E-reputation monitoring on Twitter with active learning automatic annotation

Jean-Valère Cossu, Marc El Bèze, Juan-Manuel Torres-Moreno, Eric Sanjuan

E-reputation monitoring on Twitter with active learning automatic annotation

Jean-Valère Cossu, Marc El-Bèze, Juan-Manuel Torres-Moreno and Eric Sanjuan

LIA/Université d'Avignon et des Pays de Vaucluse
** 39 chemin des Meinajaries, Agroparc BP 91228, 84911 Avignon cedex 9, France
firstname.name@univ-avignon.fr

Abstract. Opinion and trend mining on micro blogs like twitter recently attracted research interest in several fields including Information Retrieval and Machine Learning. This paper is intended to develop a so-called active learning for automatically annotating French language tweets that deal with the image (i.e., representation, web reputation) of entities : such as politicians, celebrities, companies or brands. Our main contribution is the methodology followed to build and provide an original annotated French data-set expressing opinion on two French politicians over time. Since the performance of natural language processing tasks are limited by the amount and quality of data available to them, one promising alternative for some tasks is the propagation of pseudo-expert annotations. The paper is focused on key issues about active learning while building a large annotated data set, from noise introduced by humans annotators, abundance of data and the label distribution across data and entities.

1 Introduction

In recent years, most of the opinion mining work focused on products but in a world of online networked information, where its control has moved to users, each act of a public entity is scrutinized by a powerful global audience. This requires new reputation management tools and strategies.

We can understand reputation as the general recognition by other people of some characteristics or abilities for a given entity. Specifically, in business or politics, reputation comprises the decisions taken with the perception of population. Reputation affects attitudes like satisfaction or trust, and drives behavior like support.

In turn, reputation analysis is the process of tracking, investigating and reporting an entity's actions and other entities' opinions about those actions. Currently market research using user surveys is typically performed and traditional reputation analysis is a costly task when done manually. However, the rise of online social media such as blogs and social networks and the increasing amount

** http://lia.univ-avignon.fr/

of user-generated contents in the form of reviews, recommendations, ratings and any other form of opinion, has led to the creation of an emerging trend towards online reputation analysis. It has become an interesting way to process large amount of opinions about entities but in the case of tweets there are not explicit ratings to be directly used in an opinion processing.

Although significant advances have been made [1], analyzing reputation (or image) about companies and individuals is a hard problem requiring a complex modeling of these entities (e.g. company, politician). By reputation we mean a structured and dynamic representation emitted by an entity or reflected in people's opinions about an entity. This is a real challenge not only to deal with specific requirements in information retrieval or opinion mining, but also to understand important issues in political science. Politics have already been addressed in previous works but mostly in English or Spanish [17], [20], [28], [25] and to the extend of our knowledge, nothing in French till now from a machine learning perspective. One of our targets is the reputation of French politicians, especially the two main candidates in the last French presidential election in May 2012, Nicolas Sarkozy (former president) and François Hollande (current president). This lead to build a annotated seed data-set with the involvement of specialists in political science. The annotations have two main aspects: the so-called polarity for reputation and the topic detection. In other words we want to link the opinion expressed with a specific attribute of the entity.

Reputation polarity is substantially different from standard sentiment analysis, since both author, facts and opinions have to be considered. The systems goal is to find what implications a piece of information may have on the reputation of a given entity (does the tweet have negative/positive implications for the reputation of the entity?), regardless of whether the message contains an opinion or not. Multilingual aspects, cultural factors and context awareness are among the main challenges of sentiment natural language text classification. In micro blogging, due to the 140 characters limit, messages are often allusive with few words making the task even harder.

This paper is intended to develop an efficient active learning algorithm for annotating French tweets that deal with the reputation of politicians, celebrities, companies or brands. We apply our algorithm to build an original annotated French data-set expressing opinion on two French politicians over time. Our approach relies on the propagation of a reduced set of pseudo-expert annotations among large collections of tweets. Our propagation approach deals with three key issues about active learning while building a large annotated data set:

- identify and remove noise introduced by humans annotators,
- use data abundance,
- harmonize label distribution across data and entities.

The rest of the paper is organized as following. §2 gives a focused overview of related work. In §3 we study the main characteristics of crow-sourced annotated tweets about politics. In §4 we propose a novel active learning algorithm for bias correction to improve annotation quality and to increase the final amount of labeled data. §5 is devoted to a thoroughly evaluation of our algorithm: after

pointing out that classical cross-validation is not enough accurate (it gives too high results on harmonized sets), we introduce a more severe protocol based on temporal prediction. Finally, §6 gives some conclusions about our work and opens several perspectives.

2 Related work

2.1 Tweets mining

Previous works on reputation monitoring in tweet collections and streams have been done to extract sets of messages requiring a particular attention from a reputation manager [1]. For example, recent contributions to this issue on Twitter data have been done in the context of the last editions of Replab [1] [1] and TASS [25] [2] where the lab organizers provide a framework to evaluate Online Reputation Management systems on Twitter. Participants have achieved topic detection using unsupervised clustering algorithms but most of the proposed contributions mainly rely on supervised classification methods. Clustering based on tweet content similarity have been proposed based on term expansion, term co-occurrences or wikified representation of the tweet content. Different ways to tackle the issue were to use Social Network Analysis for tweets clustering [5] and links between contents and meta-data [21].

2.2 Data building

As crowdsourcing is an increasingly popular, collaborative approach for acquiring annotated corpora. The background literature [26] puts the focus on central points which describe a current research issue. Although the use of paid-for crowdsourcing approach is intensifying, the reuse of annotation guidelines, task designs, and user interfaces between projects is still problematic, since these are usually not made available for the community despite their important role in result quality. A challenge for crowdsourcing projects is that the cost to define a single annotation task remains quite substantial. The design comprises different processes such as data selection, formal definition and instantiation of the annotated concepts. One of main project's contribution is to build and provide to the research community an original annotated data-set.

In addition to a data-set, we provide a full open-source annotation platform [3] and the design. This design comprises different processes such as data selection, formal definition and instantiation of an image. Under the given circumstances of crowdsourcing websites, it is more useful to select outstanding submissions that offer an interesting, unusual or particularly revealing set of circumstances. Literature is also full of innovative approaches about definition of crowdsourcing success, how to evaluate the results and the application of text

[1] http://www.limosine-project.eu/events/replab2013
[2] http://www.daedalus.es/TASS2013/about.php
[3] http://dev.termwatch.es/~imagiweb/index.php

mining approaches. Many recent researches focused on the reliability and applicability of crowdsourcing annotations for natural language processing [27]. In our case text mining is not only applied to handle the issue of semi-supervised annotation but also to fulfill a semi-supervised selection of the messages we want to submit for manual annotation. Previous works using so-called active learning [22] have been done to automatically build high-quality annotated data-sets on twitter monitoring in the last editions of Replab [1] or TASS [25]. In [13] and [4] the authors presented various studies on machine learning architectures and active learning applied to spoken language processing applications. Most part of research projects leave behind them small annotated corpora and big amount of unlabeled data. The small data set can be used as bootstrapping for systems [11] but how can we exploit the remaining unlabeled set ? The idea to exploit the unlabeled examples by adding them the labeled data has been well studied in the last years [19], [6], [18], [12], [23]. By re-training models on the fly, experiments in [15] are also inspired by this principle.

As manual annotation is costly work we deviously use these state-of-the-art approaches to build and improves the data-set.

3 Dataset and descriptive Statistics

This study exploits text mining to analyze a real-world data sample from Twitter. For the semi-automated step of evaluating the quality of submissions by text mining, clustering is implemented and compared to real-world results, that is to say expert committee decisions. Our main objective is to use annotated data to detect the image of a given entity with machine learning techniques. An important fact is that the image is not static and may change in time. To achieve this objective, we need a huge amount of tweets to release a deep, complete and reliable politics analysis over the time but also to validate our algorithms. Get involved on a such annotation campaign is not possible in financial terms and delay so it has been decided that both NLP and politic researchers will work jointly in a so-called active learning process. The idea of collecting annotations from volunteer contributors has been used for a variety of tasks and the advantages of crowdsourcing over expert-based annotation have already been discussed elsew here [26]. But although designing such a data-set of training examples has proven quite an interesting challenge [1], [25], it is still expensive and never fully accurate.

3.1 Statistics on the annotated set

Here we provide statistics about the data-set (more detailed statistics are available in [24]). To handle the subjectivity of annotators, we allowed a tweet to be annotated at most three times by different annotators. It also happens that a same content (in case of retweet) has been annotated several times by the same annotator. On the whole data-set the annotator's consistency is estimated to be around 80%. The polarity levels vary from very positive (*positive*) to (*negative*)

Data: Large amount of unlabeled tweet
Result: Large amount of labeled tweet
Small amount of tweets are manually labeled;
while *Not enough labeled data or insufficient classifiers' performance* **do**
  Build models with the labeled data;
  Classify a subset of unlabeled data;
  Select and send a sample of automatic classification outputs for manual confirmation;
  if *Automatic classification is sufficient* **then**
    Annotation for the whole data-set;
  else
    Go back to the beginning with more labeled data for learning;
  end
end

Algorithm 1: Annotation process

very negative opinions, with a neutral opinion (used for facts reports) there is also an ambiguous opinion for undecidable cases. The data-set consists of 11.527 manual annotations expressing the opinion on describing two French politicians over time, 5.278 tweets for François Hollande (**FH**) and 6.241 tweets for Nicolas Sarkozy (**NS**). 7.283 unique tweets (6.369 unique contents) are annotated, of which 48% are annotated only once, 46% twice, 6% three or more times. Is this enough ? How much are these examples really informative ?

As mentioned before, one difficulty is to have enough data and sufficient information to build models for employing machine-learning approaches. And despite the recent advances and good practical results, improvements remain to be achieved. How much is enough is still an open question. Some approaches have been proposed so far to apply recommendation systems's adaptation to re-build models on the fly but have shown only small improvements, not always significant.

Opinions for a reasonable analysis, we made two assumptions very positive and positive opinions are treated as "positive" (same for negative) and ambiguous opinions are considered as neutral. On the whole data-set opinions are biased to the negative with a slight difference between the two entities; for example, 47% of the opinions about **NS** are negative for 20% positive and around 53% are negative for **FH** while 14% are positive. The neutral class distribution is equivalent for both candidates with 32%. However, in the period just before the election (mid-May 2012), the negativity about FH decreases as 41% while that of **NS** increase to 52%. After the election (June to December 2012), the negativity about **FH** increases dramatically to 72% as positivity collapses to 5%. Per month distributions are resumed in Tables 1. This justifies the necessity of temporal analysis related to the image, with well-split time periods.

Table 1. Polarity distribution across the time on the annotated set from March to December 2012

	Hollande			Sarkozy		
Date	Positive	Neutral	Negative	Positive	Neutral	Negative
March	.27	.28	.45	.18	.36	.46
April	.25	.36	.39	.20	.34	.46
May	.25	.34	.41	.19	.29	.52
June	.10	.31	.59	.40	.50	.20
July	.13	.35	.52	.24	.25	.50
August	.8	.31	.61	.23	.30	.46
September	.8	.32	.60	.26	.33	.42
October	.10	.36	.54	.21	.33	.46
November	.7	.34	.59	.17	.32	.52
December	.5	.23	.72	.20	.31	.49

Annotator bias The annotator low-confidence indicator was only used for 10% of annotations and mostly for the "ambiguous" polarity level. As it was not properly used, we reconsidered the quality of mainly non-expert annotations on different aspects. We assume here that we do not explore the idea of recalibrate annotator judgment to more closely match expert behavior or to exclude some annotators from the process.

Disagreement : Language and Concept interpretation the manual annotations may reflect the subjectivity of the annotators, because of the granularity of the labels. Analyzing the annotation disagreement among annotators for each tweet would help us to understand better the opinion properties.

Since annotators may have different points of view on the same document, we try to analyze the problem at the content level. While for a machine a word sequence will match a unique model or a weighted number of models, the language acquisition skills of humans result from a multidimensional experience. The annotation is dependent from annotator's language acquisition. Considering this, by taking a closer look at the annotations from the text level, we can now observe more severe disagreement for an unique content.

This come from one part of natural language variability, the background knowledge of the annotator can make him interpreting an hidden meaning of the message, while others did not notice the irony. For example a tweet about Chuck Norris facts, to criticize the small size of NS people used to say that Chuck Norris cannot see him since he never looks down.

From a second part of the concept variability to illustrate this case a tweet related to the case Sarkozy-Kadhafi has been correctly tagged as ethic by the two annotators, but the chosen sub-aspect differs (ethic:honesty vs. ethic:case). A typical example for polarity can be a tweet that describes the result of a public poll. If the poll is in favor of a candidate, some annotators give a positive (resp. negative) polarity while others give a neutral polarity since they consider this information as a fact.

Things become interesting when looking at how annotators labeled a repeated content. For each content annotated more than five times we can observe that there is at least one annotation different from the others.

As these cases appear several times, it illustrates the fact that different aspects can be selected depending on the individual point of view but if also offers us the possibility to see the trends of an annotator.

For the learning step, instead of misleading the automatic algorithms, we can consider that this situation will reflect the diversity of interpretation. We could consider that a message convey two messages (eg. two topics with each a different opinion) by the multi-label [**?**]) but we made the questionable assumption that one tweet = one opinion and one topic. But when it comes to the evaluation set it is critical to agree about only one reference.

Data Diversity when working with learning and data mining on text contents we have to keep a high variability in the data distribution.

To prevent falling in a biased distribution that we will lead us in over-training and the overvalue our systems. And to maximize the effectiveness of the annotation step by having certified label on the largest vocabulary as possible.

This step can be seen as text processing since each content is cleaned in order to detect duplicates messages and ignore them for the further annotation steps. For a more focused work on aspects (eg. statistics about voting) we keep a track of all duplicates in order to propagate the annotation.

3.2 Harmonization

It is still possible to estimate the task difficulty with inter-annotation agreement measures such as Cohen's Kappa [7] but once disagreements have been identified what can be done ? As, in our case, each tweet has been annotated from one to three times and as we before noted severe disagreements at text level we have chosen a majority-base rule system. For each annotated content with divergent annotations, we select when it's possible the human annotation getting a relative majority according to

$$\text{Label frequency} > 1/\text{Number of labels} \tag{1}$$

For all remaining cases of disagreements (no majority) and orphan tweets (in terms of annotations) we now work at the user level (as shown in figure 1). An important aspect in social networks is the possibility for users to answer each other building this way their own network. This can be an extra feature to consider that a user belongs to a group or has the same opinion (or aspect) as the person to whom he responds or re-tweets. Considering the political dimensions of the data set, we assume that in a short time window an important contributor (in terms of messages) often expressing its revulsion about one candidates cannot find something positive in only one message. This process can be seen as smoothing the user point of view, even we if know that this is assumption is not always verified. For these users besides the correction in the train-set, the

smoothing is characterized by the addition of a negative tag in the bag-of-words of the future tested tweets, this way by looking at the past of this user we penalize the contribution the non majority class without closing doors to a further change in user's mind. As time goes, on we will return on the premise that one user has only one opinion.

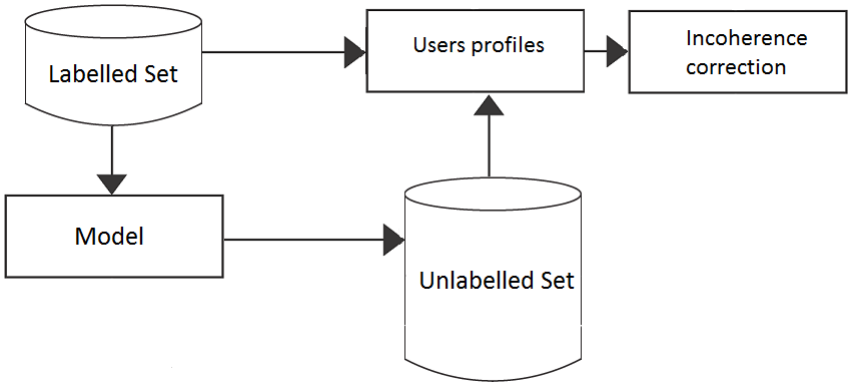

Fig. 1. Annotation errors corrections with users profiles

3.3 Labeled test set

The test set is composed of the 3 lasts months of the labeled set corresponding to 800 tweets (unique contents) for each candidate. Tweets distributions are resumed in the last lines of Table 1.

3.4 Unlabeled train set

The unlabeled set is composed of 65.968 tweets (51.020 unique contents) only concerning FH extracted with the following query "hollande OR @fhollande OR holande OR FH OR FH2012 lang:fr". Around 3.000 tweets have randomly been selected each month over 21 months from January 2012 to this December 2013.

4 Active learning : Issues and Challenges

We have seen before that human annotation of language features and concept are prone to human errors [3] [14]. That need to be considered in the model learning process. To build our models, the quality of manual annotation is critical. Selecting salient data streams or patterns (that are crucial to improve the performance of the systems) is an important issue. Something important before handle a large unannotated data set is to be sure about the training set reliability. We assume that the objective is not to build the most reliable data-set in the meaning of a particular aspect but to build a consistent data-set regarding to the analyze we want to do after.

4.1 Methodology

Classifiers For the purpose of this experiment and following the background literature we use both, normalized (tf) and inverse term frequencies (tf-idf), as features to create bag-of-words n-grams models to feed our classifiers (among those described in [9]) which we added HMM [29].

- Jaccard and Cosine distance (TF-IDF-Gini), based on a Jaccard (and Cosine) similarity to measure the distance between the bag-of-terms of a tweet and the whole bag built for each class and ranks tweets according to their similarity.
- KNN with discriminant features, matches a tweet in the test set with the k most similar tweets of the training set. Similarity is computed with Jaccard similarity on discriminant bag-of-words computed on tweet content and metadata (author, entity). k (equal to 6) has been fixed by cross-validation on the training set.

The robustness of processing also needs to be improved by being able to consider documents with low-level language quality or domain specific syntax and lexicon. Improving the robustness of the models by providing more efficient processing of noisy data is likely to be better than trying to correct or normalize the texts.

Metrics The absolute values from confusion matrix are used to calculate common metrics which measure the accuracy of the text mining approach, that is to say Precision, Recall and Macro and μF-score (equivalent to accuracy). Although it is easy to interpret, it is easy to be cheated under unbalanced test sets. For instance, returning all tweets in the same class (all "*NEGATIVE*" in our case), may have high accuracy if the set is unbalanced. To measure the data homogeneity and prevent this bias Reliability and Sensitivity [2] are also computed. Reliability and Sensitivity (R&S)can be seen as precision and recall of relationships. In brief, R&S computes the precision and recall of relationships ($<$, $>$, $=$) between the test documents produced by the systems with respect to the references set. R&S is computed for each document relationships and averaged. R&S is strict with respect to other measures: a high score according to R&S ensures a high score according to any traditional measure. In return a low score according to a particular measures infer a low R&S score, even when the system is rewarded by other measures.

4.2 Harmonization

Committee-based validation Unlike [8] and [10], we have two definition of the committee-based validation. In one part, the committee is composed by several classifiers. In the other part, the committee is a sort of very light supervision where domain non-specialist check different random sample of systems outputs to validate the process. Some studies worked well in the first direction,

such as [16] where the authors obtained 2- to 30 fold reductions of the amount of human annotation needs for text categorization. It is hard to find back the real informative examples (from the non informative ones), and the fact that it will only be possible to annotate content similar to labeled ones is an important drawback. It becomes a major issue if we care about data representativeness.

Correcting training examples with machine learning When it is not possible to select a majority label we train several classifiers described above and try to self annotate (as described in figure 2) the corpora in a leave one out process. A wide number of methods have already been explored to correct the bias of annotators and an important fact here is that we do not consider annotations as gold reference and we allow us to question them. We assume that classifiers outputs are considered as several additional referees for the committee-based validation at the same level as humans annotators and we came back to the previous step. We allow case of having multiple annotators but none of them is having the true label which systems agreed on. To make tables easier to understand we chose to present results from only two basic but complementary ones Cosine and HMM.

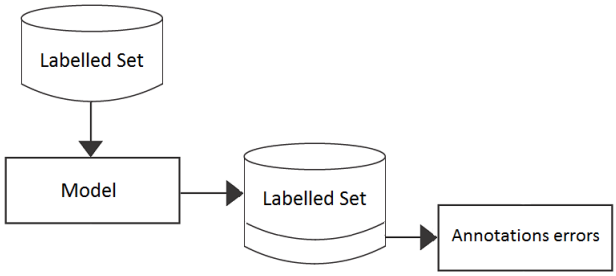

Fig. 2. Annotation errors detection with machine learning

Table 2. Annotation incoherence detection with cross-validation

		Hollande			Sarkozy		
Method	Data-set	F-Score	$\mu F - Score$	R&S F-Score	F-Score	$\mu F - Score$	R&S F-Score
Cosine	Original Set	.84	.87	.80	**.84**	**.83**	.83
HMM	Original Set	**.86**	**.88**	**.86**	**.84**	**.83**	**.86**
Cosine	Corrected Set	**.99**	**.99**	**.99**	**.99**	**.99**	**.99**
HMM	Corrected Set	.98	**.99**	.98	.97	.98	.98

In table 2 we can observe that for the original set classifiers are not able to find back the correct label for a part of the set. For these classifications errors we distinguish several cases :

1. All system agreed on a label different from the human annotations;
2. A majority agreed on a label different from the human annotations;
3. No agreement;

Based on the majority rule expressed above, we now consider for the two first cases (around 60%) the prediction of the classifiers as the new "reference" annotation for the tweet. In the last case tweets are submitted again for human verification. It is interesting to notice that except for some ironic tweets, after the correction classifiers are now able to find the correct label for a very high majority of tweets.

4.3 Expansion : harmonization using unlabeled data

Now that the training set has been corrected, we can use our classifiers to annotate a large set of unlabeled messages. The unlabeled examples can be used with unsupervised or supervised learning methods to improve the classification performance and the correction of the labeled examples by applying the above rules according to a principle of homogeneity at content and user level.

4.4 Outliers detection criterion

Outliers are examples that differ from the rest of the data, in terms of agreement or content.

Agreement we consider here outliers as tweets that neither systems or annotators agreed on the same label. Thses outliers will be ignored because manually labeling them may not help to build the models or to understand an issue since they are not likely to be useful for future observations.

A second interpretation of outliers is to respectively consider tweets for which every system agreed on the same label, as reliable and by adding them in the labeled set before iterating. At this step, as we consider them reliable enough to be used as models, theses so-called "outliers" also trend to stop being candidates for the next manual annotation step.

A sample (around 7.000 tweets) will manually labeled (as summarized in figure 3) in order to measure the process efficiency, these new annotation will also help to iterate the process.

Content this time we consider here outliers as tweets having the worse similarity (according to Jaccard discriminant bag of words representation) with the labeled set. Although we can measure the performance of the systems can be sure that in terms of homogeneity with most similar tweets will converge around a common label. But as they are not likely to be observed how can we be sure about a whole tweets cluster which is far from the corpus ?

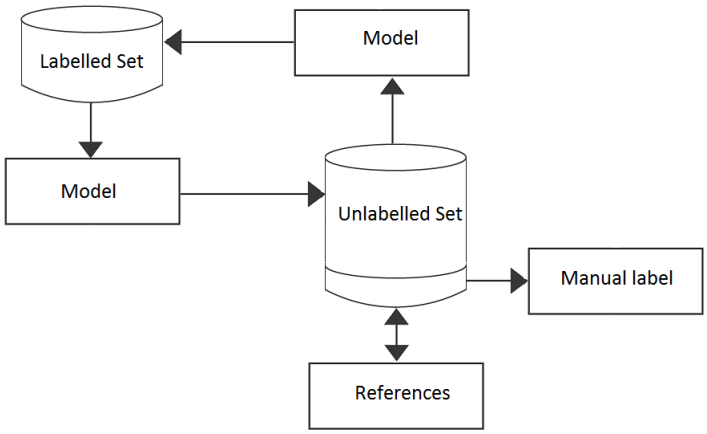

Fig. 3. Summary of the learning process

5 Evaluation

After all changes introduces by the process described above, the polarity distribution from the original set can be altered. In the period after the election (*June to December 2012*), the negativity about **FH** increases dramatically to 79% near the end of the year, as resumed in Table 3.

Table 3. Polarity distribution across the time on the annotated set after harmonization procedures

	Hollande			Sarkozy		
Date	Positive	Neutral	Negative	Positive	Neutral	Negative
March	.28	.23	.49	.21	.24	.55
April	.29	.30	.42	.17	.35	.49
May	.25	.30	.44	.19	.31	.50
June	.9	.27	.64	.19	.27	.54
July	.13	.30	.57	.31	.22	.47
August	.8	.28	.64	.25	.28	.47
September	.9	.26	.65	.27	.33	.40
October	.10	.32	.58	.19	.30	.51
November	.6	.29	.65	.17	.27	.56
December	.4	.17	.79	.19	.30	.52

5.1 Cross Validation

In this step we consider the following cross validation process to evaluate the quality of the systems annotation on the unlabeled set. We use the annotated set

to build models to automatically annotate the unlabeled set. Then we rebuild our models from the new labeled set in order to classify by the annotated set. As we do not yet have unlabeled set for **NS** results showed in table 4 with both corrected and uncorrected version of the data-set only concerns **FH**.

Table 4. Performances on 3-polarities classification task with a cross-validation process on the training set **(FH)**

Method	Train	F-Score	μ F-Score	R&S F-Score
Cosine	Original	.46	.53	.41
HMM	Original	.68	.73	.65
Cosine	Corrected	.50	.56	.44
HMM	Corrected	.73	.79	.70

As we have said the polarity task in this context is a difficult task and also as seen before in Section 4.2 the class distribution is prone to variations across the time. The results we just observed compared to the task difficulty are too good to be really representative of our process efficiency.

5.2 Temporal Expansion

We study the impact of the harmonization process described above on the results of classifiers on the last tweets of the data-set considered here as test set. In other words, we consider improvement in the output of polarity assignment to evaluate the gain offered by the harmonization process. Performances of our systems with the basic training-set (with only "*majority rule*" corrections) are reported in Table 5. We compare the performances of the classifiers with a corrected version of the training set.

Table 5. Performances on 3-polarities classification task with different training sets

		Hollande			Sarkozy		
Method	Train	F-Score	μ F-Score	R&S F-Score	F-Score	μ F-Score	R&S F-Score
Cosine	Original	.37	.60	.33	.40	.46	.38
HMM	Original	.42	.60	**.40**	.45	.48	.42
Cosine	Corrected	.44	**.69**	.36	.43	.51	.39
HMM	Corrected	**.47**	.66	**.40**	**.51**	**.55**	**.45**

5.3 Tiny class problem and propagation

In Table 6 we can observe the result of our systems now using as training the unlabeled set on which with propagate using our classifiers the annotation from the labeled set. The better performances are partly explained by the significant

difference of size between two training sets. In the next experiment, we combine these new labeled tweets to the original labeled set and observe significant improvement on the polarity results especially for positive and neutral tweets. This result proves that our propagation do contain relevant information that improves the polarity classification and that was missing on the original set. It is also important to notice that we do not observe a significant gain with the propagation of the corrected data (**CP**) compared to the raw set propagation (**Raw**).

Table 6. Performances on 3-polarities classification task with different propagated training sets (FH)

Method	Train	F-Score	μ F-Score	R&S F-Score
Cosine	Raw propagated	.42	.60	.38
HMM	Raw propagated	.46	.65	.45
Cosine	CP	.42	.62	.39
HMM	CP	.47	.67	.46
Cosine	Corrected + SP	.46	.64	.41
HMM	Corrected + SP	**.51**	**.70**	**.47**
Cosine	Corrected + TP	.44	.66	.41
HMM	Corrected + TP	.46	.64	.45

In a last experiment, we apply a temporal propagation (**TP**). We split the data-set with respect to observation in Section 4.2 about the positive opinion distribution before and after the election. Beyond the fact that results are lower than in the last step with the standard propagation of the corrected set in addition to the corrected set(**Corrected + SP**), it is very interesting to notice that this propagation in term of homogeneity is nearest to the original labeled set. The main explanation comes from the positive ratio in the test set that is now more far from the propagated set than before (cf Table 7). And if we remain polarity distribution of the 3 last months in table 1) our homogeneous train set contains too much "*POSITIVE*" opinion compared to the test set.

6 Conclusion and perspectives

In this paper, we have presented statistics and the current work in progress in the annotation process of a new French political opinion data-set. We also presented various studies in learning techniques, data-mining and technics about building an annotated data-set. As the need for in-domain annotated data still persists we hope that the methods and tools presented here will help researchers in their quest of biggest and better data-set. Because lack of space the work done on topic annotation has not been presented here; it will be presented in future publications. Solving this problem could help prevent from annotator bias and errors and minimize human oversight, by implementing more sophisticated computer-based annotation workflows, coupled with in-built control mechanisms and low supervision. Such infrastructure need also to be implement reusable.

Table 7. Polarity distribution across the time for **FH** on the unlabeled set with standard and temporal propagation

	Temporal propagation			Standard propagation		
Date	Positive	Neutral	Negative	Positive	Neutral	Negative
March	.34	.25	.40	.15	.24	.61
April	33	.28	.39	.14	.26	.60
May	.25	.28	.48	.20	.23	.66
June	.9	.28	.63	7	.23	.70
July	.11	.29	.60	.5	.26	.69
August	.9	.25	.66	.6	.21	.74
September	.8	.26	.66	.5	.21	.74

References

1. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M. *Overview of RepLab 2013: Evaluating Online Reputation Management Systems* CLEF 2013 Labs and Workshop Notebook Papers (2013)
2. Amigó, E., Gonzalo, J., Verdejo, F. *A General Evaluation Measure for Document Organization Tasks* Proceedings of SIGIR 2013 (July 2013)
3. Artstein R and Poesio M 2008 *Inter-coder agreement for computational linguistics* Computational Linguistics 34(4), 555–596.
4. Baldridge J and Palmer A 2009 *How well does active learning actually work? time-based evaluation of cost-reduction strategies for language documentation* Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.
5. Berrocal J.-L. , Figuerola C. and Rodriguez A. *REINA at RepLab2013 Topic Detection Task* CLEF 2013.
6. Blum A and Mitchell T 1998 *Combining labeled and unlabeled data with co-training* Proceedings of the Workshop on Computational Learning Theory (COLT), Madison, WI.
7. Cohen J 1960 *A coefficient of agreement for nominal scale* Educational and Psychological Measurement 20(1), 37–46.
8. Cohn D, Atlas L and Ladner R 1994 *Improving generalization with active learning* Machine Learning 15, 201–221.
9. Cossu J.-V., Bigot B., Bonnefoy L., Morchid M., Bost X., Senay G., Dufour R., Bouvier V., Torres-Moreno J.-M., El-Bèze M. *LIA@RepLab 2013* An evaluation campaign for Online Reputation Management Systems (CLEF'13), 23-26 September 2013
10. Dagan I and Engelson S 1995 *Committee-based sampling for training probabilistic classifier* Proc. of the 12th International Conference on Machine Lea rning, pp 150–157.
11. Fabbrizio GD, Tur G and Hakkani-Tür D 2004 *Bootstrapping spoken dialogue systems with data reuse* Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue.
12. Hakkani-Tür D and Riccardi G 2003 *A general algorithm for word graph matrix decomposition* Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong.
13. Hakkani-Tür D and Riccardi G 2011 *Active Learning* Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, p195-224, John Wiley & Sons, Ltd

14. Hakkani-Tür D, Riccardi G and Tur G 2006 *An active approach to spoken language processing* ACM Transactions on Speech and Language Processing (TSLP) 3(3), 1â31.
15. Kamangar K, Hakkani-Tür D, Tur G and Levit M 2008 *An iterative unsupervised learning method for information distillation* Proceedings of ICASSP.
16. Liere R and Tadepalli P 1997 *Active learning with committees for text categorization* Proceedings of the Conference of the American Association for Artificial Intelligence (AAAI), Providence, RI.
17. Malouf R. and Mullen T. 2008. *Taking sides: User classification for informal online political discourse* In Internet Research, volume 18, pages 177–190.
18. McCallum AK and Nigam K 1998 *Employing EM and pool-based active learning for text classification* Proceedings of the International Conference on Machine Learning (ICML), Madison, WI.
19. Nigam K and Ghani R 2000 *Understanding the behaviour of co-training* Proceedings of the Workshop on Text Mining at the 6th ACM SIGKDD at the KDD.
20. OĆonnor B., Balasubramanyan R., Routledge B. and Smith N. 2010. *From tweets to polls: Linking text sentiment to public opinion time series* In International AAAI Conference on Weblogs and Social Media.
21. Sanchez-Sanchez C., Jimenez-Salazar H. and Luna-Ramirez W. *UAMCLyR at Replab2013: Monitoring Task in CLEF 2013.*
22. Settles B. *Active Learning Literature Survey, 2009*
23. Tur G, Hakkani-Tür D and Schapire RE 2005 *Combining active and semi-supervised learning for spoken language understanding* Speech Communication 45(2), 171–186.
24. Velcin J., Kim Y.-M, Brun C., Dormagen J.-Y, Sanjuan E., Khouas L., Peradotto A., Bonnevay S., Roux C., Boyadjian J., Molina A. and Neihouser M. *Investigating the Image of Entities in Social Media: Dataset Design and First Results* Proceedings of Language Resources and Evaluation Conference 2014
25. Villena-Román J. and García-Morera, J. *TASS 2013-Workshop on Sentiment Analysis at SEPLN 2013: An overview*
26. Walter T. and Back A. *A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests* In Proceedings of the 2013 46th Hawaii International Conference on System Sciences
27. Wang A., Hoang C. and Kan M.-Y *Perspectives on Crowdsourcing Annotations for Natural Language Processing*
28. Wang H., Can D., Kazemzadeh A., Bar F. and Narayanan S. *A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle* In Proceedings of the ACL 2012 System Demonstrations, ACL '12, pages 115–120. Association for Computational Linguistics
29. Wang, L., and Li, L. *Automatic Text Classification Based on Hidden Markov Model and Support Vector Machine* In Proceedings of The Eighth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2013 (pp. 217-224).