Self-Attention for Twitter Sentiment Analysis in Spanish

José Ángel González [*], Lluís-F. Hurtado and Ferran Pla

Abstract.
This paper describes our proposal for Sentiment Analysis in Twitter for the Spanish language. The main characteristics of the system are the use of word embedding specifically trained from tweets in Spanish and the use of self-attention mechanisms that allow to consider sequences without using convolutional nor recurrent layers. These self-attention mechanisms are based on the encoders of the Transformer model. The results obtained on the Task 1 of the TASS 2019 workshop, for all the Spanish variants proposed, support the correctness and adequacy of our proposal.

Keywords: Twitter, Sentiment Analysis, Transformer Encoders

1. Introduction

Sentiment Analysis (SA) is one of the Natural Language Processing (NLP) problems that has been more studied in last decade. SA consists of determining if the polarity of a document is positive, negative or neutral.

Initially, SA models were trained to deal with long and nearly normative text, typically reviews of some products or services [13], [30]. As the use and the influence of Twitter have grown in last years, NLP community has incremented their efforts to address the peculiarities of the language in this social network. Nowadays, there is still great interest in the study of Sentiment Analysis in the Twitter domain. This is evidenced by the organization of different tracks devoted to this subject [18], [20].

Sentiment Analysis in Twitter presents some specific problems that do not occur in SA for normative text. On the one hand, the lack of context due to the limited length of the tweets. On the other hand, as in other social networks, the use of informal language is common in Twitter, that includes spelling errors, elongated words, the use of emoticons, special terms, user mentions, etc.

The workshop "Sentiment Analysis at SEPLN" (TASS) organized within the framework of the International Conference of the Spanish Society for Natural Language Processing (SEPLN[1]) is since 2012 the reference for the evaluation of systems for SA in Twitter for the Spanish language.

In last few years, self-attention mechanisms of the Transformer Encoder [26] have been proved as a very effective way for computing representations and complex relationships. They have been successfully used in some SA problems related to products and services reviews in English [12] [1].

In this work, we propose the use of these multi-head self-attention mechanisms, on top of pre-trained Twitter word embeddings, in order to address the Sentiment Analysis problem in Spanish on Twitter. To evaluate the adequacy of our proposal, we performed an

[*] Corresponding author. E-mail: jogonba2@dsic.upv.es

[1] http://www.sepln.org

extensive experimentation on Task 1 of the TASS 2019 workshop for several Spanish variants, where our system obtains very competitive results, being one of the best ranked systems in the competition.

The rest of the article is structured as follows. Section 2 presents the state-of-the-art for Twitter Sentiment Analysis both for English and Spanish. In Section 3, a description of the task addressed in this work is presented. In Section 4, we describe the architecture of the proposed system. Section 5 summarizes the conducted experimental evaluation, the achieved results and a qualitative analysis of the performance of the self-attention heads. Finally, some conclusions are shown in Section 6.

2. Related Work

Most works that addressed the SA problem have used polarity lexicons in some way. The construction of these lexicons is another widely explored field of research. Polarity lexicons have usually been constructed for English [13,30], but efforts have also been made to create lexicons for Spanish [19,21,15]. However, its use has declined over time due to the increase in the quality of representations, typically based on word or sentence embeddings.

The SemEval workshop has proposed several tasks related to Sentiment Analysis on Twitter from 2013 to 2017. In the last two editions [18] [20] many of the participating teams have included in their systems state-of-the-art deep learning approaches. In this respect, SemEval has become the reference for Sentiment Analysis on Twitter problem for the English language.

In SA for Twitter in Spanish, the most relevant workshop is the TASS workshop that has proposed different tasks for SA that focus on the Spanish language since 2012. An overview of the different tasks proposed, the participating teams, and the results obtained can be found in [27][28][29][7][14][4][6].

The Task 1 of TASS 2018 was focused in Sentiment Analysis at tweet level. The corpus provided by the organizers was InterTASS 2.0, including the Spain (ES), Peru (PE) and Costa Rica (CR) Spanish variants. Moreover, the organizers proposed two subtasks, Subtask 1 for monolingual SA and Subtask 2 for multilingual SA. The systems presented by [8] and [5] were the most competitive systems on the three Spanish variants for almost all the tasks while the system of [17] obtained the best results for the PE variant multilingual task.

In [8], the authors explore several deep learning architectures such as Deep Averaging Networks (DAN) [11], Attention Long Short Term Memory networks (Att-LSTM) [10] and Convolutional Neural Networks (CNN), along as different representations such as bag-of-words and Twitter word embeddings. In this case, the DAN system outperforms all the other participating systems in the ES variant.

Similarly, in [5], also were explored several deep learning architectures such as CNN and LSTM trained on top of Wikipedia word embeddings along as Support Vector Machines with a tweet representation based on word embeddings and several polarity statistics extracted from lexicons. Their LSTM and CNN systems are the first ranked systems for the CR and PE variants respectively.

The system proposed in [17] also was shown as the most competitive for the PE variant on the multilingual subtask. It is based on a genetic algorithm (EvoMSA) that orchestrates other subsystems. These subsystems are B4MSA [24] for tune input related hyper-parameters such as the normalization and the representation; and the classifier EvoDAG [9].

However, recent advances and mechanisms that have improved the NLP state-of-the-art have been published only for English SA. Meanwhile, in other languages such as the Spanish and its variants, these state-of-the-art advances are applied progressively in a slow way due to it is necessary to adjust them to work correctly in these languages.

One of these recent improvements is the proposal of the Transformer model in [26] for machine translation. This architecture is based on multi-head self-attention, dispensing with convolution and recurrences to learn relationships among words. The relationships captured by this kind of attention have shown to be effective on English SA tasks [1] [12] outperforming other systems based on Bidirectional LSTM and CNN on corpora such as Sentiment Stanford TreeBank [22] and SenTube [25].

3. Task Description

In order to validate our proposal for SA on Twitter in the Spanish language, we decided to participate in the Task 1 of TASS 2019.

This task consists on assigning global polarity to tweets on four classes $\mathbb{C} = \{N, NEU, NONE, P\}$.

Table 1

Number of tweets per class in all the sample sets of InterTASS for all the Spanish variants.

	ES			**CR**			**PE**			**UY**			**MX**		
	TR	DV	TS	TR	DV	TS	TR	DV	TS	TR	DV	TS	TR	DV	TS
N	474	266	663	310	143	459	228	107	485	367	192	587	505	252	745
NEU	140	83	195	91	55	151	170	56	368	192	90	290	79	51	119
NONE	157	64	254	155	72	220	352	230	176	94	51	82	93	48	111
P	354	168	594	221	120	336	216	105	435	290	153	469	312	159	525
Σ	1125	581	1706	777	390	1166	966	498	1464	943	486	1428	989	510	1500

Classes P and N refers to positive and negative sentiment respectively. Class NEU refers to the case where both positive and negative polarities are present in the tweet. The $NONE$ class is used for tweets which do not convey any polarity.

The organizers provided the InterTASS corpus composed by tweets from 5 different Spanish-speaking countries: Spain (ES), Peru (PE), Costa Rica (CR), Uruguay (UY) and Mexico (MX). For each Spanish variant 3 sample sets have been defined: training set (TR), development set (DV) and test set (TS). Only one Spanish variant can be used both for training and testing the system. Consequently, 5 different evaluations, one per Spanish variant, were proposed. Some statistics of the InterTASS corpus are shown in Table 1.

The InterTASS corpus is unbalanced and there is a bias towards the N and P classes, except in the training and development sets of the PE variant, where the most frequent class is $NONE$. However, in the test set of this variant, the class distributions differs, being N and P the most frequent classes. Moreover, the class NEU is usually the less populated class in all Spanish variants.

4. System Architecture

Our system is based on the Transformer [26] model. Initially proposed for machine translation, the Transformer model dispenses with convolution and recurrences to learn long-range relationships. Instead of this kind of mechanisms, it relies on multi head self-attention, where multiple attentions among the words of a sequence are computed in parallel to take into account different relationships among them. This reduces the computational complexity per layer (being also more parallelizable) and the max path length of dependencies among words to $\mathcal{O}(1)$ (instead of $\mathcal{O}(\log n)$ or $\mathcal{O}(n)$ in the cases of convolution and recurrent mechanisms respectively). This effect is particularly interesting on this task, where these dependencies can be given and there are few samples to learn them.

Concretely, we use the encoder part of the Transformer model in order to extract vector representations that are useful to perform Sentiment Analysis. We denote this encoding part of the Transformer model as Transformer Encoder (TE). Figure 1 shows the representation of the proposed architecture for the addressed task.

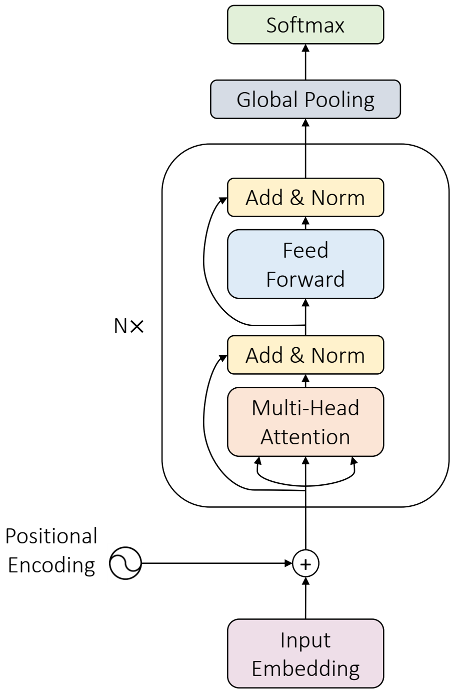

Fig. 1. System architecture based on the Transformer Encoder model.

The input of the model is a tweet $X = \{x_1, x_2, ..., x_T\ : x_i \in \{0, ..., V\}\}$ where T is the maximum length of the tweet and V is the vocabulary size. This tweet is passed through a d-dimensional pre-trained embedding layer, E, frozen during the training phase. Moreover, to consider positional information we also experimented with the sine and cosine functions proposed in [26].

This, encoded as $P \in \mathbb{R}^{T \times d}$ is added to the embedding representation of the tweet to be used as input to the first encoder layer $X^0 \in \mathbb{R}^{T \times d}$, as show in Eq 1.

$$X^0 = \{\overbrace{P_1 + E(x_1)}^{X_1^0}, ..., \overbrace{P_T + E(x_T)}^{X_T^0}\ : X_i^0 \in \mathbb{R}^d\} \tag{1}$$

After the combination of the word embeddings with the positional information, dropout [23] was used to drop input words with a certain probability p to regularize the model. On top of these representations, N transformer encoders are applied, which rely on the multi-head scaled dot-product attention shown in Eqs 2 - 4. These encoders are identical to [26], including the layer-normalized [2] residual connections.

$$MultiHead(A, B, C) = [head_1; ...; head_h]W^O \tag{2}$$

$$head_i = Attention(AW_i^Q, BW_i^K, CW_i^V) \tag{3}$$

$$Attention(Q, K, V) = softmax(\frac{QK^\intercal}{\sqrt{d_k}})V \tag{4}$$

where $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, $W_i^V \in \mathbb{R}^{d \times d_k}$, $W^O \in \mathbb{R}^{h \cdot d_k \times d}$, are the projection matrices for query, key and value of the head i and for the output of the multi-head attention respectively; and h is the number of heads for the multi-head attention mechanism.

The output for only one encoder, S, is computed as shown in Eq 8 for a given sample X^0.

$$M = MultiHead(X^0, X^0, X^0) \tag{5}$$

$$L = LayerNorm(X^0 + M) \tag{6}$$

$$F = max(0, LW_1 + b_1)W_2 + b_2 \tag{7}$$

$$S = LayerNorm(L + F) \tag{8}$$

where $M, L, F \in \mathbb{R}^{T \times d}$ are the intermediate outputs from the encoder, $W_1 \in \mathbb{R}^{d \times d_{ffw}}$, $W_2 \in \mathbb{R}^{d_{ffw} \times d}$ are the weights of the position-wise feed forward network, and $S \in \mathbb{R}^{T \times d}$ is the output of the encoder. When several encoders are stacked, the input of a encoder is used directly as input to the next encoder.

Due to a vector representation is required to train classifiers on top of these encoders, a global average pooling mechanism was applied on S. The resulting vector is used as input to a single-layer feed-forward network, whose output layer computes a probability distribution over the the four classes of the task $\mathbb{C} = \{N, NEU, NONE, P\}$.

We use Adam as update rule with $lr = 0.001, \beta_1 = 0.9$ and $\beta_2 = 0.999$ and Noam as learning rate schedule [26] with 15 $warmup\_steps$. Due to the imbalance in all the Spanish variants subsets, weighted cross entropy is used as loss function considering the distribution of each class in the training set. Concretely, we used the proportion between the most frequent class and the frequency of a given class, $w_i = \frac{\max_{c \in \mathbb{C}} n_c}{n_i}$, where n_i is the number of samples of the class i in a given set, being $w_i = 1$ if i is the most frequent class and $w_i > w_j$ if i is less frequent than the class j in the given sample set.

4.1. Resources and preprocessing

In order to initialize the embedding layer of our system with a rich semantic representation for the words of the task, a 300-d skipgram model [16] was trained on texts from the same domain of the task (Twitter). This model was trained by using 87M tweets from several Spanish variants, downloaded by streaming during several months in 2017 in our laboratory.

Regarding to the preprocessing, we have applied the same preprocess steps to all the given data, both the tweets used to learn the Word2Vec embeddings model and those provided by the organization to train the systems. Firstly, a case-folding process is applied to all the tweets, secondly, we tokenized the tweets by using TokTokTokenizer from NLTK [3]. Thirdly, user mentions, hashtags and URLS are replaced by three generic-class tokens (*user*, *hashtag* and *url* respectively). Finally, elongated tokens are diselongated allowing the same vowel to appear only twice consecutively in a token (e.g. *jaaaa* becomes *jaa*).

5. Experimental Work

In order to validate our proposal for Sentiment Analysis in Twitter and to select the best model to participate in the 2019 edition of TASS competition, we carried out some experimentation on the development set. To train the models, we fixed some hyper-parameters such as $batch\_size = 32$, $d_k = 64$, $d_{ff} = d$ and $T = 50$. Other hyper-parameters such as p, $warmup\_steps$ or h were set, considering the results obtained in previous experiments, to $p = 0.7$, $warmup\_steps = 5$ epochs and $h = 8$.

Moreover, we compared our proposal, which is based on Transformer Encoders (TE), with another deep learning systems such as Deep Averaging Networks (DAN) [11] and Attention Long Short Term Memory Networks [10] (Att-LSTM) that are commonly used in related text classification tasks and are the systems proposed by the teams that achieved best results in the 2018 edition of TASS [8][5].

We were interested to observe how the use of positional encodings or the number of encoders affect to the results obtained. Specifically, we train different models removing the positional information (TE-NoPos) and using 1 or 2 encoders. We tested all these combinations only on the ES variant and the best two configurations were also applied to the remaining variants (PE, CR, UY, MX).

The results in terms of macro-F_1 (MF_1), macro-recall (MR), macro-precision (MP) and Accuracy (Acc) achieved by all the systems considered in the development phase for all the Spanish variants are shown in Table 2. It can be seen that the best transformer encoders models (1-TE-NoPos, 2-TE-NoPos) outperform the DAN and Att-LSTM approaches by a margin of ∼5 points for MF_1 measure. This is due to the great improvement in both MR (∼6 points) and MP (∼3 points).

The use of the positional information in the TE approaches decreases the system performances (1-TE-Pos versus 1-TE-NoPos and 2-TE-Pos versus 2-TE-NoPos). This seems to indicate that the positional information, represented by sine and cosine functions added to the word embeddings, is not useful to the classification. However, the results obtained by Att-LSTM, which considers the positional information by its internal memory, obtains better results than the 1-TE-Pos and 2-TE-Pos approaches in almost all the metrics.

Table 2
Results on the development set for the different Spanish variants.

	MP	**MR**	MF_1	**Acc**
		ES		
DAN	47.66	48.46	47.94	56.28
Att-LSTM	50.00	48.14	48.83	58.00
1-TE-NoPos	52.80	**54.38**	**53.34**	60.75
1-TE-Pos	46.26	46.56	46.25	55.94
2-TE-NoPos	**52.85**	53.03	51.47	**61.27**
2-TE-Pos	47.31	48.79	47.71	56.11
		PE		
1-TE-NoPos	**49.06**	**50.43**	**49.51**	**54.62**
2-TE-NoPos	46.29	46.00	44.92	46.79
		CR		
1-TE-NoPos	**55.36**	**56.10**	**54.56**	**58.46**
2-TE-NoPos	52.14	52.36	51.71	55.13
		UY		
1-TE-NoPos	54.71	**56.63**	**54.83**	57.20
2-TE-NoPos	**55.82**	53.56	54.29	**58.64**
		MX		
1-TE-NoPos	**53.59**	55.03	**54.10**	**63.52**
2-TE-NoPos	52.78	**57.34**	54.07	60.78

The 1-TE-NoPos model obtains better results, in terms of MR and MF_1, than the 2-TE-NoPos model, outperforming its results on ∼2 points in terms of MF_1. This behavior is observed in almost all the variants, except in the MX variant, where both models obtain similar results in terms of MF_1 and 2-TE-NoPos outperforms 1-TE-NoPos in terms of MR.

Table 3 shows the results, at class level, achieved by the best model (1-TE-NoPos) for all Spanish variant. In most cases, the results obtained in the N and P classes are better than those obtained in the other classes, except in the PE variant, where the $NONE$ class is the one that obtains the best results. For all Spanish variants, as expected, the most difficult class is the NEU class due to the fact that this class corresponds to tweets that merge positive and negative sentiments.

Table 3

Results at class level for the 1-TE-NoPos model and all Spanish variants on the development set.

	N			NEU			NONE			P		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ES	**73.03**	**73.31**	**73.17**	30.56	26.51	28.39	46.34	59.38	52.05	61.25	58.33	59.76
PE	51.40	51.40	51.40	27.27	26.79	27.03	**64.88**	57.83	**61.15**	52.67	**65.71**	58.47
CR	**74.58**	61.54	**67.43**	27.87	30.91	29.31	46.09	**73.61**	56.68	72.92	58.33	64.81
UY	**69.70**	47.92	56.79	34.51	43.33	38.42	50.00	58.85	54.05	64.64	**76.47**	**70.07**
MX	**73.93**	**75.40**	**74.66**	30.91	33.33	32.08	44.07	54.17	48.60	65.47	57.23	61.07

In order to study in detail the behavior of our best system (1-TE-NoPos), we computed the confusion matrix for the ES variant that can be seen in Table 4. Note that, the NEU class is highly confused with the N and P classes. This seems to indicate that our model detects the presence of sentiment (positive or negative), but it is not capable to detect when both sentiments are present together. In addition, it can be observed that the N and P classes are confused with each other.

Table 4

Confusion matrix (1-TE-NoPos) on the ES variant development set.

	N	**NEU**	**NONE**	**P**
N	195	25	18	28
NEU	25	22	13	23
NONE	9	6	38	11
P	38	19	13	98

In light of the results of the development phase, we decided to use 1-TE-NoPos system to participate in the TASS 2019 competition. Table 5 shows the official results for all Spanish variants and the position of our system (ranked using F_1 measure) in each variant [6]. As it can be seen, our system is ranked in first place for the ES and MX variants and in second place for CR, PE and UY variants.

Highlight that although our system has been optimized for the ES variant, it has behaved reasonably well for the rest of variants.

5.1. Qualitative Analysis

With the aim of understanding the proposed model, we have analyzed the behavior of the self-attention mechanisms.

Table 5

Official results and ranking of our system on the TASS 2019 competition.

	$\boldsymbol{MF_1}$	**MP**	**MR**	**Rank**
ES	50.68	50.52	50.85	1/10
CR	49.58	49.84	49.33	2/10
PE	44.74	45.63	43.82	2/10
UY	51.54	49.68	53.55	2/8
MX	50.10	49.05	51.21	1/10

A competitive SA system should be able to combine several aspects to determine the polarity of a tweet. Among others, some of these aspects are the polarity of the words and the presence of sentiment modifiers such as polarity shifters or reversers in the tweets. We hypothesize that the attention heads of our system should capture some of these aspects.

In order to determine what heads react to these aspects, we computed the average attention that each word receives from each head considering all the occurrences of the word in a given sample set. We formalized this computation in Algorithm 1.

The development set of the ES variant is used to verify that the model generalizes and captures interesting relationships even in samples that it has never seen.

From the set of samples χ with vocabulary $\mathcal{V}$ and the trained model Θ, it is possible to calculate the attention given by the head k to a word w in the sample set χ. To do this, from each sample x of the set χ and each head k, the matrix $B \in \mathbb{R}^{|x| \times |x|}$, which contains the attentions of this head after a forward pass on the model Θ, is computed.

The columns of this attention matrix are averaged to obtain $B' \in \mathbb{R}^{|x|}$. This matrix B' contains the attention that head k gives to each word in x, computed as the average of the self-attentions in the head. Finally, the attention of each word in each head, α_{wk}, is calculated by averaging the attention given by head k to word w in all the samples.

Algorithm 1 Compute the average word attentions captured by the model on a set of samples.

Input: $\mathcal{V}$ vocabulary, set of samples χ, trained Transformer Encoder Θ

Result: α_{wk} the average attention of head k for word w

```
procedure COMPUTEWORDATTENTIONS($\chi$, $\Theta$)
    for $w \in \mathcal{V}$ do
        for $1 \leq k \leq h$ do
            $\alpha_{wk} \leftarrow 0$
        end for
    end for
    for $x \in \chi$ do
        for $1 \leq k \leq h$ do
            $B \leftarrow softmax(\frac{\Theta(x)_{Q_k} \Theta(x)^{\top}_{K_k}}{\sqrt{d_k}})$
            $B' \leftarrow \frac{1}{|x|} \sum_{i=1}^{|x|} B_{ij}$
            for $w \in x$ do
                $\alpha_{wk} \leftarrow \alpha_{wk} + B'_w$
            end for
        end for
    end for
    for $w \in \mathcal{V}$ do
        for $1 \leq k \leq h$ do
            $\alpha_{wk} \leftarrow \frac{\alpha_{wk}}{c_w}$
        end for
    end for
end procedure
```

Once α is computed, it is possible to observe if some heads are capable of taking into account some properties at word level that are necessary to determine the sentiment of a tweet.

Figure 2 shows the attention of all heads (from 1 to 8) for 6 words with high polarity. These words are extracted from the ElHuyar [21] lexicon. First row in Figure 2 shows the attention per head of three words with positive polarity (*best*, *wonderful* and *cool*) and the second row corresponds to three words with negative polarity (*worst*, *horrible* and *shit*). It can be observed that the attention heads 4 and 5 react with high intensity when the polarity is negative and positive respectively. Moreover, head 4 does not react when the polarity is positive, the same behavior is observed for head 5 when the polarity is negative. Furthermore, heads 6 and 7 seem to attend to the negative words and not to the positive ones; head 3 reacts more intensively to positive words rather than negative ones.

We extended the study to all words in the vocabulary that appear in the ElHuyar polarity lexicon [21]. Figure 3 shows average attentions per head for positive and negative words. It can be seen that the negative words receive higher attention than the positive ones. In particular, head 4 reacts more to negative words than head 5 reacts to positive words.

To confirm the capability of the heads 4 and 5 detecting the polarity of the words, we designed a classifier that uses only the attention of heads 4 and 5 (α_{w4} and α_{w5}) to determine the polarity of each word w of the vocabulary $\mathcal{V}$. This classifier is formalized in Eq. 9.

$$\mathcal{C}(w) = \begin{Bmatrix} P & \text{if } \alpha_{w4} \leq \alpha_{w5} \\ N & \text{if } \alpha_{w5} < \alpha_{w4} \end{Bmatrix} \tag{9}$$

We tested the performance of classifier $\mathcal{C}$ by classifying all words of ElHuyar lexicon that appear in the vocabulary. Note that the words in ElHuyar have only positive or negative polarity. The classifier achieved an Accuracy of 74.75% which confirms the ability of the attention heads 4 and 5 capturing the polarity at word level.

We attempted to address the Task 1 of TASS 2019 for ES variant using only the information of heads 4 and 5 and the ElHuyar lexicon. To do this, we designed a classifier based on the sum of the polarity of the words. The classifiers works as follow: if the sample does not contain any word with polarity its class is $NONE$, if the sample contains the same number of positive and negative words its class is NEU, otherwise the class of the sample is P or N depending of the number of positive and negative words.

This classifier is directly computable on any polarity lexicon (e.g ElHuyar), however to use heads 4 and 5 of our system we need to design a mechanism to discretize the polarity of each word based on the outputs of both heads. In our case, we obtain a probability distribution over the P and N classes by means of a softmax function on the output of the two heads. To discretize this function, we used a threshold $\epsilon = 0.165$ experimentally set. This classifier, SumPolClassifier, is defined in the Algorithm 2.

In order to use the SumPolClassifier with ElHuyar lexicon, $p(N|w)$ and $p(P|w)$ are obtained directly from the lexicon. Table 6 shows the results of SumPolClassifier applied to the development set of the ES variant of Task 1 of TASS 2019 both with heads 4 and 5, and ElHuyar lexicon. It can be seen how the results in terms of macro-F_1 are similar in both approaches. Both systems classify similarly the classes

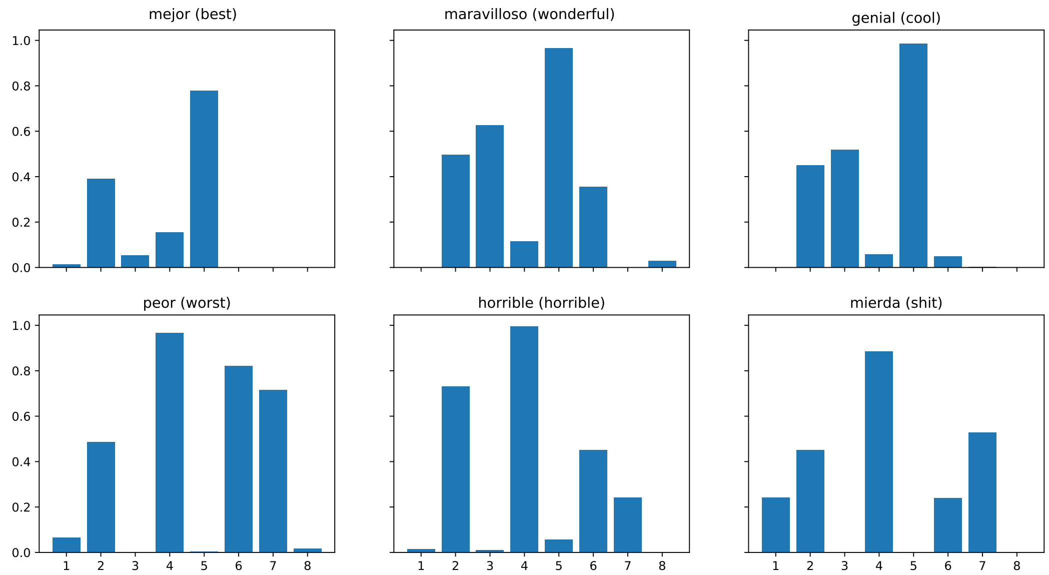

Fig. 2. Attentions for several words that contains sentiment.

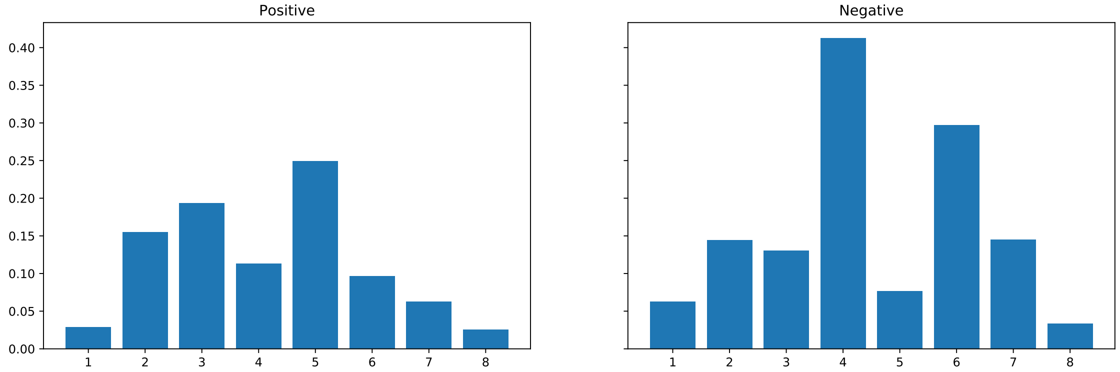

Fig. 3. Sum of attentions for all the attention heads on the words of ElHuyar.

NEU, $NONE$ and P. However, the recall on the class N with the heads 4 and 5 is significantly lower than with ElHuyar although they have more precision.

Finally, we studied how attention heads react to words that are supposed to be polarity shifters or polarity reversers. Figure 4 shows average attentions per head for eight of these words. The words in the first row (*not*, *never*, *neither* and *anybody*) are polarity reversers and the words in the second row (*very*, *nothing*, *forever* and *something*) are polarity shifters.

It can be seen that head 1 reacts to all the shifters and reversers. This head do not react to positive or negative words (see Figures 2 and 3). In addition, heads 4 and 5 do not react to shifters nor reversers because these words do not have polarity per se. However, the

Table 6

Results of SumPolClassifier both using the heads 4 and 5, and ElHuyar lexicon on the development set.

	Heads 4/5			**ElHuyar**		
	P	R	F_1	P	R	F_1
N	63.73	41.14	50.00	62.10	53.59	57.53
NEU	14.05	24.29	17.80	16.25	18.57	17.33
NONE	23.45	33.76	27.68	27.32	31.85	29.41
P	57.26	56.78	57.02	56.30	59.32	57.77
Macro	39.62	38.99	38.12	40.49	40.83	40.51

Algorithm 2 SumPolClassifier based on the heads 4 and 5 to classify the polarity of tweets.

Input: sample set χ and α the attentions per head of all word w in the vocabulary $\mathcal{V}$.

Result: $\hat{y}$, labels assigned by the classifier to all samples in the sample set.

```
procedure SumPolClassifier($\chi, \alpha$)
    for $x \in \chi$ do
        pol $\leftarrow 0$
        neutralized $\leftarrow$ false
        for $w \in x_i$ do
            $p(N|w) \leftarrow \frac{e^{\alpha_{w4}}}{e^{\alpha_{w4}} + e^{\alpha_{w5}}}$
            $p(P|w) \leftarrow \frac{e^{\alpha_{w5}}}{e^{\alpha_{w4}} + e^{\alpha_{w5}}}$
            if $|p(N|w) - p(P|w)| \geq \epsilon$ then
                neutralized $\leftarrow$ true
                if $p(N|w) > p(P|w)$ then
                    pol $\leftarrow$ pol - 1
                else
                    pol $\leftarrow$ pol + 1
                end if
            end if
        end for
        if pol > 0 then
            $\hat{y}_x \leftarrow P$
        else
            if pol < 0 then
                $\hat{y}_x \leftarrow N$
            else
                if neutralized then
                    $\hat{y}_x \leftarrow NEU$
                else
                    $\hat{y}_x \leftarrow NONE$
                end if
            end if
        end if
    end for
end procedure
```

attention values for head 1 are not relatively high except in the case of *no* and *always*. These results seem to indicate that, although it reacts fairly well to common shifters and reversers, it is necessary to reinforce the attentions dedicated to this type of words.

It is also remarkable that all the polarity reversers and the polarity shifter *nothing*, all of them with negative inertia, are attended by head 7 that was related to the negative polarity as previously discussed.

6. Conclusions

In this work, we have presented a proposal for Sentiment Analysis in Twitter for the Spanish language. Our proposal is based on the use word embedding trained from tweets in Spanish and the Transformer Encoder architecture. This architecture relies only on self-attention mechanisms to ease the learning of relationships among words, without using convolutional or recurrent mechanisms.

We have tested our system on the Task 1 of the 2019 edition of the TASS workshop for which the organizers provided 5 subsets corresponding to 5 Spanish variants. Although the hyper-parameters of the model had been tuned considering only the ES variant, our system was ranked first or second on all the Spanish variants.

These results have encouraged us to perform a thorough study of how the self-attention heads capture the information required to perform Sentiment Analysis. We have detected some heads directly related with positive and negative words and another that reacts to polarity shifters and reversers.

7. Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R) and the GiSPRO project (PROMETEU/2018/176). Work of José-Ángel González is financed by Universitat Politècnica de València under grant PAID-01-17.

References

[1] A. Ambartsoumian and F. Popowich. Self-attention: A better building block for sentiment analysis neural network classifiers. In *WASSA@EMNLP*, 2018.

[2] L. J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[3] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[4] E. M. Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. G. Cumbreras, M. G. Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and J. Villena-Román. Overview of TASS 2018: Opinions, health and emotions. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 13–27, 2018.

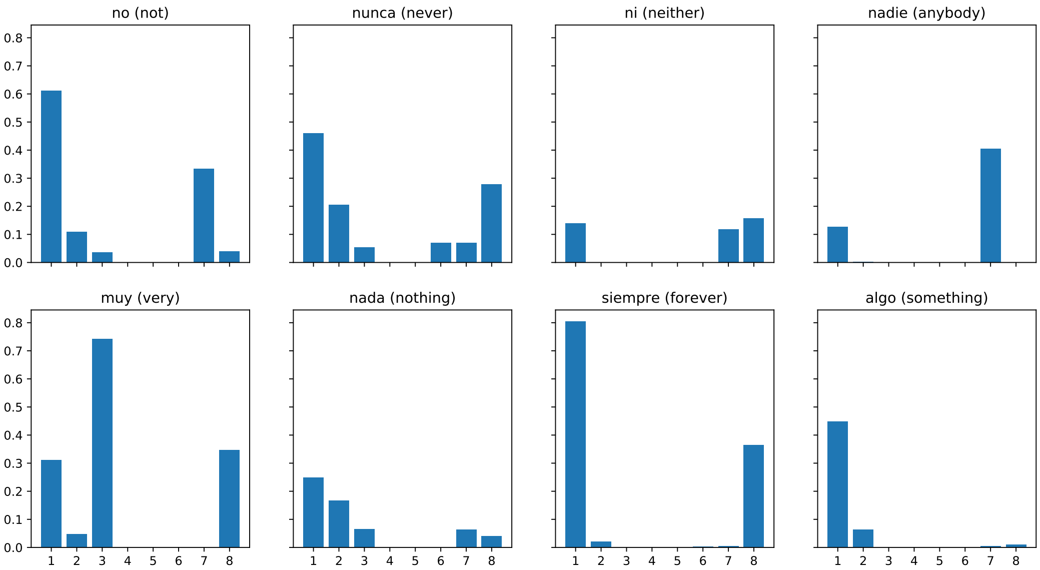

Fig. 4. Attention per head on polarity reversers and shifters.

[5] L. Chiruzzo and A. Rosá. Retuyt-inco at TASS 2018: Sentiment analysis in spanish variants using neural networks and SVM. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 57–63, 2018.

[6] M. C. Díaz-Galiano et al. Overview of TASS 2019. Bilbao, Spain, 2019. CEUR-WS.

[7] M. Á. García Cumbreras, J. Villena Román, E. Martínez Cámara, M. C. Díaz Galiano, M. T. Martín Valdivia, and L. A. Ureña López. Overview of TASS 2016. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September 13th, 2016.*, pages 13–21, 2016.

[8] J. González, F. Pla, and L. Hurtado. Elirf-upv en TASS 2018: Análisis de sentimientos en twitter basado en aprendizaje profundo (elirf-upv at TASS 2018: Sentiment analysis in twitter based on deep learning). In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 37–44, 2018.

[9] M. Graff, E. S. Tellez, S. Miranda-Jiménez, and H. J. Escalante. EvoDAG: A semantic Genetic Programming Python library. In *2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, pages 1–6, Nov. 2016.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[11] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics.

[12] G. Letarte, F. Paradis, P. Giguère, and F. Laviolette. Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 267–275, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[13] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[14] E. Martínez-Cámara, M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, and J. Villena-Román. Overview of TASS 2017. In J. Villena Román, M. A. García Cumbreras, E. Martínez-Cámara, M. C. Díaz Galiano, and M. García Vega, editors, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, CEUR Workshop Proceedings, pages 13–21, Murcia, Spain, September 2017. CEUR-WS.

[15] E. Martínez-Cámara, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López. Bilingual Experiments on an Opinion Comparable Corpus. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, page 87–93, 2013.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[17] D. Moctezuma, J. Ortiz-Bejar, E. S. Tellez, S. Miranda-Jiménez, and M. Graff. INGEOTEC solution for task 1 in tass'18 competition. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 45–49, 2018.

[18] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoy-

anov. SemEval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June 2016. Association for Computational Linguistics.

[19] V. Perez-Rosas, C. Banea, and R. Mihalcea. Learning Sentiment Lexicons in Spanish. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[20] S. Rosenthal, N. Farra, and P. Nakov. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

[21] X. Saralegi and I. San Vicente. Elhuyar at TASS 2013. In *Proceedings of the TASS workshop at SEPLN 2013*. IV Congreso Español de Informática, 2013.

[22] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[24] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, R. R. Suárez, and O. S. Siordia. A Simple Approach to Multilingual Polarity Classification in Twitter. *Pattern Recognition Letters*, 2017.

[25] O. Uryupina, B. Plank, A. Severyn, A. Rotondi, and A. Moschitti. SenTube: A corpus for sentiment analysis on YouTube social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4244–4249, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[27] J. Villena-Román, J. García Morera, M. Á. García Cumbreras, E. Martínez Cámara, M. T. Martín Valdivia, and L. A. Ureña López. Workshop on Sentiment Analysis at SEPLN 2013: An overview. In *Proceedings of the TASS workshop at SEPLN 2013*. Villena-Román, Julio; García Morera, Janine; García Cumbreras, Miguel Ángel; Martínez Cámara, Eugenio; Martín Valdivia, M. Teresa; Ureña López, L. Alfonso, 2013.

[28] J. Villena-Román, J. García Morera, M. Á. García Cumbreras, E. Martínez Cámara, M. T. Martín Valdivia, and L. A. Ureña López. Workshop on Sentiment Analysis at SEPLN: Overview. In *Proceedings of the TASS workshop at SEPLN 2014*. Villena-Román, Julio; García Morera, Janine; García Cumbreras, Miguel Ángel; Martínez Cámara, Eugenio; Martín Valdivia, M. Teresa; Ureña López, L. Alfonso, 2014.

[29] J. Villena-Román, J. García Morera, M. Á. García Cumbreras, E. Martínez Cámara, M. T. Martín Valdivia, and L. A. Ureña López. Overview of TASS 2015. In *TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015).http://ceur-ws.org/Vol-1397/*, pages 13–21. Villena-Román, Julio; García Morera, Janine; García Cumbreras, Miguel Ángel; Martínez Cámara, Eugenio; Martín Valdivia, M. Teresa; Ureña López, L. Alfonso, 2015.

[30] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.