Transformer based Contextualization of Pre-trained Word Embeddings for Irony Detection in Twitter

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

VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València
Camí de Vera sn, 46022, València, Spain

Abstract

Human communication using natural language, specially in social media, is influenced by the use of figurative language like irony. Recently, several workshops are intended to explore the task of irony detection in Twitter by using computational approaches.

This paper describes a model for irony detection based on the contextualization of pre-trained Twitter word embeddings by means of the Transformer architecture. This approach is based on the same powerful architecture as BERT but, differently to it, our approach allows us to use in-domain embeddings.

We performed an extensive evaluation on two corpora, one for the English language and another for the Spanish language. Our system was the first ranked system in the Spanish corpus and, to our knowledge, it has achieved the second-best result on the English corpus. These results support the correctness and adequacy of our proposal.

We also studied and interpreted how the multi-head self-attention mechanisms are specialized on detecting irony by means of considering the polarity and relevance of individual words and even the relationships among words. This analysis is a first step towards understanding how the multi-head self-attention mechanisms of the Transformer architecture address the irony detection problem.

Keywords: Irony Detection, Twitter, Deep Learning, Transformer Encoders

1. Introduction

Human communication using natural language in social environments is influenced by the use of figurative language. Unlike literal language, where the meaningful units used in the interactions convey exactly the meaning that the

*Corresponding author.

Email addresses: jogonba2@dsic.upv.es (José Ángel González), lhurtado@dsic.upv.es (Lluís-F. Hurtado), fpla@dsic.upv.es (Ferran Pla)

author wants to express, the figurative language aims to use words differently from the usual way, in order to transmit complex ideas in a more creative way. One of the most interesting rhetorical devices is the irony. Irony takes place in ambiguous situations where the literal meaning is opposite to the knowledge that the author have of the world and it is wanted to be transmitted [1]. Nowadays, irony is typically used in social networks with the aim of favoring social interactions, evoking humor [1], diminishing or enhancing criticism [2], and getting the attention of the readers by means of the creativity [3].

The New Princeton Encyclopedia of Poetry and Poetics [4] identify eight different types of irony: classical, romantic, tragic, cosmic, verbal, situational, dramatic and poetic irony. The most common types of irony used in the social networks are situational and verbal irony. On the one hand, situational irony is related to incongruous situations about specific events e.g. "A security company is the last victim of a malware attack". On the other hand, verbal irony has been defined by several authors [1, 5, 6] as the communication of a meaning opposite to the literal meaning, e.g. "Oh look, we are having another storm in Sydney. How unusual.". A specific form of verbal irony is the sarcasm, which also has been studied in the literature [7]. It is a subset of verbal irony where the aim of a message is to make a harmful criticism about someone or something.

The detection of the irony in text messages is a complex and subjective problem affected by a plethora of phenomena. Several relevant features have been identified to address the irony detection problem: polarity contrast [8], common sense knowledge [9], similes with "about" or "as" structures [3], punctuation marks or repetitions [10], affective features [11, 12], negation [13], contextual features [14], context incongruity [15], etc. However, computational approaches that follow the principle of text compositionality are not capable of explaining the textual irony only by means of the composition of the words of a message [16], mainly due to the fact that many irony markers are lost in the text message, such as kinesthetic (facial or hand gestures) [17] or speech features (voice tone, rhythm, silences, etc.) [18].

Irony also has a great impact on some computational approaches for Natural Language Processing (NLP) tasks in social media such as sentiment analysis [19][20][7], author profiling or deception detection [21], where the systems struggle when they are applied to ironic content. In order to boost the research on irony detection for several languages, different workshops have been organized [22, 23, 24] with the aim of improving the understanding about the irony and to diminishing the faults of the computational approaches for NLP tasks when they are applied on ironic content.

Currently, most of the systems for irony detection are based on Deep Learning architectures. Typically, these systems are built upon non-contextual word representations [25] which are contextualized by using architectures such as Convolutional Neural Networks [26] or Long Short Term Memory (LSTM) [27]. Recently, the use of pre-trained contextualized word embeddings has been widely spread by means of the BERT model [28]. However the use of the main mechanism of BERT, namely Transformers [29], has not been explored for contextualizing pre-trained word embeddings in irony detection tasks.

In this work, we propose the use of the Transformer architecture in order to contextualize pre-trained word embeddings. Specifically, we contextualize Word2Vec word embeddings, trained with several millions of tweets both for the English and the Spanish languages. This strategy, opposite to the use of pre-trained BERT, allows our system to be trained from in-domain representations using the same powerful backbone architecture as BERT. We evaluated the adequacy of our proposal on two corpora. For the Spanish language, the corpus of the Irony Detection on Spanish Variants shared task (IroSVA) [23] was used. For the English language, we used the dataset of the task 3: Irony Detection in English Tweets proposed in 2018 at the 12th International Workshop on Semantic Evaluation (SemEval) [22]. Our system was the first-ranked system in the IroSVA competition and, to our knowledge, it has achieved the second-best result on the SemEval corpus. The implementation of our system is freely available, under request, for research purposes. Additionally, we propose several analysis and algorithms in order to determine how the multi-head self-attention mechanisms of the Transformer are specialized in detecting ironic messages, with the aim of observing how the polarity, the relevance of individual words and the relationships among words, influence the irony detection problem. Next we summarize the main contributions of this paper:

- To study the irony detection problem for the English and the Spanish languages on two widely used corpora.
- To present an approach based on Transformer Encoders for contextualizing pre-trained Twitter word embeddings. This system was the first ranked system in the IroSVA competition and, to our knowledge, it has achieved the second-best result on the corpus of SemEval.
- To propose several analysis strategies towards the understanding of the behavior of Transformer Encoder models and the features captured by them when addressing the irony detection problem e.g. word polarity and relationships among words.
- To provide, under request, the implementation of our system for research purposes.

The rest of this paper is structured as follows. In Section 2, the state of the art for irony detection both for the English and the Spanish languages is presented. In Section 3, we formalized our proposal based on the Transformer Encoder architecture. In Section 4, we describe the corpora, the resources, the preprocessing and the experimental setup. In Section 5, we present an exhaustive evaluation of our proposal and a comparison with other Deep Learning approaches. In Section 6, several analyses are presented to study how our system takes into account some features related with the irony. Finally, in Section 7, the conclusions extracted after evaluating and analyzing our proposal are presented.

2. State of the Art

Several workshops have been organized to address the irony detection problem for different languages such as Spanish [23], English [22], Italian [24] and Arabic [30].

For the Spanish language, the IroSVA shared task [23] was proposed within the 35th International Conference of the Spanish Society for Natural Language Processing (SEPLN). The aim of IroSVA was to identify the presence of the irony in tweets for three Spanish variants. A peculiarity of this task is that each tweet of the corpus has an associated context that consists of a short sequence of words that identify the scope of the tweet, e.g. “flat earth” or “book of Pedro Sánchez” (referencing the controversial book written by the Spanish prime minister).

Among the systems proposed by the participants for the task, the two best systems were based on Deep Learning approaches either as classifiers or as feature extractors. The best system was presented by our team [31]. It was based in the use of Transformer encoders, relying on multi-head scaled dot-product attention mechanisms, in order to contextualize pre-trained Twitter word embeddings. A formalization of the model along with an extensive evaluation and result analysis both for Spanish and English languages is the scope of the current work.

The second-ranked team in the IroSVA competition [32] experimented with the early fusion of traditional features (TF-IDF weighted n-grams) and distributed features (pre-trained word embeddings and the internal representation of a pre-trained LSTM for the task). As classifiers, they used Support Vector Machines (SVM) and Multi Layer Perceptron on top of the input features. Contrary to these two approaches, the third most competitive approach [33] does not rely on Deep Learning architectures. In this case, the authors were interested in observing how several dependency-based features contribute to the irony detection. Concretely, they proposed the use of bag of dependency relations, bag of syntax paths and bag of dependency relations to train Random Forests and SVM models.

For the English language, the task 3 of SemEval 2018 [22], co-located with North American Chapter of the Association for Computational Linguistics (NAACL), aims to boost the work on irony detection on English tweets. In this workshop, two different subtasks were proposed. The first subtask consists in addressing the irony detection as a binary classification problem, whereas for the second subtask, the participants should distinguish among three different types of irony: verbal irony by means of polarity contrast, other verbal irony, and situational irony.

Most of the participants addressed both subtasks by using Deep Learning approaches. The best system [34] was based on the use of Densely connected Bidirectional LSTM (D-BiLSTM) on top of a combination of word embeddings with Part of Speech Tags. Moreover, the system used a late fusion of the D-BiLSTM representations, several sentiment features (generated via the AffectiveTweets package of Weka), and a vector representation of the tweets generated by averaging the word embeddings. The system was trained to simultaneously solve three tasks, the two subtasks of the competition together with a hashtag prediction task. The authors of the second-best ranked system [35] proposed an ensemble of two Attentional LSTM (Att-LSTM) which share the same architecture but operate on two different representation levels: words and characters. Both networks are only different on the first embedding layer. For the word level, the embedding layer was initialized with pre-trained word representations learned from 550M English tweets. Regarding the character level, the embedding layer was randomly initialized and learned during the training of the model for the subtask of irony detection. In order to perform the ensemble of the two Att-LSTM, the authors tested two different approaches: unweighted average and majority voting.

The third best ranked work [36] studied how the sentiment, distributional semantics, and text surface features were related with the irony. The main effort of their work relies on detecting the polarity contrast at two different levels: polarity contrast between the same element of a tweet e.g. antithetical fragmented hashtags, and polarity contrast between two different elements of a tweet e.g. words and emojis sentimentally opposed. Also, they detected that in most ironic tweets, negative polarity is preceded by neutral or positive polarity. Therefore, they decomposed the tweets to take also into account these temporal relations. Both the polarity contrast and the surface features were combined with word embeddings and they were used as input to an ensemble soft voting classifier based on Logistic Regression and SVM paradigms. It is interesting to note that, most of the participating teams addressed the tasks by using emotional and polarity features in order to enrich their systems with the aim of explaining the irony by means of polarity contrast [37][38][39].

In addition to the works proposed in conference tasks, a lot of efforts have been made in order to analyze relevant features for irony detection. The most studied phenomenon is the impact of the polarity in the irony detection problem [40][11][41]. Also, the work presented in [36] shown that, in certain context, too much of an emotion can imply the opposite sentiment, generating some kind of irony. Some works are focused on detecting implicit incongruencies among positive and negative words [42, 43]. In [42], the authors enrich the supervised learning on irony detection tasks by transferring knowledge from sentiment resources. They proposed three different Att-LSTM approaches that differ in the way of including the sentiment resources, either injecting the sentiment directly to the attention mechanisms or merging the output of different networks specialized on sentiment analysis and irony detection. In [43], the authors focused on identifying contrasting contexts, that is, positive sentiment followed by a negative situation. They learned a list of positive and negative phrases, using a bootstrapping algorithm, that are used for recognizing sarcasm in tweets.

Recently, pre-trained contextualized BERT embeddings [28][44][45] become ubiquitous in many text classification tasks, and they have been progressively applied to the irony and sarcasm detection problems [46][47][48][49]. In [47], the authors finetuned the multilingual BERT for the IroSVA task and they compare the results with classical techniques for text classification such as SVM and Gradient Tree Boosting. In [48], the authors make a further pretraining of the multilingual BERT model with the Twitter domain, and they finetune the models under a multi-task setup for addressing irony detection, author profiling and emotion detection in Arabic tweets. In [49], the sarcasm detection problem is addressed by using multimodal information such as speech, videos, and text. Pre-trained BERT was used to represent the textual utterances, showing a better performance than other strategies such as averaging GloVe word vectors [50]. In [46], the pre-trained RoBERTa [45] model was used to represent the sentences, that was further contextualized by means of a Recurrent Convolutional Neural Network to address irony and sarcasm detection. All these previous works are based on using pre-trained BERT models either for finetuning them or for extracting sentence representations. Our work differs from them because it does not use the contextual representations learned from BERT, and instead, it is based on the backbone network of the BERT models (Transformer Encoders) for contextualizing Word2Vec word embeddings pretrained on the task domain (Twitter).

3. System Architecture

The system presented in this work is based on the Transformer model [29]. Initially proposed for machine translation, the Transformer model dispenses with convolutions 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 maximum 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 for the irony detection tasks addressed in this work because word dependencies are very relevant, and the corpora used have few samples to learn these dependencies.

Concretely, we used the encoder part of the Transformer model in order to extract vector representations that are useful to perform irony detection. We denote this encoding part of the Transformer model as Transformer Encoder (TE). Figure 1 shows a scheme of the proposed architecture.

Let $\mathbb{C} = \{0, 1\}$ be the set of classes (0 denotes the non-ironic class and 1 denotes the ironic class), $X = \{x_1, x_2, ..., x_T \ : x_i \in \{0, ..., V\}\}$ be the input of the model where T is the maximum length of the tweet, $y \in \mathbb{C}$ the ground-truth of sample X, and V is the vocabulary size. This tweet is passed through a d-dimensional pre-trained embedding layer, E, frozen during the training phase, that is dependent on the language of the corpora used. Moreover, to consider positional information we also experimented with the sine-cosine function proposed in [29], defined in Eq. 1.

$$P_{(pos,2i)} = sin\left(\frac{pos}{1000^{\frac{2i}{d}}}\right) \qquad P_{(pos,2i+1)} = cos\left(\frac{pos}{1000^{\frac{2i}{d}}}\right) \tag{1}$$

where *pos* is the position and i is the dimension. This heuristic exploits the cyclic nature of sine and cosine functions to represent the positional information of the words in a text. Furthermore, unlike learned positional embeddings [51], it is able to generalize to unseen lengths and it dispenses with parameters to learn the positional information, with a negligible computational overhead before training the models. This positional information, encoded as $P \in \mathbb{R}^{T \times d}$, is added to the embedding representation of the tweet, $X^0 \in \mathbb{R}^{T \times d}$, to be used as input to the first encoder layer as shown in Eq 2.

$$X^0 = \{\overbrace{P_1 + E(x_1)}^{X_1^0}, ..., \overbrace{P_T + E(x_T)}^{X_T^0} \ : X_i^0 \in \mathbb{R}^d\} \tag{2}$$

After the combination of the word embeddings with the positional information, dropout [52] is 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 3 to 5.

$$MultiHead(A, B, C) = [head_1; ...; head_h]W^O \tag{3}$$

$$head_i = Attention(AW_i^Q, BW_i^K, CW_i^V) \tag{4}$$

$$Attention(Q, K, V) = softmax(\frac{QK^\intercal}{\sqrt{d_k}})V \tag{5}$$

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 (Q), key (K) and value (V) of the head i and for the output of the multi-head attention respectively; h is the number of heads for the multi-head attention mechanism; and $head_i \in \mathbb{R}^{T \times d_k}$ is the output of the head i. The output for only one encoder, S, is computed as shown in Eq 9 for a given sample X^0.

$$M = MultiHead(X^0, X^0, X^0) \tag{6}$$

$$L = LayerNorm(X^0 + M) \tag{7}$$

$$F = max(0, LW_1 + b_1)W_2 + b_2 \tag{8}$$

$$S = LayerNorm(L + F) \tag{9}$$

where $M, L, F \in \mathbb{R}^{T \times d}$ are the intermediate outputs from the encoder, $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$ are the weights of the position-wise feed forward network, $S \in \mathbb{R}^{T \times d}$ is the output of the encoder, and $LayerNorm$ denotes Layer Normalization [53]. When several encoders are stacked, the input of a encoder is directly used as input to the next encoder. Due to a vector representation is

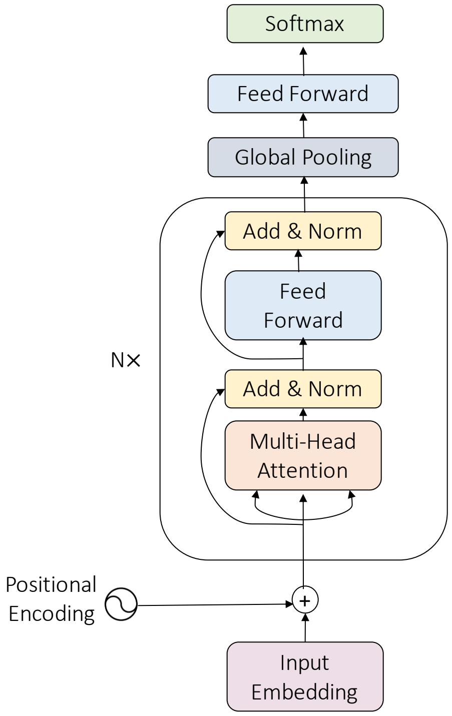

Figure 1: System architecture of our proposal based on Transformer Encoders.

required to train classifiers, on top of the output of the last encoder, a global average pooling was applied on S. The pooled vector, $G \in \mathbb{R}^d$, was used as input for a single-layer feed-forward network, whose output layer computes a probability distribution over the two classes of the task $\mathbb{C} = \{0, 1\}$, as shown in Eq. 10.

$$O = softmax(max(0, GW_3 + b_3)W_4 + b_4) \tag{10}$$

where $O \in \mathbb{R}^{|\mathbb{C}|}$ is a probability distribution over $\mathbb{C}$, $W_3 \in \mathbb{R}^{d \times d_o}$ is the weight matrix of the hidden layer applied on top of G and $W_4 \in \mathbb{R}^{d_o \times |\mathbb{C}|}$ is the weight matrix of the output layer. Due to the imbalance in all the corpora used for the experimentation, weighted cross entropy is used as loss function for training the network, considering the distribution of each class in the training set. This is shown in Eq. 11, where $\mathcal{D}$ is the dataset, $\mathcal{L}$ is the loss function and f is our model parameterized by θ. Concretely, we used the proportion between the most frequent class and the frequency of a given class, $w_j = \frac{\max\limits_{c \in \mathbb{C}} n_c}{n_j}$, where n_j is the number of samples of the class j in a given set, being $w_j = 1$ if j is the most frequent class, and $w_j > w_k$ if class j is less frequent than the class k in the sample set. We used Adam [54] as update rule and Noam [29] as learning rate schedule.

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}[\mathcal{L}(f(x;\theta), y)] = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathbb{C}|} y_{ij}\ log\ f(x_i;\theta)_j\ w_j \tag{11}$$

4. Experimental Work

In order to validate our proposal for irony detection on Twitter, we evaluated it using two different corpora, one for the Spanish language and another for the English language. They have been extensively used with the aim of training and evaluating state-of-the-art systems in both languages for irony detection tasks.

4.1. Tasks

Regarding the Spanish language, we used the corpus provided in the IroSVA shared task [23] for training and evaluating our proposal. The IroSVA shared task, framed in the Iberian Languages Evaluation Forum (IberLEF) and co-located within the SEPLN, aims of determining if a tweet is ironic or not. Three different corpora with tweets from Spain, Mexico, and Cuba were provided by the IroSVA organization. A context of the tweets is also provided, that consists of a short sequence of words that identifies the scope of each tweet, e.g. flat earth or Mexico government. However, this kind of context does not give complementary information about the tweets, beyond identifying topics that are prone to be the object of irony. Also, it is important to note that, due to the fact that the organizers considered a specific context or event to build the corpus, tweets that seem to be not ironic, become ironic considered its context.

The corpus was composed by 2400 training samples and 600 test samples for each Spanish variant. During the training phase, to adjust the models, from the original training set of the competition, we generated new training and development sets following an 87.5%-12.5% proportion for maintaining the relation of 2:1 between the non-ironic and ironic classes such as in the original training set. During the test phase, we used the original test set provided by the competition organizers. The size of each set is shown in Table 1.

Table 1: Corpus statistics for the ironic (I) and the non-ironic (No-I) classes

Corpus	Variant	Training		Development		Test	
		No-I	I	No-I	I	No-I	I
IroSVA	Spain	1400	700	200	100	400	200
	Mexico	1400	700	200	100	401	199
	Cuba	1400	700	200	100	400	200
SemEval	English	1544	1509	372	392	473	311

The official evaluation metrics proposed by the organizers were Precision, Recall, and F_1 in order to assess the performance of the systems. Due to the imbalance between the non-ironic and ironic classes, the macro-averaged F_1 measure was used to rank the participating systems.

For the English language, we used the corpus of the "Irony Detection in English Tweets" shared task [22] proposed in SemEval. The corpus was collected by crawling tweets with hashtags that indicate the presence of irony such as #irony, #sarcasm, and #not during one month. Following this process, a total amount of 4792 tweets were collected (2396 ironic tweets and 2396 non-ironic tweets). Training and test sets, following an 80%-20% proportion, were provided to the participants. It is important to highlight that the test set was modified later by the organizers in order to remove some ironic samples that require context to be understood. From this corpus, two different subtasks were proposed. The first one consists in addressing the irony detection as a binary classification problem. The second one, consists in distinguishing among three different types of irony: verbal irony by means of a polarity contrast, other verbal irony, and situational irony. We only focused on the first subtask, that is the most related with the IroSVA shared task for the Spanish language. However, unlike IroSVA, most of the ironic messages do not require a context to be understood and they are based on conveying opposite meanings. In order to carry out the experimentation, we split the original training partition into training and development partitions, following an 80%-20% proportion. The statistics of each partition are also shown in Table 1.

For evaluation purposes, standard evaluation metrics such as Precision, Recall and F_1 were also used. Concretely, in this case, the organizers consider the F_1 measure of the ironic class in order to rank the participating systems.

4.2. Resources and preprocessing

In order to incorporate task-related knowledge to our model, we initialized the embedding layer with non-contextualized pre-trained word representations. These representations are highly dependent on two main aspects: the domain and the language. Regarding the domain, we only used pre-trained representations of words that appeared on tweets from the social network Twitter. With respect to the language, we used two different word embedding models, one for each language.

For the English language, we used the pre-trained word embeddings provided in [55]. These embeddings were trained by the authors of [55] using 400M English tweets collected from 1/3/2013 to 28/2/2014. Moreover, they determined the best values for some hyper-parameters such as the dimensionality and the topology. The result of their experimentations was a 400-dimensional skip-gram model which we used directly in our proposal. For the Spanish language, we decided to use the same architecture than [55] with a slightly lower dimensionality due to the difference between the number of samples to train the model. In this case, the pre-trained representations were extracted from a 300-dimensional skip-gram model. This model was trained in our laboratory by using 87M Spanish tweets from several Spanish variants. We downloaded these tweets by means of a Twitter streamer, listening for Spanish tweets (including retweets) that contain several common Spanish words such as "que", "de" and “a”. The stream process was performed from 1/6/2017 to 1/7/2017. The competitive behavior obtained by both word embedding models have been proven in several text classification tasks [56][57][58][59].

Regarding the preprocessing, firstly, a case-folding process was applied to all the tweets, secondly, we tokenized the tweets by using the TokTokTokenizer from NLTK [60]. Thirdly, user mentions, hashtags, and URLS were 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*).

4.3. Setup

For both tasks we fixed most of the hyper-parameters, following the experimental setup proposed in [29], with the aim of minimizing the impact of the hyper-parameter tuning when comparing our proposal with other state-of-the-art systems. Specifically, $d_o = 512$, $h = 8$, $d_k = d_o/h = 64$ and $d_{ff} = d$ as stated in [29]. We defined $batch\_size = 32$ and $T = 50$ in order to be slightly higher than the maximum length of the tweets in the training set. Also, as in [29], we used the Adam update rule [54] with $lr = 0.001, \beta_1 = 0.9$ and $\beta_2 = 0.999$, and Noam learning rate schedule with $warmup\_steps = 15$ epochs. We limited the depth of the Transformer Encoder to only one layer due to the limited number of samples on both corpora available to train the models. In the training step, early stopping, with a patience of 20 epochs, was used as stopping criterion.

5. Evaluation

In this section, we present an exhaustive evaluation of the proposed approach. We compare the performance of our system, based on TE, with other deep learning systems, such as Deep Averaging Networks (DAN) [61] and Att-LSTM [62]. This comparison is only performed on the development set, while in the test set, the results of our proposal are compared against the systems of other participants. It is important to note that DAN, TE and Att-LSTM implement a pooling strategy based on averaging, either an unweighted average such as DAN and TE or a weighted average such as Att-LSTM. We used the word embeddings described in Section 4 to train all the models. Another interesting aspect to take into account for irony detection is the positional information. It is intuitive to think that this information is useful for detecting the irony, due to the sequentiality is a relevant factor for some types of irony e.g. irony by means of polarity contrast. For this reason, the effect of the positional information in the results is also studied in our experimentation. Specifically, we tested two different TE models, one with sine-cosine positional information (TE-Pos) and another one without this positional information (TE-NoPos). For all the experimentation, Precision, Recall and F_1 for the two classes, along with their macro-averaged version of F_1 (MF_1) were considered.

The results obtained for IroSVA task on the development set in the three Spanish variants are shown in Table 2. This table only shows the results of the best two systems for the Spain (SP) variant, for the Mexico (MX) and Cuba (CU) variants, due to all the other systems obtain worse results than them in MX and CU.

Table 2: Results on the IroSVA development set for the three Spanish variants.

System	$P(0)$	$P(1)$	$R(0)$	$R(1)$	$F_1(0)$	$F_1(1)$	MF_1
Spain							
DAN	84.13	**72.83**	**87.50**	67.00	85.78	69.79	77.78
Att-LSTM	84.32	61.74	78.00	71.00	81.05	66.05	73.54
TE-NoPos	**88.77**	69.91	83.00	**79.00**	**85.79**	**74.18**	**79.98**
TE-Pos	83.33	62.96	80.00	68.00	81.63	65.38	73.51
Mexico							
DAN	80.11	55.26	74.50	63.00	77.20	58.88	68.04
TE-NoPos	**82.35**	**59.29**	**77.00**	**67.00**	**79.59**	**62.91**	**71.25**
Cuba							
DAN	75.83	55.06	80.00	49.00	77.86	51.85	64.85
TE-NoPos	**82.83**	**64.71**	**82.00**	**66.00**	**82.41**	**65.35**	**73.88**

It can be seen in Table 2 that simpler models (DAN and TE-NoPos), with less parameters and without positional information, obtained the best results for all the evaluation metrics. Concretely, for the variant from Spain, the best results were obtained by TE-NoPos, although DAN outperformed it in terms of $P(1)$ and $R(0)$. For the other two variants (Mexico and Cuba), the TE-NoPos system also achieved the best results outperforming those obtained by DAN model for all the evaluation metrics.

Table 3 shows the results obtained on the development set of the SemEval task for the English language. Note that, all the systems are biased towards the ironic class and all of them obtained better results in terms of the $F_1(1)$ compared to the $F_1(0)$. The Att-LSTM system is the system that shows the most balanced behavior between both measures.

Opposite to IroSVA, Att-LSTM obtained the best results in almost all the metrics, although TE-Pos outperformed it for $P(0)$ and $R(1)$. The differences between both versions of TE are around 5 points for the MF_1 measure in favor of TE-NoPos. DAN and TE-NoPos obtained similar results of MF_1 measure, but they are not the best models. Generally, the conclusion for the English corpus is different to the conclusion for the Spanish one: the most complex model Att-LSTM (with more parameters and considering positional information by its internal memory) shows the best behavior. Nevertheless, due to the fact that TE-NoPos outperforms TE-Pos system in both corpora, it seems that the use of positional information in the Transformer architecture was not useful for the corpora considered. A deeper analysis would be necessary to determine if the negative results, achieved when including positional information, are due to the positional information itself or to the way in which this information is included in the model.

Now, we present results on the test set of both tasks. The results on the

Table 3: Results on the SemEval development set.

	$P(0)$	$P(1)$	$R(0)$	$R(1)$	$F_1(0)$	$F_1(1)$	MF_1
DAN	75.34	62.29	45.16	85.97	56.47	72.24	64.36
Att-LSTM	72.20	**67.17**	**59.41**	78.83	**65.38**	**72.54**	**68.96**
TE-NoPos	72.91	63.16	49.19	82.65	58.75	71.60	65.18
TE-Pos	**76.44**	59.49	35.75	**89.54**	48.72	71.49	60.10

Spanish IroSVA corpus of our TE-NoPos model were published in [31]. Table 4 shows the results, for all Spanish variants, of the best participating teams in the IroSVA competition ranked according the official evaluation measure (MF_1) average for all the Spain variant. Our system outperformed the second-ranked system (CIMAT) by almost to 3 points in average for all the Spanish variants.

Table 4: Results on the IroSVA test set in terms of MF_1. Our system is marked with †.

System	Spain	Mexico	Cuba	AVG
TE-NoPos†	**71.67**	**68.03**	65.27	**68.32**
CIMAT	64.49	67.09	**65.96**	65.85
LDSE	67.95	66.08	63.35	65.79
JZaragoza	66.05	67.03	63.35	64.90
W2V	68.23	62.71	60.33	63.76
ATC	65.12	64.54	59.41	63.02

Regarding the SemEval task, we evaluated two different systems on the test set. The TE-NoPos system proposed in this work and the Att-LSTM system that obtained the best results on the development set. These results have been obtained after finishing the competition. Table 5 shows the results of both systems along with those obtained by the best participating teams in the competition. The systems are ranked according to the official evaluation measure ($F_1(1)$). It is interesting to observe that, although in the validation set, the best system in terms of $F_1(1)$ is Att-LSTM, on the test set, TE-NoPos outperforms it. The best results of $F_1(1)$ obtained by TE-NoPos in comparison to Att-LSTM are due to the increment up to 10 points of $R(1)$ while both systems maintain similar precision on the ironic class $P(1)$. The two systems are biased towards the ironic class, this is the same behavior observed on the development set. Nevertheless, the results obtained by the two systems are very competitive, obtaining the second and third position of the ranking.

The performance of our systems is similar to the best ranked system in terms of $F_1(1)$. However, the difference in terms of Accuracy shows that the THU_NGN system is better detecting the non-ironic class. A deeper study is required to analyze the bias towards the ironic class in our systems compared to the THU_NGN system. Several factors, such as the weighting strategy for the cost-sensitive learning, and the use of a multi-task setup for learning the models, could influence this bias.

Table 5: Results on the SemEval test set ranked in terms of $F_1(1)$. Our systems are marked with †.

System	Acc	$P(1)$	$R(1)$	$F_1(1)$
THU_NGN	**73.50**	63.00	80.10	**70.50**
TE-NoPos †	66.96	54.83	**94.86**	69.49
Att-LSTM †	68.75	57.17	84.56	68.22
NTUA-SLP	73.20	**65.40**	69.10	67.20
WLV	64.30	53.20	83.60	65.00
NLPRL	66.10	55.10	78.80	64.80
NIHRIO	70.20	60.90	69.10	64.80

6. Analysis

In this section, we present several analyses with the aim of explaining how the TE-NoPos system is able to tackle with the irony detection problem. With this study, we pretend to analyze some useful features, captured by our model, for detecting the ironic class e.g. word polarities, relationships among words and relevant individual words. First, we intended to detect which attention heads of our system are more related with the detection of the ironic class. Considering these heads, we studied, for ironic samples, the polarity and relevance of individual words as long as the relationships among words. To carry out these analyses, we used the combination of the training and development sets, with the aim of having a higher number of samples for obtaining more robust conclusions.

6.1. Ablation of the Attention Heads

In order to detect the attention heads that play a highly relevant role in the detection of irony, we performed an ablation process of the attention heads of the trained system TE-NoPos. The main purpose addressed in this section is to answer the following question: are there attention heads specialized on detecting the irony? It is reasonable to think that the competitive results obtained by our system for both languages are due to its ability to capture relevant patterns related with the irony. Therefore, we hypothesize that there are some attention heads that react more to word relationships related to irony.

An iterative ablation process was performed to detect the attention heads whose influence in predicting the ironic class is greater. Concretely, this process consists in iteratively deactivating the output of some attention heads. To do this, the output of each head i we want to ablate is masked, and its output is propagated to the next layers of the network as a zero matrix, $head_i = \mathbf{0}^{T \times d_k}$. Then, we can observe the influence that head i have on the results obtained by the system in terms of the F_1 measure of the ironic class ($F_1(1)$).

During the ablation process, all the $2^h - 2$ combinations of h heads taken from 1 at a time, to $h-1$ at a time, are iteratively evaluated with the aim of observing the worsening of the F_1 for the ironic class. In our study, only the combinations that worsen the previous worst result were taken into account. We hypothesize that the heads that have appeared in more combinations during the successive worsening are those most related with the detection of the irony.

After finishing the iterative process, the heads that most reacted to irony were detected. In Table 6, the number of times that each attention head belongs to a combination that worsens the previous worst results are shown, both for the IroSVA and for the SemEval corpora.

Table 6: Number of times that each attention head appears in a combination that worsens the results, in terms of $F_1(1)$, after a previous worsening of the results. The total number of worsening during the process is also shown together with the number of occurrences of each head.

Corpus	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7
IroSVA	**16/18**	**11/18**	**13/18**	**10/18**	8/18	9/18	4/18	5/18
SemEval	**8/11**	0/11	**10/11**	**6/11**	2/11	3/11	3/11	**8/11**

It is clearly observable that in the English corpus the number of occurrences of the attention heads can be divided in two balanced clusters: those heads that appear in the process more than the half of times and those that appear less than the half. However, in the Spanish corpus there are some attention heads (H_4 and H_5) that are near o exactly on the half number of worsening, i.e. for the Spanish corpus the detection of irony is more scattered among all the attention heads. We considered that, the attention heads that occur more than the half of the times are highly related with the detection of the ironic class. These specialized heads were included in the set H_{ironic}. The remaining heads, less related with the ironic class, were included in the set $H_{non-ironic}$. Thus, in both corpora, there are 4 attention heads that appear more than the half of times ($H_{ironic} = \{H_0, H_1, H_2, H_3\}$ for IroSVA, and $H_{ironic} = \{H_0, H_2, H_3, H_7\}$ for SemEval) and 4 attention heads that appear less than half of the times.

Once the attention heads related to the ironic class detection are identified, it is possible to ablate them in order to observe the results of the system in terms of $F_1(0)$ and $F_1(1)$ without considering them. Table 7 shows the results of the TE-NoPos system when no heads are masked (None column in Table 7), when only H_{ironic} are masked and when only $H_{non-ironic}$ are masked, for both tasks. It can be seen that in both corpora the results in terms of $F_1(1)$ highly decrease when H_{ironic} is masked.

For the English corpus, masking $H_{non-ironic}$ almost does not affect the $F_1(1)$ results, indicating that those attention heads are not highly related with capturing the ironic class. In addition, masking $H_{non-ironic}$ leads also to a high

Table 7: Results on training+development set when masking is not applied, masking H_{ironic} and masking $H_{non-ironic}$.

Corpora	None		H_{ironic}		$H_{non-ironic}$	
	$F_1(0)$	$F_1(1)$	$F_1(0)$	$F_1(1)$	$F_1(0)$	$F_1(1)$
IroSVA	**91.98**	**85.48**	74.11	67.79	85.09	73.03
SemEval	61.88	**71.62**	**69.53**	56.56	41.08	70.17

worsening of $F_1(0)$, suggesting that these heads are related to the detection of the non-ironic class. Therefore, it seems that there are attention heads specialized in detecting the ironic class, those in H_{ironic}, and others specialized in detecting the non-ironic class, those in $H_{non-ironic}$.

For the Spanish corpus, masking H_{ironic} or $H_{non-ironic}$ leads to a high decrease of the performance over all the classes. This suggests that the detection of the ironic and non-ironic classes is highly scattered among all the attention heads, as stated before when we discussed the creation of the H_{ironic} and $H_{non-ironic}$ sets.

The worsening of the results of the ironic class when certain heads are masked seem to support our hypothesis, stated at the beginning of this section, about the specialization of the attention heads.

6.2. Polarity Analysis

In this section, we study if the attention heads in H_{ironic} implicitly capture sentiment information. The aim of this study is to determine if this information is useful for detecting the presence of irony. To achieve this goal, we propose a method to compute, for each head k, the average attention that each word w receives from all the other words w' in its context, averaged for all the occurrences of w in the set of samples $\mathcal{D}$. The context of each word w inside a tweet is determined by all the words of the tweet.

Algorithm 1 Compute the average word attention, for each head, captured by the model on a set of samples.

```
Input: $\mathcal{V}$ vocabulary, set of samples $\mathcal{D}$, trained Transformer Encoder $f$
Result: $\alpha_{wk}$ the average attention of head $k$ for word $w$
procedure ComputeWordAttentions($\mathcal{D}, f$)
    $\alpha_{wk} \leftarrow 0, \ \forall w \in \mathcal{V} \wedge \forall k \in H_{ironic}$
    for $X \in \mathcal{D}$ do
        for $k \in H_{ironic}$ do
            $B \leftarrow softmax(\frac{f(X)_{Q_k} f(X)^{\top}_{K_k}}{\sqrt{d_k}})$
            $B' \leftarrow \frac{1}{|X|} \sum_{i=1}^{|X|} B_{ij}$
            $\alpha_{wk} \leftarrow \alpha_{wk} + B'_w, \ \forall w \in X$
        end for
    end for
    $c_w \leftarrow 0, \ \forall w \in \mathcal{V}$
    $c_w \leftarrow c_w + 1, \ \forall w \in X \wedge \forall X \in \mathcal{D}$
    $\alpha_{wk} \leftarrow \frac{\alpha_{wk}}{c_w}, \ \forall w \in \mathcal{V} \wedge \forall k \in H_{ironic}$
end procedure
```

Algorithm 1 shows our proposal to compute the average attention per word, for each head. From the set of samples $\mathcal{D}$ with vocabulary $\mathcal{V}$ and the trained model f, we compute the average attention given by the head $k \in H_{ironic}$ to each word w in $\mathcal{V}$. To do this, from each sample $X \in \mathcal{D}$ and each head k, the matrix

Table 8: Top-5 attended polarity words by the H_{ironic} heads both for Spanish and English languages.

Language	Heads	(w, α_{wk})
Spanish	H_0	(incompetente, 0.93), (soberbia, 0.90), (desleal, 0.88), (vomitivo, 0.87), (indignacion, 0.86)
	H_1	(laberinto, 0.88), (recomendable, 0.85), (defensor, 0.84), (conspiraciones, 0.81), (maestro, 0.79)
	H_2	(absurda, 0.79), (desinformacion, 0.77), (cobardia, 0.75), (ambicion, 0.67), (mentirosa, 0.64)
	H_3	(salvajismo, 1.0), (corruptos, 0.99), (indecencia, 0.99), (brutal, 0.99), (vomitivo, 0.99)
English	H_0	(persuasive, 1.0), (universal, 0.99), (socialist, 0.99), (supremacy, 0.99), (loon, 0.99)
	H_2	(exhausted, 1.0), (stupidest, 0.99), (president, 0.99), (heck, 0.99), (sensitive, 0.99)
	H_3	(heck, 1.0), (desperately, 1.0), (humid, 1.0), (permission, 1.0), (fault, 1.0)
	H_7	(inspiring, 1.0), (ouch, 1.0), (manic, 1.0), (eventful, 1.0), (sweets, 0.99)

$B \in \mathbb{R}^{|X|\times|X|}$ is computed. The matrix B is the output of the softmax function applied on the scaled dot-product between Q and K matrices, as shown in Eq. 5. The rows of B are averaged to obtain $B' \in \mathbb{R}^{|X|}$. This vector 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 normalized by dividing it by the number of times that the word w appears in all the samples, c_w.

Once the matrix α is computed, it can be determined what are the most attended words by the heads in H_{ironic}. If the attention heads in H_{ironic} focus on more polarity words than the heads of $H_{non-ironic}$, then the polarity words should be more useful for detecting the ironic class than the non-ironic class. Furthermore, the more polarity words focused by H_{ironic}, the more discriminant they should be for detecting the ironic class. To do this study, we determine the most attended words w for each head k by using a threshold ϵ, i.e. a word w is highly attended by an attention head k if $\alpha_{wk} > \epsilon$. We used an $\epsilon = 0.45$ to take into account a considerable number of highly attended words to do the analysis. To determine the polarity of the most attended words, we used some polarity lexicons. For the English language, we used NRC [63], MPQA [64], AFINN [65], and BingLiu [66]. For the Spanish language, we used ElHPolar [67], ISOL [68] and NRC translated to Spanish. Table 8 shows the 5 most attended polarity words for each attention head in H_{ironic}, to illustrate the vocabulary considered in the lexicons and the attention that these words receive. Furthermore, Tables 9 and 10 show the most attended words by each head as well as which of these words are positive or negative for the Spanish and English corpora respectively.

Regarding the Spanish corpus, no heads are reacting in a higher extent to positive words than to negative ones, suggesting that the irony in IroSVA corpus is made by conveying more negative than positive feelings. Furthermore, the heads in H_{ironic} have the highest ratio of polarity words attended, meaning that many of the words highly attended by these heads convey some kind of polarity.

The main differences of the English corpus with respect to the Spanish corpus

Table 9: Number of positive and negative words for each attention head, along with the number of highly attended words and the ratio of polarity words for the Spanish language.

Head Set	Heads	$\lvert\alpha_w > \epsilon\rvert$	Negative	Positive	Ratio
H_{ironic}	H_0	240	102	24	52.50%
	H_1	221	12	18	13.57%
	H_2	73	22	8	41.09%
	H_3	603	140	47	31.01%
	Σ	1137	276	97	32.80%
$H_{non-ironic}$	H_4	276	14	28	15.21%
	H_5	116	6	9	12.60%
	H_6	281	41	11	18.50%
	H_7	237	14	18	13.50%
	Σ	910	75	66	15.50%

Table 10: Number of positive and negative words for each attention head, along with the number of highly attended words and the ratio of polarity words for the English language.

Head Set	Heads	$\lvert\alpha_w > \epsilon\rvert$	Negative	Positive	Ratio
H_{ironic}	H_0	765	92	139	30.20%
	H_2	261	54	45	37.93%
	H_3	544	111	62	31.80%
	H_7	317	29	123	47.94%
	Σ	1887	286	369	34.71%
$H_{non-ironic}$	H_1	159	8	20	17.61%
	H_4	132	14	37	54.54%
	H_5	623	72	149	35.47%
	H_6	817	180	130	37.94%
	Σ	1731	264	336	34.66%

are that in the English corpus a higher attention is given to positive words and a higher ratio of polarity words are attended by all the attention heads, both from H_{ironic} and $H_{non-ironic}$. Moreover, the attention given by the heads from H_{ironic} to polarity words is also more scattered in the English corpus, although, there are some heads mostly specialized on detecting negative (H_3) and positive (H_7) words.

All these results suggest that, in addition to the language, both corpora are quite different because of the type of irony present in them. Perhaps, the fact that the irony of the IroSVA corpus is contextualized (each sample has a certain context) makes its irony different from that of the SemEval corpus.

6.3. The Role of Individual Words for Irony

Similarly to the previous analysis, we are interested in studying if there are specific words with a high impact on the model when it decides if a sample is ironic or not. Thus, the objective is to determine which words, if any, are more relevant in the decision of the model without taking into account their relationships with other words.

This analysis has been addressed from two different perspectives. On the one hand, from the point of view of the attention mechanisms of the TE-NoPos model, by inspecting the attention matrices. To do this, the matrices B for all the attention heads in H_{ironic} are computed and averaged element-wise to obtain a matrix $\hat{B}$, finally the vector B' is computed as the average of the rows of $\hat{B}$. Therefore, in this case, the vector B' contains the averaged attention that each word w receives from all the other words in a tweet, averaged for all the attention heads in H_{ironic}.

On the other hand, from the perspective of the gradients of the loss function $\mathcal{L}$ with respect to the input X, $\nabla_X \mathcal{L}(f(X;\theta), y) \in \mathbb{R}^{T \times d}$. This concept is extensively used in the field of explainable AI [69, 70] and for generating adversarial examples [71]. We have used this information to determine the relevance of the words in the decision of our model when ironic samples are correctly classified as ironic. Thus, from a correctly classified ironic sample $X : y = 1 \wedge f(X) = y$ it is possible to compute $\nabla_X \mathcal{L}(f(X;\theta), y = 1)$ to observe what words of X have gradients with higher Euclidean norm.

Figure 2 shows some examples of ironic tweets. For each example we show, at word level, the euclidean norm of the gradients ($\nabla_X \mathcal{L}(f(.))$) and the averaged attention vector (B').

The Spanish examples translated to English are: “si la tierra fuera plana se habría caído con el lado de la mantequilla hacia abajo” → “if the earth was flat it would has fallen with the butter side down” and “el libro de pedro @user me parece la inocentada de este año en version anticipada [clap emoji]” → “pedro’s @user book seems to me to be an April Fool’s joke of this year in advance [clap emoji]”.

The examples shown in Figure 2 illustrate how the most relevant words are identified in a similar way with both techniques. It is possible to see that, in spite of relationships among words are not considered, the most relevant words

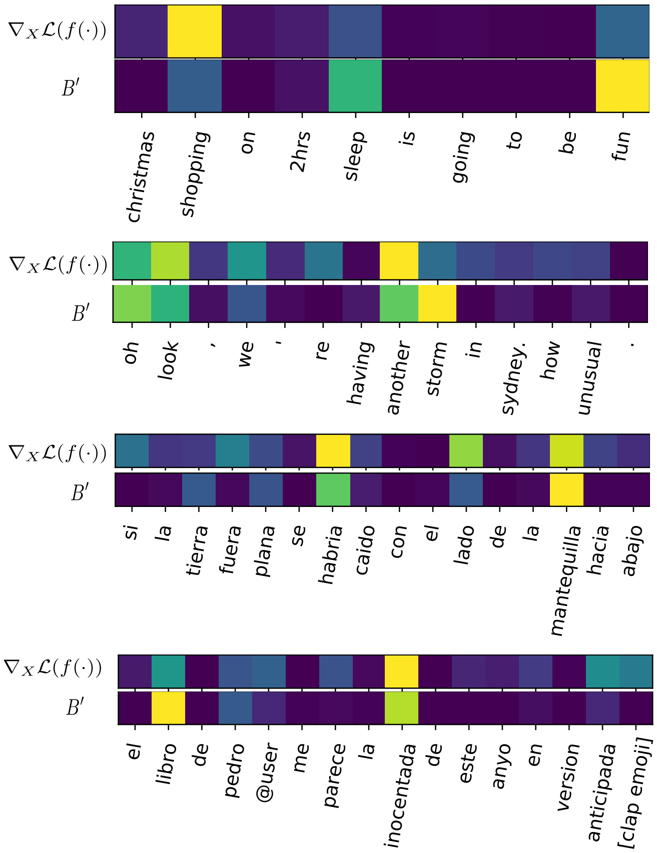

Figure 2: Examples of the word relevance measured by the euclidean norm of the gradients and the average attentions respectively (the lighter the more relevant)

seem to be part of dependencies that involve irony e.g. "shopping — sleep — fun" or "oh — look — another — storm" for the English examples and "tierra (earth) — plana (flat) — lado (side) — mantequilla (butter)" or "libro (book) — inocentada (April Fool's joke)" for the Spanish examples. This last example also illustrates the fact that, mainly in the Spanish corpus, there are some words which bias the decisions of the model to the ironic class. In this case, most of the relevance is assigned to the word "book", hinting the topic about the book of Pedro Sánchez.

6.4. Irony as Word Relationships

In some text classification tasks, such as Sentiment Analysis, some individual words tend to bias the decisions of the models. However, in Irony Detection task, the factor that generally determines the presence of irony is the relationship among words instead of the relevance of individual words.

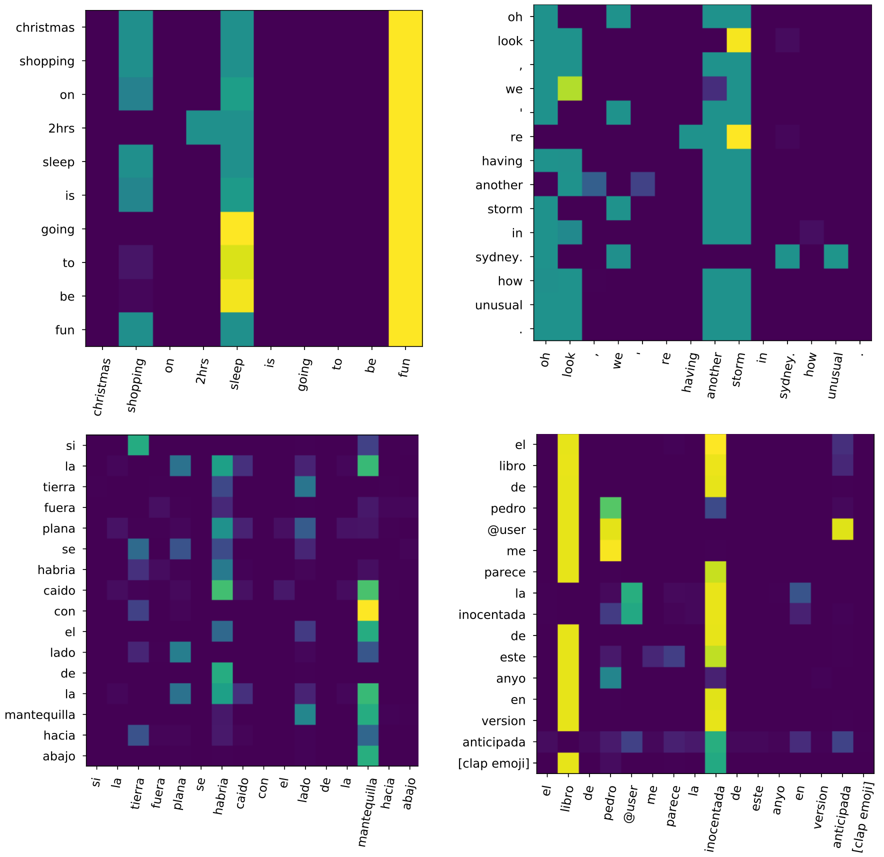

Figure 3: Attention matrices for some ironic examples in both languages (the lighter the more relevant).

In order to analyze these relationships, we computed the average of the attention matrices of all the heads in H_{ironic}, i.e. $A_{ij} = \frac{1}{|H_{ironic}|} \sum_{k \in H_{ironic}} B_{ij}$, to determine the ironic relationships between two words w_i and w_j by observing the attention that the word w_j receives from the word w_i. Thus, the maximum values of the matrix A refer to important ironic relationships between words. Figure 3 shows the matrices A for the examples of the previous section, where the first row refers to the English examples and the second refers to the Spanish examples.

In the first English example, it is remarkable the high attention that the word “fun” receives from all the other words, as well as the relationship among the segment “going to be”, that precedes the word “fun”, and the word “sleep”. Furthermore, it is interesting to observe how the words “christmas”, “shopping”, “sleep”, and “fun” attend the same words (“shopping”, “sleep”, and “fun”) with similar attentions. In the second English example, it can be observed how the model relates “look” and “storm” and how the most relevant words “oh”, “look”, “another” and “storm” are attended highly by all the other words. Regarding the Spanish language, in the first example the attentions are highly scattered, and, the highest attention is given on the word “mantequilla (butter)” by the word “con (with)”. Moreover, it is possible to see how the words of the segment “con el lado (with the side)” place their highest attention in the word “mantequilla (butter)”. Also, the attention that the words “caído (fallen)” and “plana (flat)” put on the word “habría (would have)” are also high. In the second Spanish example, the two words mostly related with the irony, “libro (book)” and “inocentada (April Fool’s joke)”, are the most attended by all the other words, highlighting the relationship among the words of the segment “el libro de (the book of)” with “inocentada (April Fool’s joke)”.

In order to observe in more detail the relationships between words captured by the model, we extracted word pairs with the highest attention from each attention matrix. This pairs excludes those relationships where one of the words is a stopword as well as the relationships where both words are the same. Table 11 shows the top-5 highest attended word pairs for the four previous examples. It can be seen that the captured relationships are highly related with the presence of irony e.g. (shopping, fun), (unusual, oh) or (book, April Fool’s joke).

Table 11: Top-5 relationships between pair of words for the previous ironic examples.

Language	Example	Top-5 Relationships
English	1	(sleep, fun), (christmas, fun) (going, fun), (2hrs, fun), (shopping, fun)
	2	(look, storm), (sydney, unusual), (’, oh), (, , storm), (unusual, oh)
Spanish	1	(fallen, butter), (down, butter), (butter, side), (side, flat), (earth, side)
	2	(book, April Fool’s joke), (pedro, book), ([clap emoji], book), (year, book), (seems, book)

7. Conclusion

In this work, we have presented a Deep Learning proposal for Irony Detection in Twitter both for the English and the Spanish languages. It is based on the contextualization of pre-trained Twitter word embeddings by using a Transformer Encoder architecture.

We have tested our proposal on two different tasks, the IroSVA shared task for the Spanish language and the Irony Detection in English Tweets task of SemEval 2018 for the English language. In both tasks, although we have not performed an extensive exploration of the model hyper-parameters, our system has shown a very competitive behavior. These results have encouraged us to perform a thorough study in order to improve our understanding about how our multi-head self-attention based system addresses the irony detection problem. We have proposed several strategies to identify attention heads specialized on detecting the irony and to analyze several features potentially related with the irony. By means of interpreting the values of the attention heads, we have analyzed the word polarity, the relevance of individual words and the relationships between pairs of words, and how these aspects influence our model in the irony detection task. This analysis is the first step towards a better understanding of the behavior of our model and the features captured by it when addressing the irony detection problem.

8. 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] D. Wilson, D. Sperber, On verbal irony, Lingua 87 (1) (1992) 53 – 76. doi:https://doi.org/10.1016/0024-3841(92)90025-E.

[2] P. Brown, S. C. Levinson, S. C. Levinson, Politeness: Some universals in language usage, Vol. 4, Cambridge university press, 1987.

[3] Y. Hao, T. Veale, Support structures for linguistic creativity: A computational analysis of creative irony in similes, in: Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 31, 2009, pp. 1376–1381.

[4] R. Greene, S. Cushman, C. Cavanagh, J. Ramazani, P. Rouzer, H. Feinsod, D. Marno, A. Slessarev, The Princeton Encyclopedia of Poetry and Poetics, Princeton reference, Princeton University Press, 2012.
URL https://books.google.es/books?id=MJVlZjIe5o8C

[5] H. Colston, R. Gibbs, A brief history of irony, Irony in language and thought: A cognitive science reader (2007) 3–21.

[6] H. P. Grice, P. Cole, J. Morgan, et al., Logic and conversation, 1975 (1975) 41–58.

[7] A. Joshi, P. Bhattacharyya, M. J. Carman, Automatic sarcasm detection: A survey, ACM Comput. Surv. 50 (5) (Sep. 2017). doi:10.1145/3124420.
URL https://doi.org/10.1145/3124420

[8] C. Van Hee, Can machines sense irony? : exploring automatic irony detection on social media, Ph.D. thesis, Ghent University (2017).

[9] C. V. Hee, E. Lefever, V. Hoste, We usually don't like going to the dentist: Using common sense to detect irony on Twitter, Computational Linguistics 44 (4) (2018) 793–832. doi:10.1162/coli_a_00337.

[10] P. Schoentjes, La poetica de la ironia., Cathedra,, 2003.

[11] D. I. H. Farías, V. Patti, P. Rosso, Irony detection in Twitter: The role of affective content, ACM Trans. Internet Technol. 16 (3) (2016) 1–24. `doi:10.1145/2930663`.

[12] S. Poria, E. Cambria, D. Hazarika, P. Vij, A deeper look into sarcastic tweets using deep convolutional neural networks, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1601–1612.
URL `https://www.aclweb.org/anthology/C16-1151`

[13] J. Karoui, F. Benamara Zitoune, V. Moriceau, N. Aussenac-Gilles, L. Hadrich Belguith, Towards a contextual pragmatic model to detect irony in tweets, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 644–650. `doi:10.3115/v1/P15-2106`.
URL `https://www.aclweb.org/anthology/P15-2106`

[14] B. C. Wallace, D. K. Choe, E. Charniak, Sparse, contextually informed models for irony detection: Exploiting user communities, entities and senti
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), Association for Computational Linguistics, Beijing, China, 2015, pp. 1035–1044. `doi:10.3115/v1/P15-1100`.
URL `https://www.aclweb.org/anthology/P15-1100`

[15] A. Joshi, V. Sharma, P. Bhattacharyya, Harnessing context incongruity for sarcasm detection, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 757–762. `doi:10.3115/v1/P15-2124`.
URL `https://www.aclweb.org/anthology/P15-2124`

[16] Z. G. Szabó, Compositionality, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, summer 2017 Edition, Metaphysics Research Lab, Stanford University, 2017.

[17] D. Muecke, Irony markers, Poetics 7 (4) (1978) 363 – 375. `doi:https://doi.org/10.1016/0304-422X(78)90011-6`.

[18] F. Poyatos, La comunicación no verbal, Vol. 13, Ediciones AKAL, 1994.

[19] S. Rosenthal, A. Ritter, P. Nakov, V. Stoyanov, SemEval-2014 task 9: Sentiment analysis in Twitter, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 73–80. doi:10.3115/v1/S14-2009.

[20] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, A. Reyes, SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 470–478. doi:10.18653/v1/S15-2080.

[21] P. Rosso, F. Rangel, I. H. Farías, L. Cagnina, W. Zaghouani, A. Charfi, A survey on author profiling, deception, and irony detection for the Arabic language, Language and Linguistics Compass 12 (4) (4 2018). doi:10.1111/lnc3.12275.

[22] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 task 3: Irony detection in English tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 39–50. doi:10.18653/v1/S18-1005.

[23] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y Gómez, J. E. Medina Pagola, Overview of the task on irony detection in Spanish variants, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. org, 2019, pp. 229–256.

[24] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, et al., Overview of the Evalita 2018 task on irony detection in Italian tweets (Ironita), in: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), Vol. 2263, CEUR-WS, 2018, pp. 1–6.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, 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, Curran Associates Inc., USA, 2013, pp. 3111–3119.

[26] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. doi:10.3115/v1/D14-1181.

[27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., USA, 2017, pp. 6000–6010.
URL http://dl.acm.org/citation.cfm?id=3295222.3295349

[30] E. Refaee, V. Rieser, An Arabic Twitter corpus for subjectivity and sentiment analysis, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 2268–2273.
URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/317_Paper.pdf

[31] J. González, L. Hurtado, F. Pla, ELiRF-UPV at IroSvA: Transformer encoders for Spanish irony detectio in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019., 2019, pp. 278–284.
URL http://ceur-ws.org/Vol-2421/IroSvA_paper_4.pdf

[32] H. U. Miranda-Belmonte, A. P. López-Monroy, Early fusion of traditional and deep features for irony detection in Twitter, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019., 2019, pp. 272–277.
URL http://ceur-ws.org/Vol-2421/IroSvA_paper_3.pdf

[33] A. T. Cignarella, C. Bosco, ATC at irosva 2019: Shallow syntactic dependency-based features for irony de in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019., 2019, pp. 257–263.
URL http://ceur-ws.org/Vol-2421/IroSvA_paper_1.pdf

[34] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, Y. Huang, THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 51–56. doi:10.18653/v1/S18-1006.

[35] C. Baziotis, A. Nikolaos, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, A. Potamianos,

NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and in: Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, 2018, pp. 245–255.
URL https://www.aclweb.org/anthology/S18-1037/

[36] O. Rohanian, S. Taslimipoor, R. Evans, R. Mitkov, WLV at SemEval-2018 task 3: Dissecting tweets in search of irony, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 553–559. doi:10.18653/v1/S18-1090.

[37] H. Rangwani, D. Kulshreshtha, A. Kumar Singh, NLPRL-IITBHU at SemEval-2018 task 3: Combining linguistic features and emoji pre-trained CNN for irony detection in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 638–642. doi:10.18653/v1/S18-1104.

[38] T. Vu, D. Q. Nguyen, X.-S. Vu, D. Q. Nguyen, M. Catt, M. Trenell, NIHRIO at SemEval-2018 task 3: A simple and accurate neural network model for irony detection in Twitter, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 525–530. doi:10.18653/v1/S18-1085.

[39] J.-Á. González, L.-F. Hurtado, F. Pla, ELiRF-UPV at SemEval-2018 tasks 1 and 3: Affect and irony detection in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 565–569. doi:10.18653/v1/S18-1092.

[40] F. Yus, Propositional attitude, affective attitude and irony comprehension, Pragmatics & Cognition 23 (1) (2016) 92–116.

[41] E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems 108 (2016) 132 – 143, new Avenues in Knowledge Bases for Natural Language Processing. doi:https://doi.org/10.1016/j.knosys.2016.05.035.

[42] S. Zhang, X. Zhang, J. Chan, P. Rosso, Irony detection via sentiment-based transfer learning, Information Processing & Management 56 (5) (2019) 1633 – 1644. doi:https://doi.org/10.1016/j.ipm.2019.04.006.

[43] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as contrast between a positive sentiment and negative situation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle,

Washington, USA, 2013, pp. 704–714.
URL https://www.aclweb.org/anthology/D13-1066

[44] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations (2019). arXiv:1909.11942.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.
URL http://arxiv.org/abs/1907.11692

[46] R. A. Potamias, G. Siolas, A. Stafylopatis, A transformer-based approach to irony and sarcasm detection, ArXiv abs/1911.10401 (2019).

[47] J. Iranzo-Sánchez, R. Ruiz-Dolz, VRAIN at irosva 2019: Exploring classical and transfer learning approac
in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, 2019, pp. 322–328.
URL http://ceur-ws.org/Vol-2421/IroSvA_paper_10.pdf

[48] C. Zhang, M. Abdul-Mageed, Multi-task bidirectional transformer representations for irony detection, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, 2019, pp. 391–400.
URL http://ceur-ws.org/Vol-2517/T4-2.pdf

[49] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, S. Poria, Towards multimodal sarcasm detection (an _obviously_ perfect paper), in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, 2019, pp. 4619–4629. doi:10.18653/v1/p19-1455.
URL https://doi.org/10.18653/v1/p19-1455

[50] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation., in: EMNLP, Vol. 14, 2014, pp. 1532–1543.

[51] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 1243–1252.
URL http://proceedings.mlr.press/v70/gehring17a.html

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
URL http://dl.acm.org/citation.cfm?id=2627435.2670313

[53] L. J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, CoRR abs/1607.06450 (2016). arXiv:1607.06450.
URL http://arxiv.org/abs/1607.06450

[54] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, pp. 1–15.

[55] F. Godin, Improving and interpreting neural networks for word-level prediction tasks in natural language processing, Ph.D. thesis, Ghent University, Belgium (2019).

[56] J. González, F. Pla, L. Hurtado, Elirf-upv en TASS 2018: Análisis de sentimientos en twitter basado en ap
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, 2018, pp. 37–44.
URL http://ceur-ws.org/Vol-2172/p2_elirf_tass2018.pdf

[57] J.-Á. González, F. Pla, L.-F. Hurtado, ELiRF-UPV at SemEval-2017 task 4: Sentiment analysis using deep learning, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 723–727. doi:10.18653/v1/S17-2121.
URL https://www.aclweb.org/anthology/S17-2121

[58] J. González, L. Hurtado, F. Pla, Elirf-upv at semeval-2019 task 3: Snapshot ensemble of hierarchical conv
in: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, 2019, pp. 195–199. doi:10.18653/v1/s19-2031.
URL https://doi.org/10.18653/v1/s19-2031

[59] J. González, F. Pla, L. Hurtado, Elirf-upv en TASS 2018: Categorización emocional de noticias(elirf-upv a
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, 2018, pp. 103–109.
URL http://ceur-ws.org/Vol-2172/p12_elirf_upv_tass2018.pdf

[60] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, 1st Edition, O'Reilly Media, Inc., 2009.

[61] M. Iyyer, V. Manjunatha, J. Boyd-Graber, 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), Association for Computational Linguistics, Beijing, China, 2015, pp. 1681–1691. doi:10.3115/v1/P15-1162.

[62] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 207–212. doi:10.18653/v1/P16-2034.

[63] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (3) (2013) 436–465. doi:10.1111/j.1467-8640.2012.00460.x.

[64] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 347–354. doi:10.3115/1220575.1220619.

[65] F. A. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs., in: M. Rowe, M. Stankovic, A.-S. Dadzie, M. Hardey (Eds.), #MSM, Vol. 718 of CEUR Workshop Proceedings, CEUR-WS.org, 2011, pp. 93–98. URL http://dblp.uni-trier.de/db/conf/msm/msm2011.html#Nielsen11

[66] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[67] X. Saralegi, I. San Vicente, Elhuyar at TASS 2013, in: XXIX Congreso de la Sociedad Espaola de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013), 2013, pp. 143–150.

[68] E. Martínez-Cámara, M. T. Martín-Valdivia, M. D. Molina-González, 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, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 87–93. URL https://www.aclweb.org/anthology/W13-1612

[69] A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 2662–2670. doi:10.24963/ijcai.2017/371.

[70] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, CoRR abs/1706.03825 (2017). arXiv:1706.03825. URL http://arxiv.org/abs/1706.03825

[71] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, pp. 1–11.