Machine Learned Annotation of tweets about politicians' reputation during Presidential Elections: the cases of Mexico and France

Jean-Valère Cossu[1], **Rocío Abascal-Mena**[2], **Alejandro Molina**[3], **Juan-Manuel Torres-Moreno**[1,4], **Eric SanJuan**[1]
[1]LIA UAPV, Avignon, France
[2]UAM-Cuajimalpa, DF, México
[3]Conabio, DF, México
[4]Ecole Polytechnique de Montréal, Montréal, Canada
firstname.lastname@univ-avignon.fr

Abstract

With regular elections challenges, opinion mining on Twitter recently attracted research interest in politics using Information Retrieval (IR) and Natural Language Processing (NLP). However, getting language and domain-specific annotated data still remains a costly manual step. In addition, the amount and quality of these annotations may be critical regarding the performance of NLP-based Machine Learning (ML) techniques. An alternative solution is to use cross-language and cross-domain sets to simulate training data. This paper describes ML approaches to automatically annotate Spanish tweets dealing with the online reputation of politicians. Our main finding is that a simple statistical NLP classifier without in-domain training can provide as reliable annotation as humans annotators can. It also outperforms more specific resources such as polarity lexicon or in-domain manually translated data.

Keywords: Natural Language Processing, Machine Learning, Opinion Mining, Twitter Mining, Political Analysis, Sentiment Analysis

1 Introduction

Modern media is changing our vision about society in many aspects. Human, Social, and Political Sciences ought to evolve to have all the necessary methodological tools in order to understand social or political trends as quick as required by modern society. Particularly Twitter[1], have been used, not only to make public opinions about different events or persons, but also as a way to participate in social movements as well.

For instance, the role of social networks during the presidential campaign of 2012 in Mexico gained great importance as the principal instrument for exercising public opinion, especially for young people. The youth organization yosoy132 was born during that campaign in and, thanks to social networks, youth from all universities, regardless their social conditions, shared a common trend topic. Moreover, studies conducted by the *Instituto Nacional de Estadística, Geografía e Informática* (INEGI)[2] claim that 40.3% of users of Information and Communication Technologies (ICT) in Mexico are young people who communicate via social networks and mobile phones and that they remain connected most of the day. This percentage equals 15.3 million people aged between 18 and 34 that are potential voters [Tello-Leal *et al.*, 2012]. Youth participation via Twitter in Mexico increased creating significant social and political communities around election subjects. Moreover, Youth Mexicans are not the only ones that massively send tweets. Twitter recently gets a great attention from the main candidates which also promote their discourse [Sandoval *et al.*, 2012] online.

As a consequence, Twitter provides the opportunity to collect, in real time, large amounts of data, directly from users; so tweets can be then analyzed in order to track reactions to events. Since Twitter provides the possibility to extract tweets and compose actual corpus there have been a lot of linguistic research applied in tweets. Using publicly available online data, to perform sentiment studies, significantly reduces the costs, efforts and time needed to administer large-scale public surveys and questionnaires [Bollen *et al.*, 2011]. But, although Sentiment Analysis (SA) is a useful area in the study of online communication, because it gives researchers the ability to automatically measure emotion in online texts [Thelwall *et al.*, 2011], political studies from a Machine Learning point-of-view in Spanish are still rare [Villena-Román *et al.*, 2013]. Nevertheless, this could be changed using ML methods to simulate human annotations and assists experts (in works such as [Sandoval-Almazán, 2015]) to label large collection of data.

Usual studies in the domain assume that a great effort of acquisition of the tweets and a subsequent manual labelling process is required. In addition, a validation process is needed to correct the errors introduced by manual labelling. Moreover important political events will always occur faster than our capacities of getting manually annotated data in several languages. In this context, we propose an approach that can provide a reliable pre-annotation using out-of-domain data which needs shallow supervision before validation in order to obtain a reliable corpus that can be used for more complex political studies like user political tendency detection or monitoring

[1] http://www.twitter.com

[2] National Institute of Statistics, Geography and Informatics of Mexico.

politician's reputation.

The rest of this paper is organized as follows: Section 2 gives an overview of related works and establishes further motivation for our work. In Section 3, we provide details about used data-sets. In section 4 we propose our approaches while Section 5 is devoted to a thoroughly evaluation. Finally, Section 6 gives some conclusions about our work and opens several perspectives.

2 State of the art

2.1 Tweets mining and sentiment analysis for politics

Politics have been addressed in previous works but mostly in English (see [Malouf and Mullen, 2008] and [Wang *et al.*, 2012a] for former studies). [O'Connor *et al.*, 2010], used a subjective lexicon that comes from the Opinion Finder in order to determine positive and negative scores for each data set corresponding to a tweet. In this case, the raw numbers of positive and negative tweets about a given topic are used to calculate a confidence score (the relation between the number of positive and negative tweets). The authors indicated that by a simple manual inspection of the tweets they have found examples that have been classified incorrectly. Nevertheless, the authors used this method to measure the "consumer confidence" (the presidential approval) in 2008 presidential elections in the United States. A different approach has been used to analyse political preferences by studying humour contained in tweets [Bollen *et al.*, 2011].

A psychometric instrument called Profile of Mood States (POMS) is used to distil six different emotional attributes: tension, depression, anger, vigour, fatigue and confusion. POMS provides a list of adjectives for which the patient has to indicate the level of approval. Each adjective is related to a state of mind and, therefore, the list can be exploited as the basis for a mood-analyser of textual data.

[Tumasjan *et al.*, 2010] presented a work in two parts: in the first one Linguistic Inquiry and Word Count (LIWC) is used to perform a superficial analysis of the tweets related to the different political parties that competed for the German Federal election in 2009. In the second part, the authors claim that the counting of tweets with references to one of the parties, accurately reflects the election results. On the other hand, they established that the Mean Absolute Error (MAE) of the "prediction" based on Twitter data was very close to the real surveys that were carried out.

An increasing number of empirical analyses of sentiment and mood based on Twitter collections have been used along with sophisticated algorithms of text pre-processing, using lexicon based classifiers, SVM and Naive Bayes. The main idea is to train a classifier using keywords from tweets to determine the mood see [Wijaya *et al.*, 2013; Martínez and González, 2013]. Moreover, several methods have been already proposed for exploiting tweets in order to detect people's mood changes throughout the day [Martínez and González, 2013; Lampos *et al.*, 2013].

In [Cha *et al.*, 2010], authors measured changes in the mood of the U.S. population, over three years, from tweets providing policy relevant indicators. More qualitative studies propose new insights about human behavior as a result showing that there is a tremendous ambition to develop opinion mining tools for social media [Maynard *et al.*, 2012; Chung and Mustafaraj, 2011; Dodds and Danforth, 2010; Gruzd *et al.*, 2011; Kramer, 2010]. Nevertheless, most of these works use English annotated corpora for experimentation, and to our knowledge, there is neither studies on Spanish nor on French about political sentiment analysis. But, how to deal with the lack of data-sets to train? In this work, we state that disposing of specific domain data in one language (French) and specific language data in another (Spanish) it is possible to transpose the learned expertise of a domain-French classifier into another in Spanish.

2.2 Multi-lingual and cross-language processing

The conversion of information expressed in different languages to a common representation is in general very complex. Cross-lingual Information Retrieval (CLIR) systems helps to retrieve documents in different languages by posing a query in source language. Then, the query is mapped into a common representation in order to retrieve pertinent documents in a destination language. The translation of documents, even phrases, into the query language requires enormous resources. Usually: (1) parallel texts, (2) machine translation systems and (3) bilingual machine readable dictionaries.

The study of multi-lingual and cross-language processing have been addressed in the Cross Language Evaluation Forum (CLEF) in the past years. [Jagarlamudi and Kumaran, 2008] describes their experiments and results of using CLEF 2007 data set in a Hindi to English cross lingual information retrieval system. By using a simple word to word translation of a query and word alignment table learned, they have obtained 73% of the performance of the monolingual system. Specially, the most important result in this work is the discovery of considering the 4 most probable word translations,with no threshold on the translation probability, gave better results than translating word to word.

[Capstick *et al.*, 2000] presents a system for supporting cross-lingual information retrieval, MULINEX, which retrieves documents from the Web by employing a dictionary based query translation. MULINEX supports French, German and English by using very large amounts of data for translation and different document categorization algorithms: n-gram categorizers for noisy input, k-nearest-neighbor algorithm for normal documents and pattern-based categorizers for very short documents. Besides the cross lingual functionality, MULINEX provides the automatic translation of documents and their summaries. MULINEX uses a query assistant that provides an opportunity for interactive query translation disambiguation. The translated query terms are translated back into the original query language. However, this approach has some clear limitations because of the lack of use of synonyms in the dictionary and because significant homonym in the target language can result in confusing back translations.

The work of [Thenmozhi and Aravindan, 2009] translates Tamil to English using statistical machine translation. They

deploy an IR system in Agriculture domain for the farmers of Tamil Nadu which helps them to specify their information need in Tamil and to retrieve the documents in English. The system is designed with dynamic learning, so any new word that is encountered in the translation process could be updated to the bilingual dictionary.

In [Wang *et al.*, 2012b], instead of using existing document representations, with additional information in a multi-view clustering setting, authors use an alternative approach of encoding the additional information as constraints. Results showed that with real data this approach is effective in improving the clustering by just using the original documents.

An interesting work that uses a cross-lingual mixture model (CLMM) for sentiment classification is presented in [Meng *et al.*, 2012]. It uses NLP tools such as alignment to reduce the bias towards that of the source language in transfer learning. The proposed model can learn previously unseen sentiment words from large unlabeled data, which are not covered by the limited vocabulary in machine translation of the labeled data. The CLMM can use unlabeled parallel data regardless of whether labeled data in the target language are used or not.

2.3 Automatic Annotations of Tweets and Agreement Issues

Recently, several researches within the Limosine project[3] [Carrillo de Albornoz *et al.*, 2014; Amigó *et al.*, 2013] lead to consider automatic annotation for corporate entities' e-Reputation analysis (mainly in English). Public figures e-Reputation also interested French researchers in the frame of the Imagiweb project[4] [Velcin *et al.*, 2014] and on the context of TASS[5] [Villena-Román *et al.*, 2013] respectively focusing on French and Spanish tweets and entities.

All agreed that human interpretation of these kind of more or less consensual contents is prone to mistakes and they all reported inter-annotator agreements quite similar to typical products-oriented Sentiment Analysis studies despite the task difficulty. Then ,it also remains difficult to obtain a strong ground-truth annotation since both facts and opinions have to be considered regardless of whether the content is opinionated or not. It is often hard to tell all the implications a message may have on the e-reputation of a given entity. And finally, the political context makes the task even harder. In this work we investigate how much ML techniques without correct training data can perform compared to humans annotators.

3 Data-Sets

3.1 Mexican political data-set

The corpus analyzed is the same used by [Jean-Valere Cossu *et al.*, 2014]. It consists in 800 tweets containing #AMLO hashtag that were extracted between the 9 and the 11 June in 2012. AMLO is the acronym for Andrés Manuel López Obrador, who was a left candidate to the Presidential elections in Mexico. AMLO has built a strong base of support among people who feel that they have been left behind as Mexico's economy grows and evolves. These tweets have manually annotated according to polarity for reputation from the author point-of-view[6]. Annotation disagreements have been solved with an extra annotator, the final annotation is considered as ground-through annotation.

The used data-set remains small because annotate a big mass of specialized tweets is a difficult and time-and-money-consuming process. In particular, the number of annotators for this task was very limited. However comparable studies [Sandoval *et al.*, 2012; Sandoval-Almazán, 2015] about Twitter and Mexican politicians are lead with the same amount of data.

Table 1: Class distribution in both complete and French subpart collection.

Class	**Class-Distribution**	**Class-Distribution** (French)
Negative	0.41	0.37
Neutral	0.29	0.30
Positive	0.30	0.33

Classes are well balanced with only a slightly difference with negative tweets between the both collections as shown in table 1.

3.2 ImagiWeb French political data-set

We use the ImagiWeb[7] collection used by [Jean-Valere Cossu *et al.*, 2014; Velcin *et al.*, 2014]. It consists in 3,184 manually annotated tweets[8] for both two main candidates (François Hollande and Nicolas Sarkozy respectively noted FH and NS later) at the last French presidential election in May 2012. Tweets were extracted between March and December 2012 and concern the two main candidates which is almost the same period as Mexican and RepLab sets.

Table 2: Class distribution in the French political collection.

Class	**Class-Distribution**
Negative	0.60
Neutral	0.12
Positive	0.28

Table 2 shows that the main tendency is negative with a very few number of neutral tweets. According to [Velcin *et al.*,

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

[4] http://mediamining.univ-lyon2.fr/velcin/imagiweb/

[5] *Taller de Análisis de Sentimientos en la SEPLN* / Workshop on Sentiment Analysis at SEPLN. See: http://www.daedalus.es/TASS2013/corpus.php

[6] Does the author have a Negative, Neutral or Positive opinion about AMLO.

[7] Data (including all annotators assessments) have recently been made publicly available at http://mediamining.univ-lyon2.fr/velcin/imagiweb/dataset.html

[8] Annotation was done by thirty people (with higher education) regarding polarity (more detailed statics about the annotation process are available in [Velcin *et al.*, 2014])

2014] the main reason is that politics in France unleash passions between people. For a reasonable analysis we only considered 3 polarity level from the 6 available in the data-set.

3.3 TASS political data-set

One part of the TASS 2013 evaluation [Villena-Román *et al.*, 2013] covers a sentiment analysis over political messages. The provided corpus is a selection of 2,500 tweets (2,150 are still available online), extracted from Twitter during the electoral campaign of the 2011 general elections in Spain (Elecciones a Cortes Generales de 2011). Tweets mentioning any of the four main national-level political parties: *Partido Popular (PP), Partido Socialista Obrero Español (PSOE), Izquierda Unida (IU) y Union, Progreso y Democracia (UPyD)* were selected. Tweets have been manually annotated according to global polarity and polarity at entity level (3 levels and no-sentiment tag). This entity level polarity is similar to Polarity for Reputation annotation from RepLab and the polarity definition in the Imagiweb data set. More details about the data-set and the annotation procedure can be found in [Villena-Román *et al.*, 2013].

Table 3: Class distribution in the TASS'2013 Political collection.

Class	Class-Distribution
Negative	0.27
Neutral	0.38
Positive	0.26
None	0.09

Table 3 shows that the main tendency is neutral with a slight difference between positive and negative values. We removed from our experiments tweets marked as having no polarity (the none tag).

3.4 RepLab data-set

We use the Spanish "polarity for reputation" side (23,100 tweets which represent around 20% of the collection) from the RepLab 2013 data-set [Amigó *et al.*, 2013]. In RepLab the goal is to decide if the tweet content has positive or negative implications for the company's reputation whether the content contains explicit sentiment words or only reports facts. Manual annotations are: positive, negative and neutral.

Table 4: Class distribution in the Spanish subset of the RepLab'2013 collection.

Class	Class-Distribution
Negative	0.24
Neutral	0.28
Positive	0.48

As shown in Table 4, the main tendency of the RepLab set is positive.

4 Experimental Approaches

We focus on improving domain and language portability to learn discriminative features that are not dependent on entities, domain or languages. This choice is motivated by many limitations identified in lexicon-based sentiment-analysis approaches and the specific annotation requirement. First, because they require the development of language-specific sentiment lexicons and annotation, which are expensive as they depend on human labour. Second, because of the short, noisy, and unedited text of social media updates limits of coverage of lexicons that result less effective than standard edited texts [Feczko *et al.*, 2008; Ohana and Tierney, 2009]. Third, and most importantly, because in politics opinion, reputation polarity is rarely encoded in sentiment-bearing words; they are also embedded in other words and short context, including. For instance, mentions of affair or financial organizations or scandals are highly correlated with negative opinion, as well as voting intention mentions to the opposite side.

In that follows, we describe our approaches the problem of detecting reputation polarity using several methods.

4.1 Lexicon Approach

Lexicon approaches start with a list of positive and negative words, which are already pre-coded. Our collection was first analyzed using a lexicon approach combined with a linguistic analysis in order to detect sentiments, during a period of time, in social and political tweets. We started with one Spanish lexicon and one English translated lexicon used to count for each tweet and for each corpus the number of positive and negative words contained in each tweet. All the process is automatically performed by using R[9].Words contained in a tweet are classified into positive or negative by those lexicon without taking into account the sarcasm that transforms the polarity of an apparently positive or negative utterance into its opposite [González-Ibáñez *et al.*, 2011]. Nevertheless, we assume that in a big corpus the sarcasm rest minimum.

4.2 Data pre-processing

We ignore all duplicate tweets (we chose to consider only the first according to the date). Each language is equally treated. Text is lower-cased and cleaned by removing hypertext links, stop-words and punctuation marks. The hash from hashtags was not removed.

4.3 Machine Learning

In [Carrillo de Albornoz *et al.*, 2014] and [Villena-Román *et al.*, 2013] machine learning was partly used to assist annotators and propose annotations. [Di Fabbrizio *et al.*, 2004] showed that a small annotated set coupled to machine learning could perform competitively to annotators to answer text mining tasks. The annotation was addressed as a classification problem that consisted of determining the polarity of each tweet. The choice of our classifiers is motivated by their good performance in many classification tasks in previous research on polarity detection and sentiment analysis [Joachims, 1998; Amigó *et al.*, 2013].

The features used by our proposals are words, bi-grams and tri-grams. They compose the tweet discriminant bag-of-words representation. We start with Term Frequency-Inverse

[9] R is an interpreted computer language designed for statistical data analysis (`http://www.r-project.org/`

Document Frequency (TF-IDF) [Robertson, 2004] combined with the Gini purity criteria [Torres-Moreno *et al.*, 2013]. This last work reports improvements using TF-IDF in association with the Gini purity criteria over n-grams (n<=3). We estimate the similarity of a given tweet by comparing it to each class and ranking it according to several distance or similarity index. Purity of a word i is defined with the Gini criteria as follows (1):

$$gini_i = \sum_{c \in \mathbb{C}} \mathbb{P}^2(i|c) = \sum_{c \in \mathbb{C}} \left(\frac{DF_i(c)}{DF_{\mathbb{T}}(i)} \right)^2 \quad (1)$$

where C is the set of classes, $DF_{\mathbb{T}}(i)$ is the number of tweets of the train set $\mathbb{T}$ containing the word i and $DF_c(i)$ is the number of tweets of the train set annotated with class c containing word i. This factor is used to weight the contribution $\omega_{i,d}$ of each term *i* in document *d* as (2):

$$\omega_{i,d} = TF_{i,d} \times \log(\frac{N}{DF_{\mathbb{C}}(i)}) \times gini_i \quad (2)$$

Where N is the number of tweets in the train set and the contribution $\omega_{i,c}$ of each term *i* in class *c* as (3):

$$\omega_{i,c} = DF_{i,c} \times \log(\frac{N}{DF_{\mathbb{C}}(i)}) \times gini_i \quad (3)$$

Cosine distance.
This distance is computed to compare similarities between the tweet bag-of-words and each class bag-of-words as follows (4) :

$$cos(d,c) = \frac{\sum\limits_{i \in d \cap c} \omega_{i,d} \times \omega_{i,c}}{\sqrt{\sum\limits_{i \in d} \omega_{i,d}^2 \times \sum\limits_{i \in c} \omega_{i,c}^2}} \quad (4)$$

SVM.
The SVM algorithms have shown their ability to handle large feature spaces and to determine the relevant ones [Joachims, 1998]. We chose to train linear multi-class Support Vectors Machine[10] [Crammer and Singer, 2002] with the objective of classifying multiple polarity classes in one pass. Classifiers have been trained with default parameters and the "bag-of-terms-weight" vectorial representation of each tweet d. Where each term weight is computed as(5):

$$\omega_i = DF_{\mathbb{T}}(i) \times log(\frac{N}{DF_{\mathbb{T}}(i)}) \times gini_i \quad (5)$$

Baseline.
The baseline algorithm was computed as simple memory test which consists in tagging each tweet d_1 with the most similar tweet d_2 in the training set (according to Jaccard index). This similarity is computed as follows:

$$sim(d_1, d_2) = \frac{\sum\limits_{i \in d_1 \cap d_2} \omega_{i,d}}{\sum\limits_{i \in d_1 \cup d_2} \omega_{i,d}} \quad (6)$$

[10]Multi-Class Support Vector Machine `http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html`

4.4 Overview Processing

Our experimental evaluation is outlined as described on figures 1, 2, 3 and 4.

4.5 Lexicon and documents translation

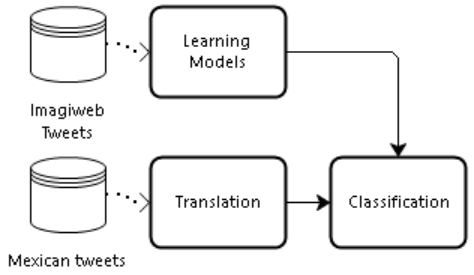

Figure 1: Classification process using translated documents

As a first experiment we choose to manually translate a sample (220 tweets) of our Mexican unlabelled set in order to perform a classification using the French annotated set as training set as shown on figure 1. The sub-part chosen for translation seems to present a better class balance (as shown in table 1). The main objective of this experiment is to verify the applicability of the same models to another one test set discussing other entities as done during RepLab 2012 [Amigó *et al.*, 2012]. We also separate both candidates from the Imagiweb set in a separate training set to evaluate if one candidate can be more similar to AMLO than the other.

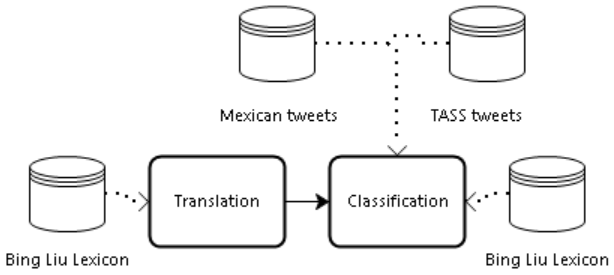

Figure 2: Classification process using lexicon approach

We chose to translate from English to Spanish a well know lexicon used in many SA tasks [Bing, 2012]. We have used the Bing Liu lexicon which is composed of around 6800 words in English. However, after making an automated translation using the Google Translator we have recovered only 2284 positive words and 1644 negative words. In this case, we left, manually, only the words that in Spanish can express a sentiment. We also compared the results of this approach with a classification using a Spanish lexicon specifically built for SA in Twitter and for politics analysis (ElhPolar Lexicon [Saralegi and Vicente, 2013]). We also evaluated both

lexicon over the TASS'2013 data-set to compare these approaches with regard to state-of-the-art (see figure 2 and Table 8).

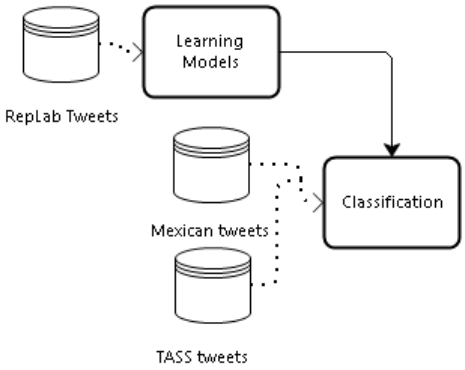

Figure 3: Classification process using RepLab'13 data-set as training set

We then investigated the classification using RepLab 2013 reputation set as training set. We questioned here the performance of a same language massive labelled set sharing the same short and noisy vocabulary specific to social media's texts. We performed the classification over Mexican tweets and TASS'2013 political set as shown in figure 2.

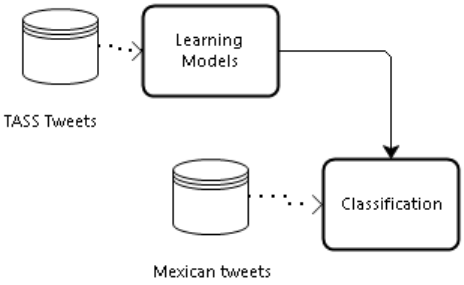

Figure 4: Classification process using TASS data-set as training set

In a last experiment (see figure 4) we used the TASS data-set as same language in-domain training set to automatically annotate the Mexican tweets.

5 Evaluation and results

5.1 Metrics

We report for each tested classifier on the overall Accuracy and as the class are not well balanced on each data set we propose to report on the Macro Averaged F-Score computed (noted F-Score in tables) as mean of each class F-Score (based on Precision and Recall) in order to give the same importance to each class. For instance, since the Negative class represents 60% of the Imagiweb set returning all tweets as Negative would give an Accuracy and global F-Score of 60% and a Macro Averaged F-Score of 33%, since it does only detect one class, which does not represent a significant and efficient system performance.

Accuracy is computed as follows:

$$Accuracy = \frac{\text{Num. of correct documents}}{\text{Num. of documents in the reference}} \quad (7)$$

Macro Averaged F-Score as:

$$\text{F\_Score} = \frac{\sum_c \frac{2 \times (\text{Precision}_c \times \text{Recall}_c)}{\text{Precision}_c + \text{Recall}_c}}{\text{Num. of classes}} \quad (8)$$

With Precision P_c for class c as:

$$P_c = \frac{\text{Num. of documents correctly assigned in class}_c}{\text{Num. of documents assigned in class}_c} \quad (9)$$

And Recall R_c for class c as:

$$R_c = \frac{\text{Num. of documents correctly assigned in class}_c}{\text{Num. of documents belonging to class}_c} \quad (10)$$

5.2 Machine Learning using translated data

The Imagiweb set provides sentiment annotation at the person level which may be more similar to our Mexican tweets and it provides an interesting experiment. According to Table 5 systems performance using same context data is really low. The main reasons are that the vocabulary used to described both French main candidates is not the same as the one used for AMLO but also that both class distributions appear to be too different. As systems performance do not increase while the size of the training set increases (when considering the complete Imagiweb set see 6), we can consider that systems performance are not limited by the amount of the training data available. Nevertheless, with the very limited size of this test set we are not able to conclude over the difference significance between systems performances.

Table 5: Classification performance on translated tweets.

Method	FH sub-set		NS sub-set	
	F-Score	Accuracy	F-Score	Accuracy
Baseline	0.29	0.33	**0.38**	0.39
Cosine	0.17	0.33	0.34	**0.42**
SVM	**0.34**	**0.36**	**0.38**	0.40

As the results are slightly better when we consider tweets from NS as training set we can consider that they Twitter' users talk about AMLO is similar to the way they talk about NS.

5.3 Translated and specialized lexicons over Spanish

Both lexicons approaches Elh-Polar lexicon and Bing Liu translated one, seem to not fit our data-set's vocabulary or this kind of analysis since they do not perform well as reported in Table 7.

Table 6: Classification performance on translated tweets.

Method	Complete Imagiweb Set	
	F-Score	Accuracy
Baseline	0.30	**0.33**
Cosine	0.26	**0.38**
SVM	**0.35**	0.37

Table 7: Lexicon classification performance on Mexican set.

Method	**F-Score**	**Accuracy**
ElhPolar Lexicon	0.25	0.32
Translated Lexicon	0.21	0.33

The lexicon approach also show its limit over the TASS'2013 data-set as reported in Table 8. We wanted to compare the best participating systems to *TASS'2013 Task 3 Sentiment Analysis at Entity Level* but as TASS organizers do not report over the Macro Averaged F-Score we are unable to evaluate the "Best" TASS performance in terms of F-Score and Accuracy. Nevertheless as TASS organizers report on a global F-Score based on Precision and Recall we suppose that results we obtained are mostly similar to the average of TASS'2013 participants (not reported in this paper).

Table 8: Lexicon classification performance on TASS'2013 set.

Method	**F-Score**	**Accuracy**
Elh-Polar Lexicon	0.30	0.41
Translated Lexicon	0.12	0.29

5.4 Machine Learning on Spanish out-of-domain data

In this experiment we performed the classification using RepLab 2013 set as training set. Although RepLab also provides annotation at the person level for some singers which may be similar the candidates level. Table 9 shows classification performance over Spanish contents according to F-Score and Accuracy. SVM is lower than baseline which performance is also higher than expected. An interesting performance is the Cosinus one's, while the Cosine similarity was outperformed using the translated tweets. It is able here to obtain quite good classification results that are close to inter-annotator agreements observed in the literature [Amigó *et al.*, 2013; Villena-Román *et al.*, 2013; Pla and Hurtado, 2014]. These results are also close to those reported in the literature which is probably our main interesting results: since with this level of performance classifiers can provide a reliable annotation faster and cheaper than what can expected from human experts.

The ML approaches have been proved to perform competitively during Replab [Amigó *et al.*, 2013] campaign. This is why classification performances will not be evaluated in this paper.

Annotation at the entity level from RepLab provides a granularity similar to the "Party/Entity" annotation from TASS.

Table 9: Classification performance on Mexican tweets.

Method	**F-Score**	**Accuracy**
Baseline	0.50	0.51
Cosine	**0.74**	**0.74**
ElhPolar Lexicon	0.25	0.32
Translated Lexicon	0.21	0.33
SVM	0.17	0.31

However when we consider TASS'2013 as test set we obtain a low level of performance as shown in Table 10. As TASS'2013 Political data-set is known to provide a difficult sentiment classification issue [Villena-Román *et al.*, 2013] although our results are disappointing it is not really a surprise. It would probably benefit from the creation of a more specific training material.

Table 10: Classification performance on TASS'2013 set.

Method	**F-Score**	**Accuracy**
Baseline	0.32	0.33
Cosine	0.32	0.33
SVM	0.33	0.33

5.5 Machine Learning using in-domain data

We performed in this experiment the classification using TASS'2013 set as training data. Table 11 shows classification performance using this set as training set.

Table 11: Classification performance on Mexican tweets.

Method	**F-Score**	**Accuracy**
Baseline	0.33	0.32
Cosine	**0.32**	**0.31**
SVM	0.31	0.29

The lower level of results could be explained by the smaller size of the training set (TASS'2013 set is really small compared to RepLab). As it is also nip and tuck between classifiers, although the limited test's size. Finally the main finding of this experiments is to show that words used this year in this context are very different from those used in both Mexican and RepLab sets.

5.6 Qualitative Analysis

Dealing with ambiguous contents often leads to note interesting errors. Some contents such as:

"RT 1.Naces 2.Eres AMLO 3. Creces 4. No eres presidente. 5. No eres presidente. 6. No eres presidente. 7. No eres presidente. 8. Mueres. JAJA" (In Englih: 1. You're born 2.You are AMLO 3. You grow 4. You're not president. 4. You're not president. 6. You're not president. 7. You're not president 8. You die. LOL LOL") are tagged positive by the systems while they are really negative.

Here it is another example:

"AMLO gran orador cada vez que abre la boca sueña #elpejeaburrehastaalospejezombies" ("AMLO great speaker everytime he opens the mouth he dreams" in English). It is also an

irony because people are not dreaming about a better country instead they are becoming tired and almost falling at sleep every time that AMLO speaks.

Automatic systems would also benefit from hashtag splitting since they are not able to understand aggregated words such as *"#esunhonortuitearporobrador or #alpejenolesalenlascuentas"*. Nevertheless, besides linguistic rules it will require a deeper processing including langugage knowledge to *"#elpejeaburrehastaalospejezombies #elpejeaburrehastaasuspejezombies and #elpejeaburrehastalospejezombies* as being the same statement.

These are typical examples of humoristic contents that systems are not able to handle properly. Lexicons will probably never be able to consider correctly this kind of messages. While ML approaches could handle them once they have seen similar examples in the training set or in an active learning procedure.

6 Conclusions

In this paper we described and compared several approaches for a fast political classification of Spanish tweets concerning last presidential election in Mexico. This kind of content is often hard to understand and annotation is prone to human error. Our experimental evaluation (although our test set was limited) establishes that without specific training material Mahcine Learning approaches can achieve state-of-the-art result while the literature insists on the need of specific training data.

Annotating this tweets dealing with politics is known to be a costly and difficult task. Our experiments have shown that the need of costly expert's annotation can be reconsidered. The presented ML approaches are mainly language- and domain-independents. So, only a little effort it is necessary to adapt these methods to another domain, such as the popularity of products or corporate entities, and handle large amount of data unlike experts. Another outcome from our experiments is probably the annotated data-set itself which can be used for further researches.

At first we intend to apply this process to others candidates at the Mexican election. This will allow us to investigate the correlation between polls results and the evolution of class distribution between candidates over the time. We also want to consider more specific work on the existing data-set. We have several ideas on how to improve our approach to identifying the polarity in political tweets using information carried in hashtags (although hashtag splitting will require extra-linguistic knowledge and linguistic rules) and Twitter users' name. The detection of irony and the study of re-tweet phenomena [Morchid *et al.*, 2014] can be two important elements to improve tweet classification. Then in forthcoming works, we think to study in detail the impact of these phenomena in the Micro-Blogs classification.

Acknowledgments

This work is funded in part by the project ImagiWeb ANR-2012-CORD-002-01 (France).

References

[Amigó *et al.*, 2012] Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Maarten De Rijke. Overview of replab 2012: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 333–352. Springer, 2012.

[Amigó *et al.*, 2013] Enrique Amigó, Jorge Carrillo De Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten De Rijke, and Damiano Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 333–352. Springer, 2013.

[Bing, 2012] Liu Bing. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and A P. Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[Capstick *et al.*, 2000] Joanna Capstick, Abdel Kader Diagne, Gregor Erbach, Hans Uszkoreit, Anne Leisenberg, and Manfred Leisenberg. A system for supporting cross-lingual information retrieval. *Information Processing and Management*, 36(2):275–289, 2000.

[Carrillo de Albornoz *et al.*, 2014] Jorge Carrillo de Albornoz, Enrique Amigó, Damiano Spina, and Julio Gonzalo. ORMA: A semi-automatic tool for online reputation monitoring in twitter. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014*, pages 742–745. Springer International Publishing, 2014.

[Cha *et al.*, 2010] Meeyoung Cha, Hamed Haddadi, FabrÃcio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[Chung and Mustafaraj, 2011] Jessica Elan Chung and Eni Mustafaraj. Can collective sentiment expressed on twitter predict political elections? In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[Crammer and Singer, 2002] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[Di Fabbrizio *et al.*, 2004] Guiseppe Di Fabbrizio, Gokhan Tur, and Dilek Hakkani-Tur. Bootstrapping spoken dialog systems with data reuse. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 72–80, Cambridge, Massachusetts, USA, 2004. Association for Computational Linguistics.

[Dodds and Danforth, 2010] PeterSheridan Dodds and ChristopherM. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[Feczko *et al.*, 2008] Matthew Feczko, Andrew Schaye, M Marcus, and A Nenkova. Sentisummary: Sentiment summarization for user product reviews. In *proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 265–271, 2008.

[González-Ibáñez *et al.*, 2011] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: A closer look. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 581–586, 2011.

[Gruzd *et al.*, 2011] Anatoliy A. Gruzd, Sophie Doiron, and Philip Mai. Is happiness contagious online? a case of twitter and the 2010 winter olympics. In *HICSS*, pages 1–9, 2011.

[Jagarlamudi and Kumaran, 2008] Jagadeesh Jagarlamudi and A. Kumaran. Cross-lingual information retrieval system for indian languages. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 80–87. Springer Berlin Heidelberg, 2008.

[Jean-Valere Cossu *et al.*, 2014] Rocıo Abascal Jean-Valere Cossu, Alejandro Molina Mena, Juan-Manuel Torres-Moreno, and Eric SanJuan. Bilingual and cross domain politics analysis. *Avances en la Ingeniería del Lenguaje y del Conocimiento*, page 9, 2014.

[Joachims, 1998] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[Kramer, 2010] Adam D. I. Kramer. An unobtrusive behavioral model of "gross national happiness". In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, pages 287–290, 2010.

[Lampos *et al.*, 2013] Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *ACL '13*, page 993–1003, Sofia, Bulgaria, 08/2013 2013. Association for Computational Linguistics, Association for Computational Linguistics.

[Malouf and Mullen, 2008] R. Malouf and T. Mullen. Taking sides: User classification for informal online political discourse. *In Internet Research*, 18:177–190, 2008.

[Martínez and González, 2013] Víctor Martínez and Víctor M. González. Sentiment characterization of an urban environment via twitter. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction - 7th International Conference, UCAmI 2013, Carrillo, Costa Rica, December 2-6, 2013, Proceedings*, pages 394–397, 2013.

[Maynard *et al.*, 2012] D. Maynard, K. Bontcheva, and D Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of NLP can u tag # user generated content*, 2012.

[Meng *et al.*, 2012] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 572–581, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[Morchid *et al.*, 2014] Mohamed Morchid, Richard Dufour, Pierre-Michel Bousquet, Georges LinarÃ¨s, and Juan-Manuel Torres-Moreno. Feature selection using principal component analysis for massive retweet detection. *Pattern Recognition Letters*, 49(0):33 – 39, 2014.

[O'Connor *et al.*, 2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[Ohana and Tierney, 2009] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.

[Pla and Hurtado, 2014] F. Pla and L. Hurtado. Political tendency identification in twitter using sentiment analysis techniques. In *COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 183–192, 2014.

[Robertson, 2004] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[Sandoval-Almazán, 2015] Rodrigo Sandoval-Almazán. Using twitter in political campaigns: The case of the PRI candidate in mexico. *IJEP*, 6(1):1–15, 2015.

[Sandoval *et al.*, 2012] Rodrigo Sandoval, Rodolfo Torres Matus, and Rosa Nava Rogel. Twitter in mexican politics: Messages to people or candidates? In *18th Americas Conference on Information Systems, AMCIS 2012, Seattle, Washington August 9-11, 2012*, 2012.

[Saralegi and Vicente, 2013] X. Saralegi and I. San Vicente. Elhuyar at tass 2013. In *XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 143–150, 2013.

[Tello-Leal *et al.*, 2012] E. Tello-Leal, D. A. Tello-Leal, and C. M. Sosa Reyna. Reflexiones sobre el uso de las tecnologías de información y comunicación en las campañas electorales en méxico: e-campañas. *Revista Virtual Universidad Católica del Norte*, (36):33–47, 2012.

[Thelwall *et al.*, 2011] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 2(62):406–418, 2011.

[Thenmozhi and Aravindan, 2009] D. Thenmozhi and C. Aravindan. Cross lingual information retrieval system for agriculture society. In *International Forum for Information Technology in Tamil Conference (INFITT)*, 2009.

[Torres-Moreno *et al.*, 2013] JM Torres-Moreno, M El-Beze, and P Bellot. Bechet, opinion detection as a topic classification problem in in textual information access. chapter 9, 2013.

[Tumasjan *et al.*, 2010] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, pages 178–185, 2010.

[Velcin *et al.*, 2014] Julien Velcin, Caroline Brun, Jean-Yves Dormagen, Young-Min Kim, Claude Roux, Julien Boyadjian, Stephane Bonnevay, Marie Neihouser, Eric SanJuan, Leila Khouas, Molina A., and Neihouser M. Investigating the image of entities in social media: Dataset design and first results. In *Language Resources and Evaluation Conference (LREC)*, 2014.

[Villena-Román *et al.*, 2013] Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González Cristóbal. TASS - workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44, 2013.

[Wang *et al.*, 2012a] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 115–120, 2012.

[Wang *et al.*, 2012b] X. Wang, B. Qian, and I. Davidson. Improving document clustering using automated machine translation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 645–653, 2012.

[Wijaya *et al.*, 2013] V. Wijaya, A. Erwin, M. Galinium, , and W Muliady. Automatic mood classification of indonesian tweets using linguistic approach. In *International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 41–46, 2013.