-Einstein 2 (Bo2).

PRM is integrated into the language model formalism that takes advantage of the remoteness of candidate terms for QE from query terms within feedback

\(^{11}\)enwik9, One Billion Word Language Modeling Benchmark, the Brown corpus, English GigaWord from 1 to 5, eswik, parallel es-en europarl
documents. In this approach the distance is computed in terms of sentences from the query terms and its combinations rather than in terms of tokens.

PRM-W is similar to PRM but the distance is calculated at word level. As in (Ermakova et al., 2016), we equalize the context length by multiplying by the average sentence length.

We selected 50 expansion candidates from 5 best scores documents from the English Wikipedia dump provided by INEX organizers (Bellot et al., 2014). The initial retrieval was performed by default retrieval L2 model from Terrier platform which is based on \( \text{TF} \cdot \text{IDF} \) measure with \( L^2 \) term frequency normalization. The other parameters were set as in (Ermakova et al., 2016).

3.2.3 Contextualization

Contextualization can be considered as adding new links between the nodes of the classifier graph. INEX/CLEF TC organizers defined context as a small readable summary which contains relevant information (called passages) extracted from Wikipedia and related to both the entity and tweet’s topic. Hereafter we consider context information as the outputs provided by systems that participated at the INEX/CLEF Tweet Contextualization task 2014 (Bellot et al., 2014).

Theses context were generated using a corpus based on a dump of the English Wikipedia from November 2012. Since notes and bibliographic references are difficult to handle, they were removed to facilitate the extraction of plain text answers as well as were also removed empty pages (having less than one section). We select information from the INEX output that will be added in the tweet bag of words. We selected the most relevant sentence as a short context (equivalent to the tweet length). In addition to the tweet content, the following selections were proposed:

- The most relevant sentence as a short context (equivalent to the tweet length);
- The 5 most relevant sentences as a long context;
- The 5 most relevant sentences are also evaluated alone in order to estimate if a context can stand in for a tweet.

We perform RDC with this new document and we look forward improvement of the classification results.

4 Evaluation Framework

In this section we provide a detailed description of the evaluation framework. We present the used data set, evaluation measures and the systems we compared.

4.1 Data Set

The corpus is a bilingual collection of tweets (Amigo et al., 2014) related to 31 entities from Automotive and Banking domains. For each entity 2 200 tweets covering a period going from the 1st of June 2012 to the 31st of December 2012 were extracted by querying Twitter with entities’ canonical names. The standard categorization provided by the Reputation Institute is used as a gold-standard. 15 489 tweets compose the training set, while the test set contains 32 446 tweets. Overall, RepLab test set vocabulary size is twice as big as the one of the annotated set. The organizers also provide more than 300 000 unlabeled tweets containing entities mentions. This set is considered as a large micro-blog text archive to build a related lexical context. It has also been considered with Active Learning approaches (Spina et al., 2015). We used this set to train our Word2Vec Model. Both train and test sets are annotated regarding one of the following reputation dimensions: Products/Services, Innovation, Workplace, Citizenship, Governance, Leadership, Performance and Undefined. As reported in Table 1, the “Products & Services” class represents around 50% of the data-set. Note that, Undefined is not a real class, but a label used by annotators to denote they were not able to assign a class with the information given in the tweet. As it is ignored by the Reptrak experts, this class was excluded from RepLab official results. Undefined may be viewed as a

<table>
<thead>
<tr>
<th>Label</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citizenship</td>
<td>2209</td>
<td>5027</td>
</tr>
<tr>
<td>Governance</td>
<td>1303</td>
<td>3395</td>
</tr>
<tr>
<td>Innovation</td>
<td>216</td>
<td>306</td>
</tr>
<tr>
<td>Leadership</td>
<td>297</td>
<td>744</td>
</tr>
<tr>
<td>Performance</td>
<td>943</td>
<td>1598</td>
</tr>
<tr>
<td>Products &amp; Services</td>
<td>7898</td>
<td>15903</td>
</tr>
<tr>
<td>Workplace</td>
<td>468</td>
<td>1124</td>
</tr>
<tr>
<td>Undefined</td>
<td>2228</td>
<td>4349</td>
</tr>
</tbody>
</table>

Table 1: Label distribution in the training and test set

\[ \text{http://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework} \]
noisy-class. In this paper, we report results of our approaches for:
- all classes including Undefined (+U);
- all classes excluding Undefined (-U - as well as RepLab official evaluation);

We selected the 3,000 last (across the time) tweets from the training collection to build a development set for parameters optimization.

INEX TC 2014 task may be viewed as complementary to CLEF RepLab and it was partially based on RepLab dataset. For the INEX Track 2014, the INEX’s organizers manually selected from the RepLab collection a set of 240 tweets based on their readability. These tweets, in English, have more than 80 characters and do not contain URLs in order to focus on content analysis. The entity name is used as an entry point into Wikipedia or DbPedia to give the contextual perspective. From these 240 tweets, 77 tweets match with the RepLab 2014 Dimensions Track test set.

4.2 Evaluation Measures

We compare our proposal to RepLab baselines and best submitted systems using RepLab official metrics. We use the absolute values from confusion matrix to calculate the accuracy of the text mining approach, that is to say average F-Score (AvgF) and Accuracy (Acc). Although accuracy is easy to interpret, it does not represent the informativeness of non-informative system under unbalanced test sets such as those returning all tweets in the same class (here all “Products & Services”).

4.3 Competitors

We conducted experiments with the following systems:
- **Naive**: baseline which assigns the most frequent class to a tweet.
- **SVM_base**: linear Support Vectors Machine (SVM) classifier with binary representation of tweet content (1 if the word occurs in the tweet, 0 otherwise).
- **Best_F**: best system according to F-Score participated in RepLab 2014. Best_F uses semantic expansion (Rahimi et al., 2014).
- **Best_Acc**: best system according to accuracy participated in RepLab 2014. In Best_Acc tweet enrichment is done via pseudo-relevant document (McDonald et al., 2015).
- **CRF**: linear Conditional Random Fields (Lafferty et al., 2001) represent log-linear models, normalized at the entire tweet level, where each word has an output class associated to it. CRF can localize specific positions in tweets that carry information and highlight continuous contextual information. In this setup the probability between words and classes for the whole tweet (of N words) is defined as follows:

\[
P(c^N_1 | w^N_1) = \frac{1}{Z} \prod_{n=1}^{N} \sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, s) \]

(4)

Log-linear models are based on M feature functions \(h_m\) computed at each position from the previous class \(c_{n-1}\), current class \(c_n\) and the whole observation sequence \(s\) (tweet). \(\lambda_m\) are the weights estimated during the training process and \(Z\) is a normalization term defined as:

\[
Z = \sum_{c^N_1} \prod_{n=1}^{N} \sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, s) \]

(5)

The tweets from the training set were used to train our CRF tagger with uni-gram (window’s length neighborhood of 2 words around the current word) and bi-gram features. Then the CRF tagged each word in every tweet and decision for the final tweet’s label is made by a majority vote.

- **SVM**: multi-class SVM trained with default parameters and the BoW representation of each tweet \(d\) (each term weight is computed as \(\frac{1}{N}\)).
- **COS**: similarity between the tweet BoW \(d\) and each class BoW \(c\) is computed as follows (6):

\[
\cos(d, c) = \frac{\sum_{i \in d, c \in c} \omega_{i,d} \times \omega_{i,c}}{\sqrt{\sum_{i \in d} \omega_{i,d}^2 \times \sum_{i \in c} \omega_{i,c}^2}} \]

(6)

where \(\omega_{i,:}\) is the frequency of the \(i\)-th term in \(d\). \(\omega_{:,i}\) is the frequency of the \(i\)-th term in \(c\).

- **TUNED**: while COS considers \(\alpha\), \(\beta\) and \(\gamma\) set to 1 by default (as it is described in Section 3), TUNED represents the cosine similarity with optimized parameters, i.e. \(\alpha\), \(\beta\) and \(\gamma\) values which

\[\text{http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html}\]
led to the best AvgF on the development set (see 3.1).

We applied QE (PRM, PRM-W, Bo1, Bo2, KL, CS) and TC (Context) methods to COS and TUNED, hereinafter referred to as -COS and -TUNED respectively. We also analyze the influence of the expansion based on Word2Vec models on SVM, COS, CRF and TUNED hereinafter called $W2V-\bullet$.

5 Results

5.1 Contextualization Evaluation

Generating elaborated contexts is costly task. Before considering to invest time in generating 32 000 contexts we set a limited evaluation protocol using INEX TC framework. We started with investigating the performance our NLP approach on tweets known to be readable, that is to say, contents that a machine can handle easily. RDC Performance on this small evaluation set is reported in Table 2. This experiment reveals interesting results: the systems perform better on this small set than on the whole set, meaning INEX selected tweets may be easier to classify. We see two major reasons to explain this. First the label distribution changed since Products/Services represent now less than 30% while it was 56% for the complete test set. The second explanation is that these tweets were manually selected for their readability and their length which mean that they are certainly easier to handle from a machine learning point-of-view. We can observe that the basic version of the classifier already performs well but the parameters adjustment provides additional information that is reliable to find information about the small classes. We then investigate how much we can improve our results on this sub-set by adding external context information provided by systems which participated to the INEX TC.

We first considered RDC performances considered the large context substituting the tweet content. The baseline (tweet content only) outperformed each proposed context. We then considered the tweet content embedded in its large context. Results were slightly better, but the baseline is still outperforming all other combinations. We finally experimented RDC using the short context added to the tweet content. Then, some context shown improvements of the RDC performances according to F-Score. Probably because the information added was really related to the tweet’s topic however accuracy remains lower than without additional data. Considering this, the short context seems to be better than a larger one, probably less noise is introduced regarding to the original content.

The Context context that has shown improvements regarding RDC is provided by one of the best systems according by both informativeness and readability that participated at the INEX/CLEF Tweet Contextualization task 2011-2014 (Ermakova, 2015). The system is based on the cosine similarity smoothed from the local context. This system takes into account named entity recognition, part-of-speech weighting and sentence quality measure. For each tweet, each INEX output proposes different sentences more or less related to the tweet that compose a context. In the next experiments we only consider this system as context generator.

5.2 Global Evaluation

Table 3 compares approaches performances according to average F-Score (AvgF) and Accuracy (Acc) over the test set for all classes including (+U) and excluding (-U) Undefined. We report here only results obtained with tweet enrichment for COS and TUNED since they demonstrated better scores than other classifiers according to F-score. As interesting results, we can see that the basic version of COS already outperforms SVM and CRF-based classifiers in terms of F-Score as well as the best systems participated in RepLab. However, according to Acc, CRF-based classifiers outperformed all other systems. It is also obvious that CRF models are much more robust to noise imported by Undefined class. The optimized weights obtained on the development set (TUNED) gave a light improvement according to F-Score and Accuracy. In contrast, AvgF and Acc for the contextualization-based systems are lower than the corresponding baselines. Word2Vector expansion slightly ameliorated results for only CRF baseline on -U data set. However, the introduction of the noisy class impaired the results whatever the system considered. The use of the expansion methods from the Terrier platform decreased performance of the baselines. In contrast, PRM and PRM-W methods showed the highest improvement. According to AvgF, the best results were obtained by the tuned co-

<table>
<thead>
<tr>
<th>System</th>
<th>AvgF</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>-TUNED</td>
<td>.650</td>
<td>.710</td>
</tr>
<tr>
<td>-COS</td>
<td>.560</td>
<td>.600</td>
</tr>
<tr>
<td>-COS (complete test)</td>
<td>.505</td>
<td>.735</td>
</tr>
<tr>
<td>SVM base (complete test)</td>
<td>.380</td>
<td>.622</td>
</tr>
</tbody>
</table>

Table 2: Performance Comparison over the INEX Tweets Ordered by F-Score.
Table 3: Dimensions detection performances. Best performances are highlighted in bold. Statistical significant improvements (averaged across entities) over the SVM(-U) (two-sided pairwise t-test p < 0.05) are denoted by *.

<table>
<thead>
<tr>
<th>Method</th>
<th>AvgF</th>
<th>Acc</th>
<th>AvgF</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRM-TUNED</td>
<td>.527*</td>
<td>.755*</td>
<td>.463</td>
<td>.645</td>
</tr>
<tr>
<td>PRM-W-TUNED</td>
<td>.523*</td>
<td>.750*</td>
<td>.458</td>
<td>.645</td>
</tr>
<tr>
<td>Bo1-TUNED</td>
<td>.503*</td>
<td>.740</td>
<td>.447</td>
<td>.635</td>
</tr>
<tr>
<td>Bo2-TUNED</td>
<td>.508*</td>
<td>.745</td>
<td>.443</td>
<td>.635</td>
</tr>
<tr>
<td>KL-TUNED</td>
<td>.504*</td>
<td>.740</td>
<td>.443</td>
<td>.635</td>
</tr>
<tr>
<td>CS-TUNED</td>
<td>.503*</td>
<td>.740</td>
<td>.443</td>
<td>.635</td>
</tr>
<tr>
<td>W2V-TUNED</td>
<td>.514*</td>
<td>.740</td>
<td>.451</td>
<td>.635</td>
</tr>
<tr>
<td>Context-TUNED</td>
<td>.479</td>
<td>.725</td>
<td>.421</td>
<td>.620</td>
</tr>
<tr>
<td>TUNED</td>
<td>.519*</td>
<td>.740</td>
<td>.456</td>
<td>.635</td>
</tr>
<tr>
<td>PRM-COS</td>
<td>.516*</td>
<td>.755*</td>
<td>.449</td>
<td>.645</td>
</tr>
<tr>
<td>PRM-W-COS</td>
<td>.514*</td>
<td>.750*</td>
<td>.447</td>
<td>.645</td>
</tr>
<tr>
<td>Bo1-COS</td>
<td>.467</td>
<td>.730</td>
<td>.410</td>
<td>.630</td>
</tr>
<tr>
<td>Bo2-COS</td>
<td>.474</td>
<td>.735</td>
<td>.414</td>
<td>.630</td>
</tr>
<tr>
<td>KL-COS</td>
<td>.468</td>
<td>.730</td>
<td>.411</td>
<td>.625</td>
</tr>
<tr>
<td>CS-COS</td>
<td>.468</td>
<td>.730</td>
<td>.411</td>
<td>.625</td>
</tr>
<tr>
<td>W2V-COS</td>
<td>.502*</td>
<td>.735</td>
<td>.435</td>
<td>.625</td>
</tr>
<tr>
<td>Context-COS</td>
<td>.466</td>
<td>.720</td>
<td>.406</td>
<td>.615</td>
</tr>
<tr>
<td>COS</td>
<td>.505*</td>
<td>.735</td>
<td>.446</td>
<td>.625</td>
</tr>
<tr>
<td>SVM</td>
<td>.469</td>
<td>.732</td>
<td>.461</td>
<td>.679</td>
</tr>
<tr>
<td>W2V-SVM</td>
<td>.468</td>
<td>.732</td>
<td>.456</td>
<td>.679</td>
</tr>
<tr>
<td>W2V-CRF</td>
<td>.492</td>
<td>.771*</td>
<td>.481</td>
<td>.761</td>
</tr>
<tr>
<td>CRF</td>
<td>.491</td>
<td>.760*</td>
<td>.483</td>
<td>.762</td>
</tr>
<tr>
<td>Best_F</td>
<td>.489</td>
<td>.695</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Best_Acc</td>
<td>.473</td>
<td>.731</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM_base</td>
<td>.380</td>
<td>.622</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Naive</td>
<td>.152</td>
<td>.560</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The performed experiments showed that contextualization does not improve the best lexical classifiers on RepLab although the provided context is much easier to read for reputation experts than the expansion terms provided by a QE system. Moreover, the expansion of already well readable tweets does not improve the classification performance. Therefore, tweet lexical content can be sufficient as long as enough training data is available. In contrast, effective query expansion techniques improved the results of various classifiers and do help to compensate for a lack of optimization. CRF-based systems appeared to be more robust to noisy data than other classifiers, although they showed lower F-score on the pure data set.

Finally, the thorough analysis of INEX TC run performances over the common subset of tweets from RepLab revealed a close relationship between sentence informativeness and efficient expansion for Profiling Reputation, and, consequently, an indirect way of evaluating informativeness. In an operational system we then face a dilemma as, on the one hand, we need to improve the classification performance and on the other hand, businesses want information to understand why a reaction to these contents is needed.

In future works, we will look more into details where the systems fail. We will also analyze the relationship between tweet readability and classification performance since readable tweets (i.e. long ones without URLs) may be easier to handle from a machine learning point-of-view as well as it is the case with humans.

ACKNOWLEDGEMENTS

This work was funded by French ANR project ImagiWeb (under ref. ANR-2012-CORD-002-01). The authors would like to thank Dr Eric Sanjuan, Dr Juan-Manuel Torres-Moreno and Pr Marc El-Beze.

REFERENCES


