Maestría en Ciencia de la Computación
Twitter Sentiment Analisis for Spanish in short texts

David Choqueluque Roman
Universidad Católica San Pablo
david.choqueluque@ucsp.edu.pe
Arequipa,Perú

Abstract: This paper presents the comparative analysis of three approaches to the classification of sentiments in Spanish tweets using the TASS 2017-2018 corpus. The classifiers are based on Recurrent Neuronal Network (RNN), Sentence embedding and Transfer Learning.

Keywords: Sentiment Analysis, RNN, Snetence Embedding, Transfer Learning.

I. INTRODUCTION

Sentiment analysis is one of the most important tasks related to subjectivity analysis within Natural Language Processing. The sentiment analysis of tweets is especially interesting due to the large volume of information generated every day, the subjective nature of most messages, and the easy access to this material for analysis and processing. The existence of specific tasks related to this field, for several years now, shows the interest of the NLP community in working on this subject. The Workshop on Semantic Analysis at the $SEPLN$ (in Spanish Taller de Analisis Semantico en la SEPLN, TASS) is the evolution of the Workshop on Sentiment Analysis at the SEPLN which is being held since 2012. In this line, Task 1 is focused on the evaluation of polarity classification systems at tweet level of tweets written in Spanish.

In this paper we describe different approaches for Spanish tweet classification based on RNN's, Sentence Embeddings [1] and Transfer Learning [2].

II. CORPUS PRE-PROCESSING

In the TASS dataset, training, development and test data was provided. The training and development sets were annotated with four possible polarity categories per tweet: P, N, NEU and NONE. The test corpora had no annotations. For our experiments we used the training, development and test sets from Spain(ES) from TASS2017 and TASS2018. In addition, we tested the three classifiers with the General1L dataset.

Table 1 shows the sizes of the different corpora and the number of tweets for each class.

Corpus	Category	Train	Dev	Test
General3L	P	1851	591	442
	N	1358	469	355
	NEU	437	131	102
	NONE	951	296	235
Total		4597	1487	1134
InterTASS-ES (2017)	P	318	156	-
	N	418	219	-
	NEU	133	69	-
	NONE	139	62	-
Total		1008	506	1899
InterTASS-ES (2018)	P	318	156	-
	N	418	219	-
	NEU	133	69	-
	NONE	139	62	-
Total		1008	506	1899

TABLE I
SIZE AND CATEGORIES DISTRIBUTION FOR THE DIFFERENT CORPUS.

For RNN and Sentence Embeddings classification methods each corpus was processed as follow:

- User mentions were removed.
- URLs links, punctuation, numbers and digits were removed.
- Sequences of three or more occurrences of the same character or syllable were replaced by a unique occurrence of that character. For instance, “holaaaa” was replaced by “hola”, ”jajaja”was replaced by ”jaja”.
- The emojis were removed.
- The text was converted to lowercase.

For the Transfer Learning method we apply the preprocessing recommended by the author as follow:

- User mentions were replaced by the token: ’user-ref’.
- Urls were replaced by the token: ’hyp-link’.
- For Hastags we add a prefix ’hash-tag’.
- For words combined with numbers we add a prefix ’int-string’.
- For slang words we add a prefix ’slang-string’.
- Sequences of three or more occurrences of the same syllable were replaced by token. For instance, ”jajajaj”was replaced by ’risa$_{j}a'$.

III. RESOURCES

III-A. Word Embeddings

For RNN and Sentence Embedding methods we used a 300 dimension word vectors of FastText in its Spanish version.

III-B. Pre-Training Model

For the Transfer Learning method we used a Wikipedia Spanish Language Model .

IV. Classifiers

IV-A. RNN - Bidirectional LSMT

The RNN classifier is a two layer Bidirectional LSTM recurrent network. The network has four layers: Embedding Layer, Spatial Droput 1D layer, bidirectional1 layer, bidirectional2 layer and dense layer. Table 2 shows the configuration of the keras RNN Architecture.

Layer	**Parameter**	**Value**
Bidirectional layers	LSTM units	200
Spatial 1D layer	Dropout	0.5
Bidirectional layers	Recurrent Dropout	0.5
Bidirectional layers	Spatial Dropout	0.4
Batch size		8
Epochs		15

TABLE II

RNN Arquitecture parameters.

IV-B. SIF - Sentence Embeddings

For this classifier we based on [1] that propose an unsupervised method to build sentence embeddings from each individual word embedding in a sentence. the metod is composed of two steps.

- First, compute the weighted average of the word vectors (where the weight is the SIF: Smooth Inverse Frequency) in the sentence.

$$SIF(w) = \frac{a}{(a + p(w))}$$

where a is a hyper-parameter and $p(w)$ is the estimated word frequency in the corpus.
- Subtract from the sentence embedding obtained in step before the first principal component of the matrix with all sentence embeddings as columns.

For the sentiment classification each tweet is convert to feature number vector using SIF, the classifier is a simple Neuronal Network of tree layers (Input layer and two dense layers).

IV-C. Transfer Learning

Universal Language Model Fine-tuning for Text Classification (ULMFiT) [2] is a new approach based on transfer learning for Computer Vision but by text. There are three stages in ULMFiT (Figure 1)

- General domain language model pre-training: This step is analogous to the ImageNet pre-training in computer vision. Since language modeling (next word prediction) can capture general properties of a language it serves as an ideal source task for pre-training a network. The network is pre-trained on a Wikitext pretrained model.
- Target task language model fine-tuning: In this step, the language model is fine-tuned on data from the target task (on which classification will be performed). This step improves the classification model (discussed in the third step) on small datasets.
- Target task classifier fine-tuning: The final stage in ULMFiT involves training the model with two additional linear blocks. ReLU activation is used for the intermediate layer and softmax for the final linear layer. Each block uses batch normalization and dropout.

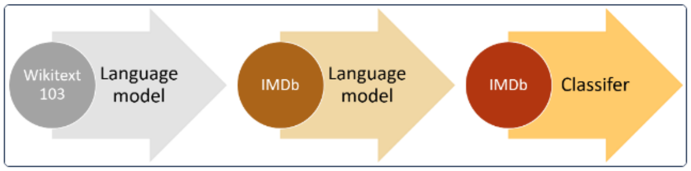


Fig. 1. ULMFiT stages [2].

V. Results

For compare the three methods we use the accuracy and F1 score metrics. The results is shows in Table 3.

Dataset	**General3L Corpus**		**TASS2017**		**TASS2018**	
Method	Accuracy	F1	Accuracy	F1	Accuracy	F1
RNN	0.5978	0.4548	0.5532	0.4613	0.5230	0.3750
SIF	0.6228	0.4331	0.5059	0.3988	0.5340	0.4200
ULMFiT	0.7504	0.5375	0.5450	0.4736	0.5100	0.3530

TABLE III

Accuracy and F1 score.

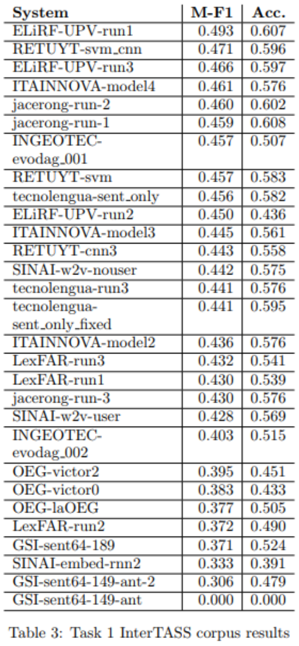
For TASS2017 and General3L corpus we compare our results with [3] (Figure 2-3). With the Task 1-TASS2017 corpus the best classifier in F1 metric is ULMFiT and would be second in the ranking of [3] (Figure 2).

For General corpus the best classifier is also ULMFiT in F1 metric would be first in the ranking of [3] (Figure 3).

For TASS2018 corpus the classifiers was compare with [4] (Figure 4). The best classifier for the dataset was SIF with 0,42 but it would be fifteenth in the ranking.

VI. Conclusions

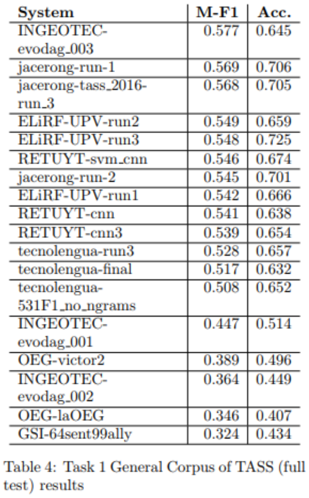
We presented three methods for Spanish Sentiment Analysis for TASS2017 and TASS2018. The approaches we used are: A Bidirectional-LSTM recurrent network using FastText word embeddings, a Sentence Embedding based classifier using FastText word embeddings and a Transfer Learning based classifier. Only the Transfer Learning based classifier is a considerably good model for TASS2017 and General corpus. We perform our experiments only with the corpus of a variant of Spanish (Monolingual ES). For future work, the three models should be further refined by modifying the hyperparameters of each one. According to the state of the art, NLP models based on ULMFiT give better results, which is why we would recommend paying more attention to the refinement of this.



System	M-F1	Acc.
ELiRF-UPV-run1	0.493	0.607
RETUYT-svm_cnn	0.471	0.596
ELiRF-UPV-run3	0.466	0.597
ITAINNOVA-model4	0.461	0.576
jacerong-run-2	0.460	0.602
jacerong-run-1	0.459	0.608
INGEOTEC-evodag_001	0.457	0.507
RETUYT-svm	0.457	0.583
tecnolengua-sent_only	0.456	0.582
ELiRF-UPV-run2	0.450	0.436
ITAINNOVA-model3	0.445	0.561
RETUYT-cnn3	0.443	0.558
SINAI-w2v-nouser	0.442	0.575
tecnolengua-run3	0.441	0.576
tecnolengua-sent_only_fixed	0.441	0.595
ITAINNOVA-model2	0.436	0.576
LexFAR-run3	0.432	0.541
LexFAR-run1	0.430	0.539
jacerong-run-3	0.430	0.576
SINAI-w2v-user	0.428	0.569
INGEOTEC-evodag_002	0.403	0.515
OEG-victor2	0.395	0.451
OEG-victor0	0.383	0.433
OEG-laOEG	0.377	0.505
LexFAR-run2	0.372	0.490
GSI-sent64-189	0.371	0.524
SINAI-embed-rnn2	0.333	0.391
GSI-sent64-149-ant-2	0.306	0.479
GSI-sent64-149-ant	0.000	0.000

Table 3: Task 1 InterTASS corpus results

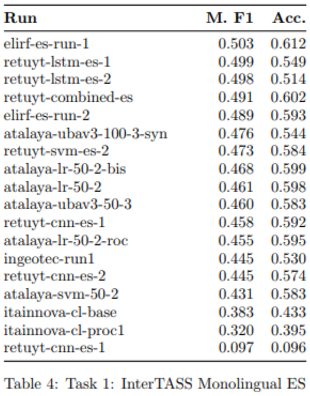
Fig. 2. TASS2017 Task 1 Results [3].



System	M-F1	Acc.
INGEOTEC-evodag_003	0.577	0.645
jacerong-run-1	0.569	0.706
jacerong-tass_2016-run_3	0.568	0.705
ELiRF-UPV-run2	0.549	0.659
ELiRF-UPV-run3	0.548	0.725
RETUYT-svm_cnn	0.546	0.674
jacerong-run-2	0.545	0.701
ELiRF-UPV-run1	0.542	0.666
RETUYT-cnn	0.541	0.638
RETUYT-cnn3	0.539	0.654
tecnolengua-run3	0.528	0.657
tecnolengua-final	0.517	0.632
tecnolengua-531F1_no_ngrams	0.508	0.652
INGEOTEC-evodag_001	0.447	0.514
OEG-victor2	0.389	0.496
INGEOTEC-evodag_002	0.364	0.449
OEG-laOEG	0.346	0.407
GSI-64sent99ally	0.324	0.434

Table 4: Task 1 General Corpus of TASS (full test) results

Fig. 3. TASS2017 General Corpus Results [3].



Run	M. F1	Acc.
elirf-es-run-1	0.503	0.612
retuyt-lstm-es-1	0.499	0.549
retuyt-lstm-es-2	0.498	0.514
retuyt-combined-es	0.491	0.602
elirf-es-run-2	0.489	0.593
atalaya-ubav3-100-3-syn	0.476	0.544
retuyt-svm-es-2	0.473	0.584
atalaya-lr-50-2-bis	0.468	0.599
atalaya-lr-50-2	0.461	0.598
atalaya-ubav3-50-3	0.460	0.583
retuyt-cnn-es-1	0.458	0.592
atalaya-lr-50-2-roc	0.455	0.595
ingeotec-run1	0.445	0.530
retuyt-cnn-es-2	0.445	0.574
atalaya-svm-50-2	0.431	0.583
itainnova-cl-base	0.383	0.433
itainnova-cl-proc1	0.320	0.395
retuyt-cnn-es-1	0.097	0.096

Table 4: Task 1: InterTASS Monolingual ES

Fig. 4. TASS2018 Monolingual ES Results [4].

References

[1] Sanjeev Arora, Yingyu Liang, Tengyu Ma,“A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR SENTENCE EMBEDDINGS”, 2017.

[2] Jeremy Howard, Sebastian Ruder, “Universal Language Model Fine-tuning for Text Classification”, 2018.

[3] Eugenio Martınez-Camara, Manuel C. Dıaz-Galiano, “Overview of TASS 2017”, 2017.

[4] Eugenio Martınez-Camara, Yudivian Almeida-Cruz, “Overview of TASS 2018: Opinions, Health and Emotions”, 2018.