Lexical Context for Profiling Reputation of Corporate Entities

Jean-Valère Cossu[1,2] and Liana Ermakova[3,4]
[1]*Vodkaster, Paris, France*
[2]*MyLI - My Local Influence, Marseille, France*
[3]*LISIS, Université Paris-Est Marne-la-Vallée, France*
[4]*Université de Lorraine, France*

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[1] or TripAdvisor[2], 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 includes 5 sub-tasks:

1. Filtering;
2. Reputation Polarity Detection;

[1]https://www.yelp.com
[2]https://www.tripadvisor.com/

Cossu, J-V. and Ermakova, L.
Lexical Context for Profiling Reputation of Corporate Entities.
DOI: 10.5220/0006284505670576
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 2*, pages 567-576
ISBN: 978-989-758-248-6

3. Topic Detection;
4. Priority Ranking;
5. Reputation Dimensions Classification.

In this research we mainly focus on the 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[3] 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[4] (Bellot et al., 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[5]?
- Is there a way to automatically evaluate the quality of these additional information?
- 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 these 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 (Peleja et al., 2014). Regarding politics, many projects have also recently been started up, such as POPSTAR in Portugal[6] (Saleiro et al., 2015) and IMAGIWEB in France (Velcin et al., 2014). In this context, RepLab made a significant impact in ORM by providing a large annotated data-set and defining new tasks such as extracting sets of messages requiring a particular attention from a reputation manager (Amigó et al., 2013) (that is to say the Priority Ranking task) to improve opinion analysis (Peetz et al., 2016).

Within RepLab, participants shown that it is possible achieved strong performances to Topic Detection using both unsupervised clustering algorithms and supervised classification methods. We can cite among them, Formal Concept Analysis (Cigarrán et al., 2016), Topic-Specific Similarity Function (Spina et al., 2014) and well known Latent Dirichlet Allocation to discover latent topics in tweets (Yang and Rim, 2014). Different approaches tackling the classification 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

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

[4]https://inex.mmci.uni-saarland.de/tracks/qa/

[5]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).

[6]http://www.popstar.pt

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' 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[7] 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[8] (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[9], 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:

$$G(i) = \sum_{c \in C} (p(c|i))^2 = \sum_{c \in C} \left(\frac{DF_c(i)}{DF(i)} \right)^2 \quad (1)$$

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 $\omega_{i,d}$ of each term i in a class c as (2):

$$\omega_{i,c} = DF_{(i),c}^{\alpha} \times log(\frac{N}{DF_{(i),c}})^{\beta} \times G_{(i)}^{\gamma} \quad (2)$$

where N is the number of tweets in the training set and the contribution $\omega_{i,d}$ of each term i in document d by replacing the word # of occurrences $DF_{(i),c}$ by $TF_{(i),d}$.

We assume that words are not equally informative. Thus, terms weighting can be improved by adding weights to the main features (α, β and γ) which will

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

[8] non-consecutive bi-grams

[9] using short stop-lists from Oracle's website (http://docs.oracle.com) for English and Spanish

be later called optimized parameters. The optimized values of these weights are obtained with a learn-to-rank approach using the development set. We then perform RDC with the expansion of initial tweets by traditional query expansion mechanisms as well as tweet contextualization considering both case, no parameters and optimized parameters. Details about the employed tweets expansion mechanisms are given in Subsection 3.2

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)[10], 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):

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-c<j<c, j\neq 0} log(\frac{exp(i^T_{w_{t+j}} o_{w_t}))}{\sum_{w=1}^{N} exp(i^T_w o_{w_t})}) \quad (3)$$

where N is the number of words 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 (Amigó et al., 2013) (as (Amir et al., 2014) did) supplemented by a large amount of easily available corpora[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 Continuous distributed words representation which exists in the training vocabulary and has a sufficient purity (0.5) as defined in Formula (1).

[10]Out Of Vocabulary words are words occurring in the test set while they were never seen in the training set.

[11]enwik9, One Billion Word Language Modeling Benchmark, the Brown corpus, English GigaWord from 1 to 5, eswik, parallel es-en europarl

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:

- Divergence From Randomness (DFR) models implemented in Terrier(Amati, 2003).
- Proximity Relevance Model (PRM) (Ermakova et al., 2016).
- Word-based Proximity Relevance Model (PRM-W)(Ermakova et al., 2016).

Terrier is a 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 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 *InL2* model from Terrier platform which is based on $TF - IDF$ measure with *L2* 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 (Amigó 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 [12] 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 provides 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 noisy-class. In this paper, we report results of our approaches for:

Table 1: Label distribution in the training and test set.

Label	Train	Test
Citizenship	2209	5027
Governance	1303	3395
Innovation	216	306
Leadership	297	744
Performance	943	1598
Products & Services	7898	15903
Workplace	468	1124
Undefined	2228	4349

- 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

[12] http://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework

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 testset.

4.2 Evaluation Measures

We compare our proposal to RepLab baselines [13] 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**)[14] 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:

$$P(c_1^N|w_1^N) = \frac{1}{Z}\prod_{n=1}^{N}\sum_{m=1}^{M}\lambda_m \cdot h_m(c_{n-1}, c_n, s)) \quad (4)$$

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). λ_m are the weights estimated during the training process and Z is a normalization term defined as:

$$Z = \sum_{c_1^N}\prod_{n=1}^{N}\sum_{m=1}^{M}\lambda_m \cdot h_m(c_{n-1}, c_n, s)) \quad (5)$$

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**: linear multi-class SVM[15] trained with default parameters and the BoW representation of each tweet d (each term weight is computed as (2)).
- **COS**: similarity between the tweet BoW d and each class BoW c is computed as follows (6):

$$cos(d,c) = \frac{\sum_{i\in d\cap c}\omega_{i,d}\times\omega_{i,c}}{\sqrt{\sum_{i\in d}\omega_{i,d}^2 \times \sum_{i\in c}\omega_{i,c}^2}} \quad (6)$$

where $\omega_{i,\bullet}$ is the frequency of the $i-$th term in $\bullet$.

- **TUNED**: while **COS** considers α, β and γ set to 1 by default (as it is described in Section 3) **TUNED** represents the cosine similarity with optimized parameters, i.e. α, β and γ values which 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−•**.

[13] The organizers provided two baselines in the RDC task. A Naive one that assigns the most frequent class to each tweet. A ML-based classification using a linear SVM for each entity with Bag-of-Word's (BoW) binary representation.

[14] Macro Averaged F-Score, based on Precision and Recall

[15] http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

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

Table 2: Performance Comparison over the INEX Tweets Ordered by F-Score.

System	AvgF	Acc
-TUNED	**.650**	.710
-COS	.560	.600
-COS (complete test)	.505	**.735**
SVM_base (complete test)	.380	.622

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 cosine classifier with the sentence-based PRM query expansion method. Since Contextualization introduces non-informative words to improve readability, it did not improve results which is not surprising regarding INEX organizers' conclusions. Indeed they mention that finding the right compromise between readability and informativeness remains the main issue of these systems.

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 *.

	-U		+U	
Method	AvgF	Acc	AvgF	Acc
PRM-TUNED	**.527***	**.755***	.463	.645
PRM-W-TUNED	.523*	.750*	.458	.645
Bo1-TUNED	.503*	.740	.445	.635
Bo2-TUNED	.508*	.745	.447	.635
KL-TUNED	.504*	.740	.443	.635
CS-TUNED	.503*	.740	.443	.635
W2V-TUNED	.514*	.740	.451	.635
Context-TUNED	.479	.725	.421	.620
TUNED	.519*	.740	.456	.635
PRM-COS	.516*	.755*	.449	.645
PRM-W-COS	.514*	.750*	.447	.645
Bo1-COS	.467	.730	.410	.630
Bo2-COS	.474	.735	.414	.630
KL-COS	.468	.730	.411	.625
CS-COS	.468	.730	.411	.625
W2V-COS	.502*	.735	.435	.625
Context-COS	.460	.720	.406	.615
COS	.505*	.735	.446	.625
SVM	.469	.732	.461	.679
W2V-SVM	.468	.732	.456	.679
W2V-CRF	.492	**.771***	.481	.761
CRF	.491	.769*	**.483**	**.762**
Best_F	.489	.695	-	-
Best_Acc	.473	.731	-	-
SVM_base	.380	.622	-	-
Naive	.152	.560	-	-

6 CONCLUSIONS

In this paper we introduced a supervised classification method for RDC task based a threshold intersection graph. Overall, the literature claims that the performance of statistical NLP approaches are limited by the amount and quality of text available to them. We observed statistical NLP perform significantly better on a selection of readable tweets from RepLab, than on the whole RepLab collection. Consequently, we compared the performance impact of parameters optimization against two lexical context expansions: one using typical Query Expansion techniques; a second one based on the insertion of most informative Wikipedia sentences provided by the state-of-the-art contextualization system that demonstrated the best results on INEX TC track.

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 relation 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

Amati, G. (2003). *Probability Models for Information Retrieval Based on Divergence from Randomness: PhD Thesis*. University of Glasgow.

Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., and Spina, D. (2014). Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 307–322.

Amigó, E., De Albornoz, J. C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., De Rijke, M., and Spina,

D. (2013). Overview of replab 2013: Evaluating online reputation monitoring systems. In *CLEF 2013*.

Amir, S., Almeida, M., Martins, B., Filgueiras, J., and Silva, M. J. (2014). Tugas: Exploiting unlabelled data for twitter sentiment analysis. *Proceedings of SemEval*, pages 673–677.

Anwar Hridoy, S. A., Ekram, M. T., Islam, M. S., Ahmed, F., and Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2(1):8.

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., and Tannier, X. (2014). Overview of INEX tweet contextualization 2014 track. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings*, pages 494–500. CEUR-WS.org.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155.

Buckley, C. (1995). Automatic query expansion using SMART : TREC 3. In *Proceedings of The third Text REtrieval Conference (TREC–3). NIST Special Publication 500-226*, pages 69–80. National Institute of Standards and Technology.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50.

Cigarrán, J., Castellanos, Á., and García-Serrano, A. (2016). A step forward for topic detection in twitter: An fca-based approach. *Expert Systems with Applications*, 57:21–36.

Cossu, J.-V., Janod, K., Ferreira, E., Gaillard, J., and El-Bèze, M. (2015). Nlp-based classifiers to generalize experts assessments in e-reputation. In *Experimental IR meets Multilinguality, Multimodality, and Interaction*.

Deveaud, R., Mothe, J., and Nia, J.-Y. (2016). Learning to rank system configurations. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2001–2004. ACM.

Ermakova, L. (2015). A method for short message contextualization: Experiments at clef/inex. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 352–363. Springer.

Ermakova, L., Mothe, J., and Nikitina, E. (2016). Proximity relevance model for query expansion. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1054–1059, New York, NY, USA. ACM.

Karisani, P., Oroumchian, F., and Rahgozar, M. (2015). Tweet expansion method for filtering task in twitter. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 55–64. Springer.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ling, W., Chu-Cheng, L., Tsvetkov, Y., and Amir, S. (2015). Not all contexts are created equal: Better word representations with variable attention.

Malaga, R. A. (2001). Web-based reputation management systems: Problems and suggested solutions. 1(4):403–417.

McDonald, G., Deveaud, R., McCreadie, R., Macdonald, C., and Ounis, I. (2015). Tweet enrichment for effective dimensions classification in online reputation management. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 654–657.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, C. (2016). Why evan williams of twitter demoted himself - the new york times, http://www.nytimes.com/2010/10/31/technology/31ev.html. Last accessed on 2017-02-18.

Peetz, M.-H., de Rijke, M., and Kaptein, R. (2016). Estimating reputation polarity on microblog posts. *Information Processing & Management*, 52(2):193–216.

Peleja, F., Santos, J., and Magalhães, J. (2014). Reputation analysis with a ranked sentiment-lexicon. In *Proceedings of the 37th SIGIR conference*.

Qureshi, M. A. (2015). *Utilising Wikipedia for text mining applications*. PhD thesis.

Rahimi, A., Sahlgren, M., Kerren, A., and Paradis, C. (2014). Stavicta group report for replab 2014 reputation dimension task. In *CLEF (Working Notes)*, pages 1519–1527. Citeseer.

Rocchio, J. (1971). Relevance feedback in information retrieval. In *The SMART Retrieval System*, pages 313–323.

Saleiro, P., Amir, S., Silva, M., and Soares, C. (2015). Popmine: Tracking political opinion on the web. In *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, pages 1521–1526. IEEE.

Sánchez-Sánchez, C., Jiménez-Salazar, H., and Luna-Ramírez, W. A. (2013). Uamclyr at replab2013: Monitoring task. In *CLEF (Working Notes)*. Citeseer.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Spina, D., Gonzalo, J., and Amigó, E. (2014). Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 527–536. ACM.

Spina, D., Peetz, M.-H., and de Rijke, M. (2015). Active learning for entity filtering in microblog streams. In *SIGIR 2015: 38th international ACM SIGIR conference on Research and development in information retrieval*.

Torres-Moreno, J.-M., El-Bèze, M., Bellot, P., and Béchet, F. (2013). Opinion detection as a topic classification problem.

Twitter (2016). https://about.twitter.com/company. Last accessed on 2017-02-18.

Velcin, J., Kim, Y., Brun, C., Dormagen, J., SanJuan, E., Khouas, L., Peradotto, A., Bonnevay, S., Roux, C., Boyadjian, J., et al. (2014). Investigating the image of entities in social media: Dataset design and first results. In *LREC*.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2014). A linguistic approach for determining the topics of spanish twitter messages. *Journal of Information Science*, page 0165551514561652.

Yang, M.-C. and Rim, H.-C. (2014). Identifying interesting twitter contents using topical analysis. *Expert Systems with Applications*, 41(9):4330–4336.