Lyapunov Filtering of Objectivity for Spanish Sentiment Model

Iti Chaturvedi and Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University, Singapore
{iti,cambria}@ntu.edu.sg

David Vilares
Grupo LyS, Departamento de Computación,
Universidade da Coruña, Spain
david.vilares@udc.es

***Abstract*—Objective sentences lack sentiments and, hence, can reduce the accuracy of a sentiment classifier. Traditional methods prior to 2001 used hand-crafted templates to identify subjectivity and did not generalize well for resource-deficient languages such as Spanish. Later works published between 2002 and 2009 proposed the use of deep neural networks to automatically learn a dictionary of features (in the form of convolution kernels) that is portable to new languages. Recently, recurrent neural networks are being used to model alternating subjective and objective sentences within a single review. Such networks are difficult to train for a large vocabulary of words due to the problem of vanishing gradients. Hence, in this paper we consider use of a Lyapunov linear matrix inequality to classify Spanish text as subjective or objective by combining Spanish features and features obtained from the corresponding translated English text. The aligned features for each sentence are next evolved using multiple kernel learning. The proposed Lyapunov deep neural network outperforms baselines by over 10% and the features learned in the hidden layers improve our understanding subjective sentences in Spanish.**

I. Introduction

Subjectivity detection can prevent the sentiment classifier from considering irrelevant or potentially misleading text [1]. This is particularly useful in multi-perspective question answering summarization systems that need to summarize different opinions and perspectives and present multiple answers to the user based on opinions derived from different sources. It is also useful to analysts in government, commercial and political domains who need to determine the response of the people to different crisis events [2], [3].

Subjectivity detection approaches in the 1990's used well established general subjectivity clues to generate training data from un-annotated text [1]. In addition, features such as pronouns, modals, adjectives, cardinal number, and adverbs showed to be effective in subjectivity classification. Some existing resources contain lists of subjective words, and some empirical methods in natural language processing (NLP) have automatically identified adjectives, verbs, and n-grams that are statistically associated with subjective language [4]. However, several subjective words such as may occur infrequently, consequently a large training dataset is necessary to build a broad and comprehensive subjectivity detection system.

Existing approaches to subjectivity detection can be grouped into three main categories: keyword spotting, lexical affinity, and statistical methods [5].

Keyword spotting is the most naïve approach and probably also the most popular because of its accessibility and economy. Text is classified into categories based on the presence of fairly unambiguous words. One scheme uses the concept of $private - state$ that is a general term for opinions and emotions that are positive or negative [1]. The phrase "Injustice cannot last long" contains a negative private state. Human annotators are used to judge the strength of each private state as low, medium, high or extreme. A sentence is subjective if it contains a private state and all other sentences are objective.

Similarly, [6] created a rule-based subjectivity dataset using a list of subjectivity clues and patterns. Next, a Naïve Bayes classifier was trained on extraction patterns as well as pronouns, adjectives, cardinal numbers and adverbs features in subjective and objective sentences. The drawback of their approach was that they assume that low subjectivity score sentences may be objective. However, it is difficult to identify objective sentences since any objective sentence can be made subjective using a subjective modifier. Some authors have shown that in some patterns may be highly correlated with objectivity in a particular domain. For example in Wall Street journals, sentences containing 'price' or 'profit' are likely to be objective.

Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious words, it assigns to arbitrary words a probabilistic 'affinity' for a particular category. In [1], the authors ranked patterns using the conditional probability given by the frequency of a pattern in a subjective sentence given the total frequency of each pattern in all training sentences. For example, all sentences that contain the verb 'asked' in the passive voice are subjective. Similarly, expressions involving the noun 'fact' are highly correlated with subjective expressions. The drawback with this approach is that it is unable to identify objective patterns effectively resulting in false positives.

Subjectivity detection in foreign languages was proposed by translating English lexicons in [7]. However, translation requires the lemmatized form of words which can lead to loss of subjective form. For example, the lemma form of 'memories' is 'memory', when translated to Romanian becomes 'memorie' with an objective meaning 'the power of retaining information'. Spanish sentences were first translated to English and then used to train a subjectivity classifier in [8].

However, translation of sentences can lead to loss of lexical information such as word sense resulting in a low accuracy. Similarly, in [9] the authors tried to minimize the resources to build a subjectivity lexicon in foreign languages. They used bootstrapping to sample subjective clues from a few manually selected seed words. In each iteration, candidate words with low similarity with the original seed list are discarded. The method is limited by the fact that suitable seed words may be difficult to determine in some domains or languages. Further, with each bootstrap iteration the noise in the subjectivity dataset keeps increasing.

Recently, the idea to learn a shared deep representation from multiple languages in a common space has been proposed [10]. Here, the objective function minimizes the distance between two parallel sentences in both languages. However, this can lead to loss of information. The main difference of our method from [11] is that instead of using a tree structure to determine causality while training an auto-encoder with aligned sentences from English and German, here we consider a hierarchy of feature detectors where lower level features are learned using convolution and the higher level features are learned using recurrent neural networks (RNNs) guided by a Lyapunov stability constraint. We refer to the resulting framework as Lyapunov deep neural network (LDNN). In the next section, we describe related work and outline of the paper.

II. Related Work and Contributions

Endogenous NLP methods automatically learn concepts from documents by training state space graphs where nodes are the words and the arc determine causal dependencies among them in large documents [12], [13]. In this way, no prior semantic understanding of documents or linguistic databases are needed [14]. For example, conditional random fields (CRFs) are commonly used for sequence labeling tasks such as part-of-speech (POS) tagging, named-entity recognition, and shallow parsing [15]. Here, shallow parsing is used to summarize relevant information from documents by labelling each word as +1 or -1 denoting that it is included or excluded from the summary. A Bayesian network is able to represent subjective degrees of confidence. The representation explicitly explores the role of prior knowledge and combines pieces of evidence of the likelihood of events. In order to compute the joint distribution of the belief network, there is a need to know $p(P|parents(P))$ for each variable P. It is difficult to determine the probability of each variable P and also difficult a statistical table for large-scale inference. Semantic networks, on the other hand, represent knowledge in patterns of interconnected nodes and arcs. Definitional networks focus on IsA relationships between a concept and a newly defined sub-type. The result of such a structure is called a generalization, which in turn supports the rule of inheritance for copying properties defined for a super-type to all of its sub-type. WordNet is an example of a well known semantic network [16]. The methods described above focus on English language, hence to allow for portability to foreign languages such as Spanish or Arabic, deep convolutional neural networks (CNNs) that can learn a dictionary of common features are suitable. For instance, we can assume that synonyms convey the same orientation and antonym relations convey an inverse sentiment in the foreign language after translation. Next, feature relatedness graphs are built for the foreign language using mappings from foreign senses to the English senses available in WordNet. In a deep neural network (DNN), the lower layers learn abstract concepts and the higher layers learn complex features for subjective sentences.

CNNs are sensitive to the order of words in a sentence and do not depend on external language specific features such as dependency or constituency parse trees [17]. Here narrow or wide convolution is achieved by applying filters such as pattern templates across the input sequence of words. A convolution layer in the network is obtained by convolving a matrix of weights with the matrix of activations at the layer below and the weights are trained using back propagation [18]. Next to model sentiment flow, in [19], the authors used recurrent CNN to model the dynamics in dialogue tracking and question answering systems. However, they assume that the data is unimodal.

The significance and contributions of the research work presented in this paper can be summarized as follows:

- We propose a framework for subjectivity detection in Spanish by automatically extracting convolution features in Spanish and the translated English form of each sentence. The aligned features of both languages for each sentence are then combined using RNNs and multiple kernel learning (MKL).
- A linear matrix inequality has been developed to derive the stability criteria for multi-lingual subjectivity detection. Our results show that the proposed framework outperforms baselines on two benchmark datasets.
- We show how lexical resources in English such as subjectivity clues, POS tags and word sense disambiguation (WSD) can be effectively transferred from English to Spanish.

To verify the effectiveness of LDNN in capturing dependencies in high-dimensional data and its portability on language translation task we consider MPQA Gold corpus of 504 sentences manually annotated for subjectivity in Spanish [7], [6]. Here, we try to develop a subjectivity lexicon for Spanish language using the available resources in English.

Next, to evaluate the method on a very large dataset we use the TASS corpora that is a collection of Spanish tweets commonly used for the evaluation of social media analysis tasks [20]. It is the evaluation framework used in different editions of the TASS workshop on Sentiment Analysis for Spanish. It includes collections for sentiment analysis, topic modeling, political analysis or aspect-based sentiment analysis, among other challenges. The classification accuracy obtained using the proposed LDNN is shown to outperform the baseline by over 10% on both real datasets.

The rest of the paper is organized as follows: Section III provides the preliminary concepts necessary to comprehend the proposed LDNN algorithm of the present work. In section

IV, we introduce the Lyapunov linear matrix inequality for learning the weights of a RNN from both Spanish and English features. Lastly, in section V, we validate our method on real world benchmark datasets.

III. Preliminaries

We briefly review the theoretical concepts necessary to comprehend the present work. This is followed by the linear matrix inequalities that ensure stable convergence of the multilingual model. We begin with a description of conditional restricted Boltzmann machines (CRBM). Layers of CRBM result in a deep model for sentence classification. Next, we explain RNNs and the relation of convolution features to temporal features. We also describe MKL to combine features from different languages.

A. Deep Neural Networks

A DNN can be viewed as a composite of simple, unsupervised models such as restricted Boltzmann machines (RBMs) where each hidden layer serves as the visible layer for the next RBM. RBM is a bipartite graph comprising two layers of neurons: a visible and a hidden layer; where the connections among neurons in the same layer are not allowed.

To train such a multi-layer system, we must compute the gradient of the total energy function E with respect to the weights in all the layers. To learn such weights and maximize the global energy function, the approximate maximum likelihood contrastive divergence approach can be used. This method employs each training sample to initialize the visible layer. Next, it uses the Gibbs sampling algorithm to update the hidden layer and then reconstruct the visible layer consecutively, until convergence. As an example, here we use a logistic regression model to learn the binary hidden neurons and each visible unit is assumed to be a sample from a normal distribution. The continuous state $\hat{h}_j$ of the hidden neuron j, with bias b_j, is a weighted sum over all continuous visible nodes $\boldsymbol{v}$ and is given by:

$$\hat{h}_j = b_j + \sum_i v_i w_{ij}, \tag{1}$$

where w_{ij} is the connection weight to hidden neuron j from visible node v_i. The binary state h_j of the hidden neuron can be defined by a sigmoid activation function:

$$h_j = \frac{1}{1 + e^{-\hat{h}_j}}, \tag{2}$$

Similarly, in the next iteration, the continuous state of each visible node v_i is reconstructed. Here, we determine the state of visible node i, with bias c_i, as a random sample from the normal distribution where the mean is a weighted sum over all binary hidden neurons and is given by:

$$v_i = c_i + \sum_j h_i w_{ij}, \tag{3}$$

where w_{ij} is the connection weight to hidden neuron j from visible node i. This continuous state is a random sample from $\mathcal{N}(v_i, \sigma)$, where σ is the variance of all visible nodes.

Unlike hidden neurons, visible nodes can take continuous values in a Gaussian RBM. Lastly, the weights are updated as the difference between the original and reconstructed visible layer labelled as the vector $\boldsymbol{v}_{recon}$ using:

$$\triangle w_{ij} = \alpha(< v_i h_j >_{data} - < v_i h_j >_{recon}), \tag{4}$$

where α is the learning rate and $< v_i h_j >$ is the expected frequency with which visible unit i and hidden unit j are active together when the visible vectors are sampled from the training set and the hidden units are determined by (1). Finally, the energy of a DNN can be determined in the final layer using:

$$\mathrm{E} = -\sum_{i,j} v_i h_j w_{ij}, \tag{5}$$

To extend the DNN to a convolutional deep neural network, we simply partition the hidden layer into Z groups. Each of the Z groups is associated with a $n_x \times n_y$ filter where n_x is the height of the kernel and n_y is the width of the kernel. Let us assume that the input image has dimension $\mathrm{L_x} \times \mathrm{L_y}$. Then the convolution will result in a hidden layer of Z groups each of dimension $(\mathrm{L_x} - n_x + 1) \times (\mathrm{L_y} - n_y + 1)$. These learned kernel weights are shared among all hidden units in a particular group. The energy function of layer l is now a sum over the energy of individual blocks given by:

$$\mathrm{E}^l = -\sum_{z=1}^{\mathrm{Z}} \sum_{i,j}^{(\mathrm{L_x}-n_x+1),(\mathrm{L_y}-n_y+1)} \sum_{r,s}^{n_x,n_y} v_{i+r-1,j+s-1} h_{ij}^z w_{rs}^l. \tag{6}$$

Hence, each layer of a deep convolution neural network is referred to as a convolution RBM (CRBM). In such a model the lower layers learn abstract concepts and the higher layers learn complex features for subjective sentences.

B. Recurrent Neural Networks

The delay equation for a RNN with distributed time delays and several system modes that follows a Markov process $r(t)$ is given as follows:

$$\begin{aligned} \mathbf{x(t+1)} &= -A\mathbf{x(t)} + \mathbf{W}_0(r(t))f(\mathbf{x}(t)) \\ &+\mathbf{W}_1(r(t))g_1(\mathbf{x}(t-\tau)) + \mathbf{W}_2(r(t))\int_{t-\tau}^{t} g_2(\mathbf{x}(t)dt \\ y(t) &= \mathbf{C}(r(t))\mathrm{x}(t) + f(t,\mathrm{x}(t)) \\ \text{where } &|g_k(x) - g_k(y)| \leq |G_k(x-y)| \forall x, y \\ \text{and } &|f(x) - f(y)| \leq |F(x-y)| \forall x, y \end{aligned} \tag{7}$$

where A is a diagonal matrix of degradation rates of the neurons, $\mathbf{W}_0(r(t))$, $\mathbf{W}_1(r(t))$ and $\mathbf{W}_2(r(t))$ are the connection weight matrix, the discretely delayed connection weight matrix, and the distributively delayed connection weight matrix, C is the output weight matrix, the system is in mode $r(t)$ at time instant t, g_k, f are the activation functions that satisfy the Lipschitz condition with known constant scale matrix G_k, F, $x(t)$ is the state of the neurons and $\mathbf{x}(t-\tau)$ is the input shifted in time by τ delays.

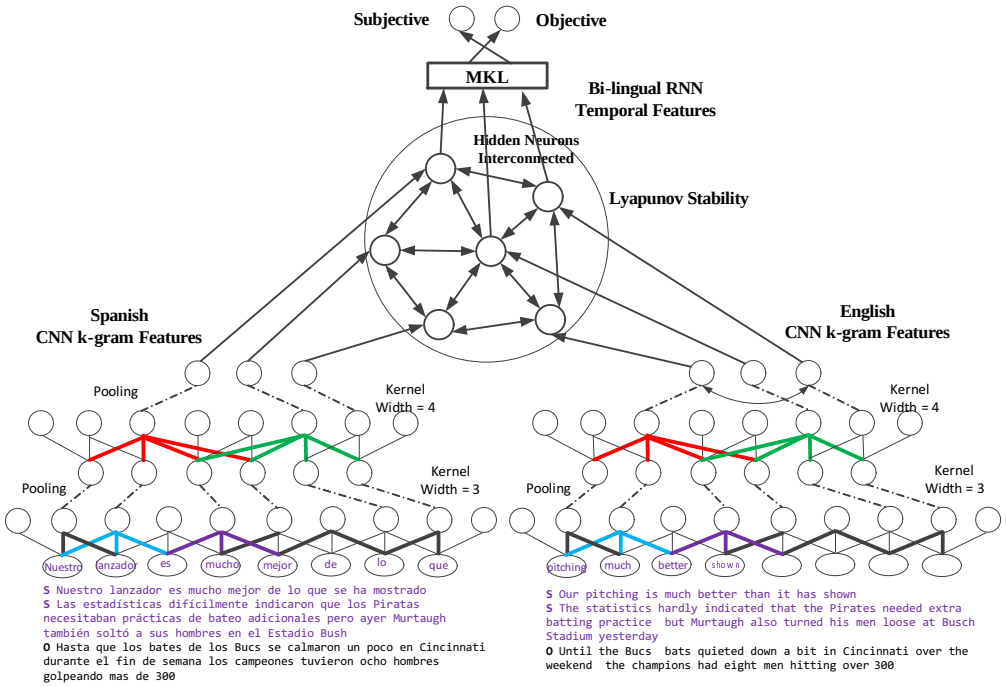

Fig. 1. Illustrates the state space of a LDNN for a subjective sentence in a movie review. Features are extracted in Spanish as well as English using two deep CNNs. The extracted features are combined using a RNN with Lyapunov stability constraint. Lastly, MKL classifier is trained using the features learned by the RNN. The bold lines correspond to kernels. The second layer has kernels of width 3, and the fourth layer has kernels of width 4. The third and fifth layer are pooling layers shown as dashed lines.

In this paper, we propose to learn distributed time-delayed dependence $\mathbf{W}_2(r(t))$ using CNNs. Hence, a kernel of width k is able to capture distributed delays of up to k and is given by the covariance matrix of features learned in the penultimate layer using (4). To learn the weights $\mathbf{W}_0(r(t))$ and $\mathbf{W}_1(r(t))$ of the RNN, back propagation through time is used where the hidden layer is unfolded in time using duplicate hidden neurons.

C. Multiple Kernel Learning

MKL uses the sequence of sentences $s(1), s(2), \ldots, s(T)$ and the corresponding target labels $y(t) \in \{Subj, Obj\}$ to train a classifier of the dual form :

$$\max_{\beta} \min_{\alpha} \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} \alpha_i \alpha_j y(i) y(j) \left(\sum_{m=1}^{M} \beta_m K_m(s(i), s(j)) \right) - \sum_{i=1}^{T} \alpha_i,$$

$$\text{s.t} \sum_{i=1}^{T} \alpha_i y(i) = 0, \ \sum_{m=1}^{\mathrm{M}} \beta_m = 1, \ 0 \leq \alpha_i \leq C \forall i. \tag{8}$$

where M is the total number of positive definite Gaussian kernels $K_m(s(i), s(j))$ each with a set of different parameters and α_i, b and $\beta_m \geq 0$ are co-efficients to be learned simultaneously from the training data using quadratic programming.

IV. Lyapunov Deep Neural Network Framework

In this section we propose a novel linear matrix inequality (LMI) that ensures stable convergence of multi-lingual DNN to the optimal solution. The proposed framework is hence referred to LDNN. Next, we describe the complete flow chart for merging lexical resources in Spanish and English.

A. Lyapunov Stability Condition

In this section, we provide the LMI for stable convergence of RNN using $V(t)$ as the Lyapunov function for the error function $e(t)$ of (7):

$$V(t) = x(t)^T P_i e(t) + \int_{t-\tau}^{t} e(s)^T Q_1 e(s) ds + \int_{\tau}^{0} \int_{t+s}^{t} e(\eta)^T Q_2 e(\eta) d\eta ds \tag{9}$$

where i is the state of the Markov chain $r(t)$, $P_i > 0$, $Q_1 \geqslant 0$, $Q_2 \geqslant 0$, $Q_1 = \epsilon_{1i} G_1^T G_1$, $Q_2 = \epsilon_{2i} G_2^T G_2$, for positive scalars $\epsilon_{1i}, \epsilon_{2i} > 0$ and G_1, G_2 are known scale matrices of the activation functions. To this end, we first provide a lemma that combines the different time-delay weight matrices, similar to [21].

Lemma 1 : If there exist matrix P_i, R_i and positive scalars $\epsilon_{0i}, \epsilon_{1i} > 0$ such that the linear matrix inequality in Table I holds, where γ_{ij} are the transition probabilities of Markov process $r(t)$, then the RNN is globally asymptotically stable.

Proof: As given in [21]

where $K_i = P_i^{-1} R_i$, is the gain matrix to be designed.

TABLE I
LINEAR MATRIX INEQUALITY FOR LYAPUNOV STABILITY AND GAIN CORRECTION

$$\begin{bmatrix} -A_iP_i - P_iA_i - R_iC_i + C_i^T R_i^T + \sum \gamma_{ij} P_j & P_iW_{0i} & \epsilon_{0i}G_0^T & P_iW_{1i} & \epsilon_{1i}\tau G_1^T & P_iW_{2i} & \epsilon_{2i}\tau G_2^T & R_i & \epsilon_{3i}F^T \\ W_{0i}^T P_i & -\epsilon_{0i}I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \epsilon_{0i}G_0 & 0 & -\epsilon_{0i}I & 0 & 0 & 0 & 0 & 0 & 0 \\ W_{1i}^T P_i & 0 & 0 & -\epsilon_{1i}I & 0 & 0 & 0 & 0 & 0 \\ \epsilon_{1i}G_1 & 0 & 0 & 0 & -\epsilon_{1i}I & 0 & 0 & 0 & 0 \\ W_{2i}^T P_i & 0 & 0 & 0 & 0 & -\epsilon_{2i}I & 0 & 0 & 0 \\ \epsilon_{2i}G_2 & 0 & 0 & 0 & 0 & 0 & -\epsilon_{2i}I & 0 & 0 \\ R_i^T & 0 & 0 & 0 & -0 & 0 & 0 & -\epsilon_{3i}I & 0 \\ \epsilon_{3i}F & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\epsilon_{3i}I \end{bmatrix} < 0$$

For the case of CNNs, we initialize the distributed time-delay weight matrix W_{2i} with the covariance matrix of deep CNN output, the predicted class in each layer can hence be corrected by multiplication with gain matrix K_3 for distributed time delays τ in that layer.

Figure 1 the state space of a LDNN for a subjective sentence in a movie review. Features are extracted in Spanish as well as English using two deep CNNs. The extracted features are combined using a RNN with Lyapunov stability constraint. Lastly, MKL classifier is trained using the features learned by the RNN. The bold lines correspond to kernels. The second layer has kernels of width 3, and the fourth layer has kernels of width 4. The third and fifth layer are pooling layers shown as dashed lines.

B. Spanish Sentiment Model

In this section, we describe the entire framework of combining Spanish and English resources. Subjectivity clue words such as 'good', 'happy', 'sad' are available for English language. These were translated to obtain the corresponding list in Spanish using the Bing translator API. Since words may have different subjectivity when used in different forms such as 'noun' or 'verb', hence POS tagging was done for all training sentences in Spanish and English. Gaussian Bayesian networks were constructed over subjectivity clues with highest frequency as described in [22]. The phrases corresponding to high ML structures were used to select important sentences that are used to pre-train the deep neural network. The deep CNN where each layer is a CRBM was used to extract features in the form of 3-grams and 4-grams in each language separately. Next, we align the English and Spanish features for a single sentence to form a single sample of features in the training set. The new training data is used to train a RNN. In order to ensure stable convergence the RNN output is multiplied with a suitable gain matrix as explained in Section IV-A. Lastly, the low-dimensional features learned at the hidden neurons of the RNNs are further evolved using a multiple-kernel learning classifier (MKL). Since the word sense changes when translating from Spanish to English, we can get rid of some false positives by verifying the subjectivity of each sentence using Spanish WSD database. Figure 2 illustrates the Spanish Sentiment Model that combines lexical resources in Spanish with English.

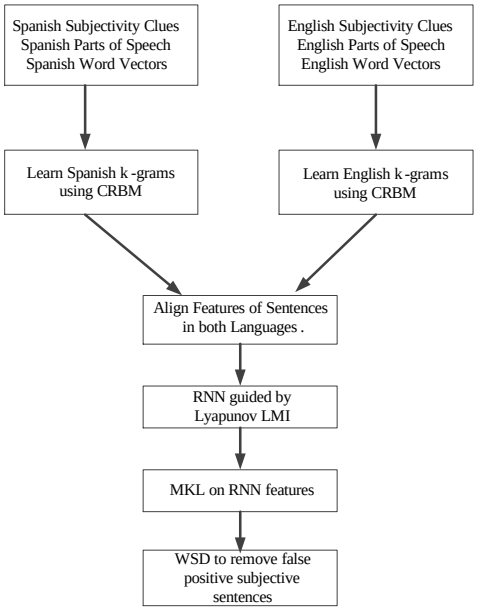

Fig. 2. Illustrates the Spanish Sentiment Model that combines lexical resources in Spanish with English.

V. EXPERIMENTS

We applied our proposed algorithm to two real world problems. The two datasets were real world data collected from Spanish news articles and Spanish tweets. The first was a small dataset to classify the sequence of sentences in a news article as subjective or objective. The second was a very large dataset of Spanish tweets that can belong to any of four categories namely positive, negative, neutral or none. Performance measures such as F-measure[1] and mean square classification error were evaluated using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), respectively in the network reconstructed at the visible layer.

A. Preprocessing

The data pre-processing included removing top 50 stop words and punctuation marks from the sentences[2]. Next, we used a POS tagger to determine the POS for each word in a sentence.

[1]F-measure =2 $\frac{Precision \times Recall}{Precision + Recall}$

[2]http://www.ranks.nl/stopwords/

Subjectivity clues dataset [1] contains a list of over 8,000 clues identified manually as well as automatically using both annotated and un-annotated data. Each clue is a word and the corresponding POS. The frequency of each clue was computed in both subjective and objective sentences of the MPQA corpus. Here, we consider the top 50 clue words with highest frequency of occurrence in the subjective sentences. We also performed chunking of the sentences into top 25 concepts that are correlated in a vector space of emotions as described in [23], [24], [25].

In order to determine the optimal structure among the top words and concepts in subjective and objective sentences, each of the sentences in the training set was transformed to a binary feature vector where presence of a top word is denoted as '1' and absence is denoted as '0'. Since each sentence is dependent on the previous sentence in the article; the resulting matrix of words versus frequency is a time series. It must be noted that each word in a sentence is also dependent on the preceding words.

We use multivariate Gaussian Bayesian fitness function to extract the maximum likelihood (ML) probabilities of each word given up-to three parent words and up-to two time points delay. Such sub-structures are referred to as network motifs. Top 20% of Motifs with high ML are used to select the training sentences for the CNN.

1) Translating subjectivity clues: MPQA subjectivity clues are just available for English. To get the correspondence for Spanish, we followed a cross-lingual approach using an statistical machine translation system [26]. Particularly, we relied on the free API provided by https://www.bing.com/translator/.

2) Taggers: To exploit the subjectivity clues, we trained POS taggers, both for English and Spanish. To do that, we consider the universal POS tag set introduced in [27] and that is included as a part of the Universal Dependency Treebanks v2.0 [28]. The latter already provides official splits for training, development and test sets for training taggers and dependency parsers. There exist specific resources for POS tagging of tweets [29]. However, to the best of our knowledge, there only are available resources for training English models. In this context, we preferred to used the universal tag sets, to make more homogeneous and comparable the results between the two languages. To train the taggers we relied on the free distribution of the maximum-entropy tagger proposed in [30], that showed improvement on the labeling of unknown words, which used to drop accuracy of classical taggers [31], and can be considered as an advantage when dealing with tweets, where a larger percentage of unknown words tend to appear.

3) Word Sense Disambiguation for Spanish: We rely on Babelfy to apply word-sense disambiguation on the Spanish texts. Babelfy[3] is an state-of-the-art multilingual framework for WSD and entity linking [32]. In terms of WSD, it follows a knowledge-based approach by exploiting relations between word meanings from BabelNet [33].

[3]http://babelfy.org

Given an input text, the substrings matching a BabelNet entity are analyzed and their candidate meanings are ranked, by previously computing a graph-based semantic interpretation. As a result, we obtain the most coherent meaning of that expression within the input text, including additional information such as its POS tag. In this way, we can determine if a word is a real subjectivity clue or if it has a different meaning. For example, the word 'fine' can be a noun with negative sentiment in 'They put me a fine, because I was driving too fast' or a positive adjective in 'I feel really fine today'. Selecting the right subjectivity clue is crucial to correctly analyze both sentences.

B. Subjectivity Classification on the MPQA Gold Corpus

In order to evaluate the portability of the proposed method on language translation task we consider another MPQA Gold corpus of 504 news sentences manually annotated for subjectivity in Spanish [6]. The annotation resulted in 273 subjective and 231 objective sentences as described in [7]. The sentences are machine translated into English to obtain the training dataset. This corpus is small, as the annotators need to be trained with annotation guidelines in Spanish. Some sentences are difficult to annotate as Objective or Subjective and, hence, are annotated by several different annotators. However, it is a popular benchmark used by previous authors, and can evaluate the robustness of LDNN when few training samples are present.

The CNN is collectively pre-trained with both subjective and objective sentences that contain high ML word and concept motifs. The word vectors are initialized using a context window of size 5 and 30 features. Each sentence is wrapped to a window of 50 words to reduce the number of parameters and, hence, the over-fitting of the model. A deep CNN with three hidden layers of 100 neurons and kernels of size $\{3, 4, 5\}$ and one logistic layer of 300 neurons is used. The output layer corresponds to two neurons for each class of sentiments. The 300 feature outputs of deep CNN from both languages are used to train a recurrent NN with 10 hidden neurons and up-to 2 time point delays. These 10 features are then used to train the MKL classifier. Lastly, we check the word sense of each sentence predicted as subjective using the WSD lexicon by Babelfy. We used 10-fold cross validation to determine the accuracy of the trained CNN classifier on new sentences.

Table II shows that LDNN outperforms previous methods by 5-15% in accuracy. A comparison was done with baseline classifiers such as rule-based classifier [7], bootstrapping based classifier [9], SVM and Naïve Bayes [8]. The Bootstrapping method starts with a set of seed words in Spanish and iteratively includes new words into the lexicon with maximum similarity in each Bootstrap or iteration. Such a method is unable to capture the temporal dependence between sentences. By using a layer of recurrent neurons, we are able to learn time-delayed features for polarity changes within a single review. Lastly, WSD and rule based classifiers are heavily dependent on templates and do not consider the relative positions between nouns and verbs.

TABLE II
F-MEASURE BY DIFFERENT MODELS FOR CLASSIFYING SPANISH SENTENCES IN A DOCUMENT AS SUBJECTIVE AND OBJECTIVE IN MPQA GOLD DATASET.

Model	Type	F-measure
Rule Based [7]	Obj	0.52
	Subj	0.32
	Total	0.44
Bootstrapping [9]	Obj	0.52
	Subj	0.32
	Total	0.43
SVM [8]	Naïve Bayes	0.62
	SVM	0.62
LDNN	Obj	0.81±0.03
	Subj	0.87±0.02
	Total	**0.84±0.02**

Visualizing learned text features: To visualize the learned features we consider the 4-grams in the test set that show highest activation when convolved with the learned kernels. Here, we simply consider the root mean square error between predicted 4-gram kernel vectors and the prior word-vectors for each 4-gram learned using co-occurrence data. Table III shows features with highest activation at the hidden neurons in proposed LDNN for 'Subjective' and 'Objective' sentences in the Gold MPQA corpus. It can be seen that our method captures subjective and objective sentiments in 3-grams very accurately, the objective 3-grams are factual while the objective 3-grams are strongly positive or negative comments. It is apparent that by using both languages we can have a larger feature set.

C. Sentiment Classification on the TASS 2015 Corpus

The TASS corpora is a collection of Spanish tweets commonly used for the evaluation of social media analysis tasks. It is the evaluation framework used in different editions of the TASS workshop on Sentiment Analysis for Spanish [20].

In this paper, we are using the training set of 7219 tweets by 150 public figures coming from politics, sports or communication. The tweets were collected during the year 2011-2012. Each one is annotated with one of these four categories: *positive*, *neutral*, *negative*, or *without opinion*. This allow both to carry out coarse- and fine-grained sentiment analysis evaluations.

A test subset containing 1000 tweets with a similar distribution to the training corpus and manually labelled. This allows to counteract some of the limitations of the corpus made by pooling (e.g., most of the systems might fail the same tweet, assigning to this one a wrong label and the frequency distribution of the classes was not representative of the training set). Table IV shows accuracy by different models for classifying sentences in a document as Positive (Subjective), Negative (Subjective), Neutral (mixed) or None in TASS test dataset. A simple CNN model for sentences learns features of two or three words using sliding window kernels. We also compare our approach with different models evaluated at the TASS workshop (see the overview paper [20] for a detailed description of all approaches).

In LYS [34], the authors used classical logistic regression with linguistic features. Their approach was limited as they relied heavily on polarity lexicons that are not available in Spanish, instead in this paper we use convolution deep learning to automatically learn features from both English and Spanish. Further, instead of a single layer classifier, we learn hierarchy of feature detectors that can capture long reviews efficiently.

Visualizing Gain Correction in Features: Figure 3 shows shows a single predicted feature with (Y*) and without (Y) gain correction. It can be seen that by using a linear matrix inequality it is possible to amplify the underlying signal for easier classification.

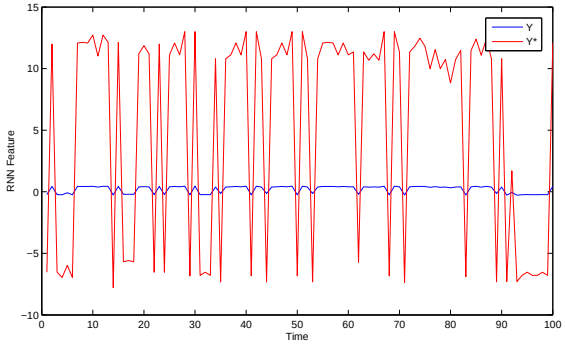

Fig. 3. This figure shows a single predicted feature with (Y*-red) and without (Y-blue) gain correction. It can be seen that by using a linear matrix inequality it is possible to amplify the underlying signal for easier classification.

VI. CONCLUSION

In this paper, we have proposed a framework for filtering noise or objective sentences automatically from Spanish text. This is achieved by extracting features from both Spanish and the translated English form using two separator deep neural networks. The features are combined using MKL.

The DNN has layers of CRBM followed by layers of RNN to capture temporal features. A linear matrix inequality has been developed to derive the stability criteria for multi-lingual subjectivity detection. Our results show that the proposed LDNN outperforms baselines on two benchmark datasets.

In this way we show that lexical resources in English such as Subjectivity Clues, Parts of Speech Tags and WSD can be effectively transferred to Spanish for large scale datasets such as Twitter.

ACKNOWLEDGEMENT

David Vilares is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180)

REFERENCES

[1] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 105–112.

[2] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification." in *ACL*, 2014, pp. 1555–1565.

TABLE III
FEATURES LEARNED AT THE HIDDEN NEURONS IN PROPOSED LDNN FOR 'SUBJECTIVE' & 'OBJECTIVE' SENTENCES IN THE GOLD MPQA CORPUS.

Model	English	Spanish
Subjective	sweet good considerate mother	Esto es un acertijo para (This is a riddle for)
	upset financially wise professional	de justicia social tan grande (as great social justice)
	money side suffers losses	guiar al golpeador más duro (guide the hardest puncher)
Objective	1899 Parliament erected statue	este siglo principalmente por (This century mainly by)
	plot modest rural cemetery	empleado americano de Edison Edwin (American Edison employee Edwin)
	well dominated television end	Nieman dijo que me quedara con (Nieman said stay with)

TABLE IV
ACCURACY BY DIFFERENT MODELS FOR CLASSIFYING SENTENCES IN A DOCUMENT AS POSITIVE(SUBJECTIVE), NEGATIVE(SUBJECTIVE), NEUTRAL(OBJECTIVE) OR NONE IN TASS DATASET.

	CNN [17]	LYS [34]	LIF [20]	LDNN
TASS 2015	0.66	0.637	0.692	**0.884**

[3] M. Bonzanini, M. Martinez-Alvarez, and T. Roelleke, "Opinion summarisation through sentence extraction: An investigation with movie reviews," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 1121–1122.

[4] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Statistical approaches to concept-level sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 6–9, 2013.

[5] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[6] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, 2005, pp. 486–497.

[7] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[8] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 127–135.

[9] C. Banea, R. Mihalcea, and J. Wiebe, "A bootstrapping method for building subjectivity lexicons for languages with scarce resources." in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. European Language Resources Association, 2008.

[10] S. Chandar, M. Khapra, B. Ravindran, V. Raykar, and A. Saha, "Multilingual deep learning," in *NIPS Deep Learning Workshop*, 2013.

[11] S. Chandar, S. Lauly, H. Larochelle, M. Khapra, R. Balaraman, V. Raykar, and A. Saha, "An autoencoder approach to learning bilingual word representations," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 1853–1861.

[12] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems," ser. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, 2010, vol. 5967, pp. 148–156.

[13] E. Cambria and A. Hussain, *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Cham, Switzerland: Springer, 2015.

[14] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain, "Sentiment data flow analysis by means of dynamic linguistic patterns," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 26–36, 2015.

[15] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *NAACL*, 2003, pp. 134–141.

[16] C. Fellbaum, Ed., *WordNet: an electronic lexical database*. MIT Press, 1998.

[17] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014, pp. 655–665.

[18] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 160–167.

[19] N. Kalchbrenner and P. Blunsom, "Recurrent convolutional neural networks for discourse compositionality," *CoRR*, vol. 1306.3584, 2013.

[20] J. Villena-Román, J. García-Morera, M. García-Cumbreras, E. Martínez-Cámara, M. Martín-Valdivia, and L. Ureña-López, "Overview of tass 2015," in *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, 2015.

[21] Z. Wang, Y. Liu, and X. Liu, "State estimation for jumping recurrent neural networks with discrete and distributed delays," *Neural Networks*, vol. 22, no. 1, pp. 41 – 48, 2009.

[22] I. Chaturvedi, E. Cambria, F. Zhu, L. Qiu, and W. K. Ng, "Multilingual subjectivity detection using deep multiple kernel learning," *Proceedings of KDD Wisdom, Sydney*, 2015.

[23] E. Cambria, J. Fu, F. Bisio, and S. Poria, "AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis," in *AAAI*, Austin, 2015, pp. 508–514.

[24] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, "EmoSenticSpace: A novel framework for affective common-sense reasoning," *Knowledge-Based Systems*, vol. 69, pp. 108–123, 2014.

[25] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space," in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, R. Setchi, I. Jordanov, R. Howlett, and L. Jain, Eds. Berlin: Springer, 2010, vol. 6279, pp. 385–393.

[26] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[27] R. Petrov, S. and Das, D. and McDonald, "A universal part-of-speech tagset," *arXiv preprint arXiv:1104.2086*, 2011.

[28] R. McDonald, J. Nivre, and et al, "Universal Dependency Annotation for Multilingual Parsing," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2013, pp. 92–97.

[29] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, vol. 2. Association for Computational Linguistics, 2011, pp. 42–47.

[30] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000, pp. 63–70.

[31] E. Brill and P. Resnik, "A Rule-based Approach to Prepositional Phrase Attachment Disambiguation," in *Proc. of COLING'94*, Kyoto, Japan, 1994, pp. 1198–1204.

[32] A. Moro, F. Cecconi, and R. Navigli, "Multilingual word sense disambiguation and entity linking for everybody." in *International Semantic Web Conference*, M. Horridge, M. Rospocher, and J. van Ossenbruggen, Eds., vol. 1272. CEUR-WS.org, 2014, pp. 25–28.

[33] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 216–225.

[34] D. Vilares, Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez, "Lys at tass 2014: A prototype for extracting and analysing aspects from spanish tweets," *Proceedings of the TASS workshop at SEPLN*, 2014.