Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

TWilBert: Pre-trained deep bidirectional transformers for Spanish 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

ARTICLE INFO

Article history:
Received 16 April 2020
Revised 29 July 2020
Accepted 17 September 2020
Available online xxxx
Communicated by Zhaopeng Tu

Keywords:
Contextualized Embeddings
Spanish
Twitter
TWilBERT

ABSTRACT

In recent years, the Natural Language Processing community have been moving from uncontextualized word embeddings towards contextualized word embeddings. Among these contextualized architectures, BERT stands out due to its capacity to compute bidirectional contextualized word representations. However, its competitive performance in English downstream tasks is not obtained by its multilingual version when it is applied to other languages and domains. This is especially true in the case of the Spanish language used in Twitter.

In this work, we propose TWiLBERT, a specialization of BERT architecture both for the Spanish language and the Twitter domain. Furthermore, we propose a Reply Order Prediction signal to learn inter-sentence coherence in Twitter conversations, which improves the performance of TWilBERT in text classification tasks that require reasoning on sequences of tweets. We perform an extensive evaluation of TWilBERT models on 14 different text classification tasks, such as irony detection, sentiment analysis, or emotion detection. The results obtained by TWilBERT outperform the state-of-the-art systems and Multilingual BERT. In addition, we carry out a thorough analysis of the TWilBERT models to study the reasons of their competitive behavior. We release the pre-trained TWilBERT models used in this paper, along with a framework for training, evaluating, and fine-tuning TWilBERT models.

1. Introduction

In recent years, the Natural Language Processing community have been moving from uncontextualized word representations [1,2] towards contextualized word representations [3–6]. In the first case, each word is represented by one embedding that condenses information of all the contexts where the word appears. While in the second case, each word is represented by different embeddings depending on the context of the word. This allows to model complex features of the words e.g. coreference or polysemy. Among these contextualized architectures, BERT [6] stands out due to its capacity to compute bidirectional contextualized word representations.

BERT is a neural bidirectional language model which uses Transformer Encoders [7] as backbone. It is able to compute bidirectional word representations due to the use of a Masked Language Model (MLM) as pre-training objective. MLM is based on Cloze tasks, where tokens are randomly masked, forcing the model to learn the bidirectional context of a token to predict it. Furthermore, the authors of BERT considered the sentence coherence as an important aspect of language understanding. For this reason, they proposed the Next Sentence Prediction (NSP) signal with the aim of learning coherence by means of determining if a text segment A precedes a text segment B in the original source.

Due to the competitive performance of this model in English downstream tasks, the authors of BERT provided a multilingual version (M-BERT), trained with the Wikipedia dumps of 104 different languages. However, this multilingual model exhibits systematic deficiencies that affect certain language pairs [8]. Furthermore, the competitive performance of BERT in English downstream tasks is not achieved by M-BERT when it is applied to tasks on other languages. or this reason, specializations of BERT for several languages have proliferated [9–13].

In addition to the language, the domain of the downstream tasks is also a key aspect which degrades the performance of this kind of models. The more different the target domain is compared to the source domain, the more remarkable is the degradation. This is especially true for the Twitter domain in which we are interested, where, usually, users communicate with each other informally, and using social networks slang.

* Corresponding author.
E-mail addresses: jogonba2@dsic.upv.es (J.Ángel González), lhurtado@dsic.upv.es (L.-F. Hurtado), fpla@dsic.upv.es (F. Pla).

https://doi.org/10.1016/j.neucom.2020.09.078
0925-2312/

In this work, we propose a specialization of BERT both for the Spanish language and the Twitter domain, that we called as TWilBERT. This specialization consists on training a BERT model from scratch to obtain coherent contextualized embeddings of Spanish tweets. In order to learn inter-sentence coherence, we propose Reply Order Prediction (ROP), an adaptation of the NSP signal, similar to [14], to Twitter conversations. To our knowledge, this is the first work that proposes a full specialization of BERT for the Twitter domain, taking coherence into account. In addition, we implemented and freely released a Keras [15] framework to train, evaluate and fine-tune TWilBERT models. Moreover, we release the pre-trained TWilBERT models used in this article, that can be easily used in the provided framework.

The following items summarize the main contributions of our work:

- We propose an adaptation of BERT to address text classification tasks in Spanish Twitter, that obtains significant improvements on several datasets provided in international workshops.
- We adapted the Next Sentence Prediction signal for learning coherence between pairs of tweets inside Twitter conversations, which significantly improves the results on downstream tasks.
- We performed an extensive analysis to study the competitive performance of TWilBERT in Spanish Twitter text classification tasks.
- We provided a framework [1] for training, evaluating, and fine-tuning TWilBERT models, that also allows to use several improvements on Transformer models recently published in the literature.
- We released the pre-trained TWilBERT models used in this paper, that are freely available.[1]

The rest of this paper is structured as follows. In Section 2, several related works for BERT models are presented. In Section 3, we discuss the motivation for specializing BERT models on the Twitter domain and in the Spanish language. In Section 4 a description of the TWilBERT models used in this article is presented. In Section 5 we present the evaluation of the TWilBERT models on a broad set of downstream text classification tasks on Spanish Twitter. In Section 6 several analyses to study the competitive behavior of the TWilBERT models are performed. Finally, in Section 7, we present the conclusions extracted after evaluating and analyzing our proposal, along with the future work.

2. Related work

BERT and several variants of its underlying structure are the state of the art for learning contextual representations that are useful in many Natural Language Processing tasks. In this section, we discuss some of these variants of the BERT architecture and its hiper-parameters, recently published in the literature, that improved BERT in several directions [14,16,17].

In SpanBERT [17], several masking strategies were proposed. Their best results were obtained by masking contiguous random token spans instead of single tokens, and using a span boundary objective for predicting each token in a masked span using the tokens on its boundary. The performance of several pre-training masking schemes in span selection tasks such as question answering and coreference resolution, was also studied. They found that using a geometric distribution for sampling random spans provides substantial gains on span selection tasks.

The authors of RoBERTa [16] made a careful measurement of the impact of BERT hyper-parameters and training corpora on the performance of the model. Specifically, they found three interesting aspects that had a great impact on the BERT performance: the NSP signal, the masking strategy, and the batch size. First, the NSP signal consistently degrades the results on downstream tasks, showing that this signal does not provide additional information to the MLM. Regarding the masking, they found that a dynamic masking strategy achieved better results than static masking, i.e. it is better to use different maskings rather than use a small fixed set of masks for each sample during training as in BERT. With respect to the batch size, they found that using large batch sizes improves the perplexity of the MLM objective, as well as the performance on downstream tasks.

In AlBERT [14], the authors found that there is some point where further increasing the model size degrades the behavior of the system in downstream tasks. This degradation was observed empirically when a BERT model with $L = 24$ layers and hidden size $H = 2048$ was trained and fine-tuned for the ReAding Comprehension from Examinations dataset [18], obtaining significantly lower results than another model trained with $L = 24$ and $H = 1024$ (BERT large [6]). To overcome this degradation when the model size increases, while maintaining the training time and the memory consumption, the authors of AlBERT proposed three different strategies. Firstly, the factorized embedding parameterization. This strategy was proposed because of, usually, in this kind of models it is required a higher dimensionality for the contextualized representations than for the subword embeddings. In BERT, increasing the dimensionality of the contextual embeddings forces to increase also the dimensionality of the incontextual subword embeddings due to the residual connections between each pair of subsequent layers. Nevertheless, factorized embedding parameterization allows to untie the dimensionality of both kinds of embedding, reducing considerably the number of parameters of the model. Secondly, cross-layer parameter sharing, to improve the parameter efficiency by means of tying the weights among a pre-defined set of layers. The authors of the research shown that this strategy was able to smooth the transitions from layer to layer, thus stabilizing the network parameters. Thirdly, they proposed an alternative to the NSP signal, the so-called Sentence-Order Prediction (SOP). The SOP signal is a reformulation of NSP where pairs of unordered sentences are used as negative samples. The benefits from the NSP signal have been a controversial issue in the literature [19,17], it seems that the NSP signal captures only topic coherence which does not provide additional information to the MLM task. For this reason, the authors of [14] proposed SOP as pre-training signal to learn better the inter-sentence coherence.

Recently, several strategies for improving Transformer models have been proposed. These strategies can be used to increase the performance of BERT by means of modifying its underlying architecture. Some works in this regard are the LAMB optimizer [20] and Product Key Memory layers [21]. In [20] the authors proposed a layerwise adaptative large batch optimization technique which allows the models to be trained with very large mini-batches without any degradation of the performance. This way, the training time of the BERT models was reduced from 3 days to 76 min in a TPUv3 Pod. In [21], the authors proposed a novel structured memory layer which can be integrated in any neural network with the aim of increasing the capacity of the models without computational overhead. This mechanism has been especially useful in Transformer language models, where a 12-layered Transformer with only one memory layer, under a specific setup of its hyper-parameters, was able to outperform a 24-layered baseline Transformer.

In order to use BERT in other languages different from the English language, the authors of BERT [6] also provided a multilingual pre-trained model (M-BERT). This model was trained with the Wikipedia datasets of 104 different languages. However, the

[1] https://github.com/jogonba2/TWilBERT.

competitive performance of BERT in English downstream tasks is not achieved by M-BERT when it is used on other languages. Several works have focused on training specialized BERT models, from scratch or from pre-trained weights, for several languages: Dutch [9], French [10,11], Finnish [12], and Italian [13]. In FlauBERT [10] and in CamemBERT [11], pre-trained BERT-based language models were proposed for the French language, which obtained better results than M-BERT, under similar settings, for a wide range of downstream tasks. In [12], a thorough evaluation of M-BERT compared with a BERT model trained from scratch with Finnish texts, was made. The authors shown that the language-specialized version constitutes the state of the art in several Finnish tasks, systematically outperforming M-BERT, which largely fails to reach competitive performance. In AlBERTo [14], a BERT language model was pre-trained with Italian tweets (without coherence signal) and evaluated in several text classification tasks such as irony detection, sentiment analysis, and subjectivity classification.

In addition to the language, another aspect that degrades the performance of pre-trained BERT models is the domain. The more different the target domain is from the pre-training domain, the more remarkable is the degradation of the performance. Several works studied this issue, mainly focusing on training BERT models, from-scratch or on from some pre-trained weights, in the target domain [13,22,23]. In [22], a BERT-based language model was pre-trained by using a large-scale dataset of scientific papers. Their experimentation with the domain specialized model shown significant improvements over BERT in a broad set of tasks. In [23], the authors proposed to use combinations of general and biomedical domain corpora in order to train BERT-based language models specialized on addressing named entity recognition, relation extraction, and question answering downstream tasks.

In this work, we propose the specialization of BERT for both the Spanish language and the Twitter domain. We called this approach TWilBERT. TWilBERT leverages recent modifications of the BERT architecture, published in RoBERTa [16] and AlBERT [14], that shown systematic improvements on the MLM objective and downstream tasks. Specifically, our proposal aggregates the inter-sentence coherence loss of AlBERT, applied on (tweet, reply) pairs, along with most of the hyper-parameter choices of RoBERTa that allow for successfully pre-training BERT models such as: dynamic masking, which is crucial for pre-training on large datasets; the use of large batch sizes for improve the perplexity of the MLM objective and the performance in downstream tasks, and the value of the Adam β_2 hyper-parameter for improving stability with large batch sizes.

Beyond the similarities of TWilBERT, AlBERT and RoBERTa in terms of the underlying architecture and its hyper-parameters, the most related work presented in this section is ALBERTo [13], because of we also attempted to address the Twitter domain. In addition to the specialization language (Spanish language in our case), our systems are different from ALBERTo models in a crucial aspect of the BERT architecture. In ALBERTo, the model does not learn coherence among tweets because the cognition of a flow of tweets cannot be automatically identified on a sequence of tweets from the same author. However, we considered that inter-sentence coherence is an important aspect of language understanding that could improve the performance on downstream tasks that require reasoning on pairs of tweets. For this reason, differently from [13], we propose to use coherence signals in Twitter conversations, where a flow of tweets can be easily identified as (tweet, reply) pairs.

In addition, we implemented and freely released a Keras [15] framework to train, evaluate and fine-tune TWilBERT models. All the techniques and improvements discussed in this section are implemented within the framework. Also, we release the pre-trained TWilBERT models used in this article, that can be easily used in the provided framework.

3. Motivation

The competitive behavior of BERT-based models in downstream tasks whose features are similar to the dataset used for pre-training have encouraged the scientific community to use BERT ubiquitously in a broad range of tasks. However, its performance is drastically reduced when this kind of models, specially M-BERT, are used in non-English tasks [11,10,13] where some properties like syntaxis and grammar are different from those which the models were trained. This is the case of the Twitter domain, where M-BERT have to deal with Spanish tweets [24–27]. Typically, these proposals have obtained lower results than other Deep Learning architectures based on the use of incontextual word embeddings trained with Spanish Twitter datasets [28,29].

Our motivation in this research is to adapt and improve the language modeling capacity of the BERT architecture to boost the state of the art in text classification tasks in the domain of Twitter for the Spanish language. To achieve this goal, it is necessary to tackle with two main challenges.

The first challenge is the language dependency. Although the authors of [6] provided multilingual models pre-trained with large amounts of texts in many languages (M-BERT), which presupposes that all these languages share structural properties e.g. typological (similar subwords) or grammatical properties. However, despite the fact that M-BERT provides a deeper representation than simply memorizing vocabulary, contextual representations exhibit systematic deficiencies that affect certain language pairs, as shown in [8]. This entails to a reduction in the results when fine-tuning is performed on some languages. This is so much so that, in order to obtain more competitive performance, usually, it is better to use simpler models trained in the target language than the M-BERT model.

The second challenge we must tackle is the domain dependency. M-BERT was trained using the Wikipedia dataset from 104 different languages. Consequently, the use of M-BERT in other domains can degrades the performance if the target domain is very different to the domain used for pre-training. This is the case of Twitter, where users communicate with each other informally, using typical expressions of social networks slang, many times with lexico-syntactic errors, or adding special tokens such as hashtags, user mentions, and emojis. Therefore, there is a great mismatch between Twitter (target domain) and Wikipedia (source domain).

Another problem related with the domain is the strategy used to learn coherence. We consider that, as discussed in [14], the inter-sentence modeling is an important aspect of language understanding, and we want to take it into account for learning coherence in Twitter. To learn coherence in M-BERT, the self-supervised Next Sentence Prediction (NSP) was used, that allows to improve the performance in downstream tasks which require reasoning between pairs of sentences. However, the benefits of the NSP signal have been a controversial topic in the literature [19,17]. To address the NSP problems, the SOP signal was proposed in [14]. In addition, in the Twitter domain, this signal cannot be used directly, due to there is no sequentiality between sentences like in a document (or tweets in the history of tweets from a given user). Nevertheless, there is a sequentiality among a given tweet and a reply to this tweet in Twitter conversations. For these reasons, in this work, we propose the Reply Order Prediction (ROP) signal, which is an application of SOP to learn coherence between (tweet, reply) pairs in order to improve the performance in downstream tasks that requires reasoning on pairs of tweets. The

definition of this signal is identical to SOP, but using positive and negative pairs extracted from Twitter conversations instead of subsequent sentences of a document.

4. TWilBERT

TWilBERT is provided as a framework [1] that allows training, evaluating, and fine-tuning BERT-based models. It also includes several techniques and improvements published in recent works such as: cross-sharing parameter layers [14], factorized embedding parameterization [14], Product Key Memory layers [21], LAMB optimizer [20], gradient accumulation. In addition, we provided two different pre-trained models for the Twitter domain in Spanish, freely available in.[1]

Similarly to what the authors made in [6], we defined two different TWilBERT models, with different number of Transformer layers and attention heads in the multi-head self-attention mechanism of each layer.

On the one hand, TWilBERT-Base (TW-Base) was defined to have half of the Transformer layers and attention heads than M-BERT. TW-Base has $L = 6$ Transformer layers, $A = 6$ attention heads, $d_q = d_k = d_v = 64$ the dimensionality of the Query, Key and Value projections [7], $E = 768$ the dimensionality for the subword embedding layer and $H = E$ hidden size.

On the other hand, TWilBERT-Large (TW-Large) was defined to have the same number of parameters than M-BERT [6]. Specifically, TW-Large have $L = 12,\ A = 12,\ d_q = d_k = d_v = 64,$ and $E = H = 768$. We did not use any kind of dropout [30] in the models, due to it can adversely affect the performance of Transformer-based models, as stated in [14]. As pre-training objectives, we used MLM and ROP for both TWilBERT models, in order to learn coherent bidirectional representations of (tweet, reply) pairs.

We used dynamic masking [16] for generating the MLM targets using n-gram masking [17] with a maximum span of $m = 3$ subwords and a maximum of 15% subwords masked for each sample. The probability for masking a span of length $0 < l \leqslant m$ is defined following Eq. (1). The probabilities of each kind of token masking ([MASK] token, random subword and keep subword) are the same as in the BERT model [6].

$$p(l) = \frac{1/l}{\sum_{i=0}^{m} 1/i} \tag{1}$$

To build the corpus, a total of 91 million of Spanish tweets were streamed from September 2019 to January 2020. We applied a post-process in order to get the replies for all the tweets collected by the streamer. Those tweets that have not got reply, or are not reply of a tweet, or have less than 3 words, were discarded. The result of this post-process was 47 million of (tweet, reply) pairs (7.65 Gb of text and 1.16 billion words) which generates 94 million of positive and negative pairs for the ROP signal. All these tweets were segmented as subwords units by using SentencePiece [31] with a vocabulary size of 30,000 subwords. Furthermore, in order to reduce the number of subwords required for representing the (tweet, reply) pairs, user mentions and urls were replaced by a generic token.

We used Adam [32] with gradient accumulation for minimizing the cross-entropy both for the MLM and ROP signals, with an effective batch size of 2048 samples (64 batch size and 32 accumulation iterations). We also used Noam learning rate annealing [7] with 10,000 warmup steps for TW-Large and 8,000 for TW-Base, due to the faster convergence of TW-Base compared to TW-Large. To deal efficiently with pairs of variable-length sequences, we implemented a bucketing strategy based on the lengths of the (tweet, reply) pairs, with a maximum length of 128, in order to reduce, as much as possible, the amount of padding. The buckets were ordered by length in ascending order to be used as input for the TWilBERT models. Two Nvidia Geforce RTX 2080 Ti were used for training TW-Large and TW-Base during 15 days.

Table 1
Differences among M-BERT, TW-Base, and TW-Large.

	M-BERT	TW-Base	TW-Large
Language	104 languages	Spanish	Spanish
Domain	Wikipedia	Twitter	Twitter
Objectives	MLM+NSP	MLM+ROP	MLM+ROP
Tokenization	WordPiece	SentencePiece	SentencePiece
Vocabulary	110k	30k	30k
Masking	Static subword	Dynamic spans	Dynamic spans
L	12	6	12
A	12	6	12
E	768	768	768
H	768	768	768
d_q	64	64	64
d_k	64	64	64
d_v	64	64	64

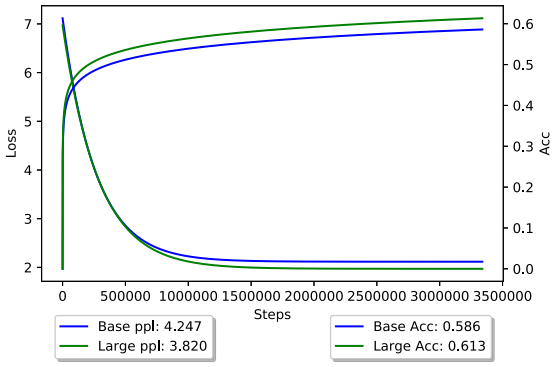

Fig. 1. Loss, including perplexity (ppl) and accuracy (Acc) of the MLM signal.

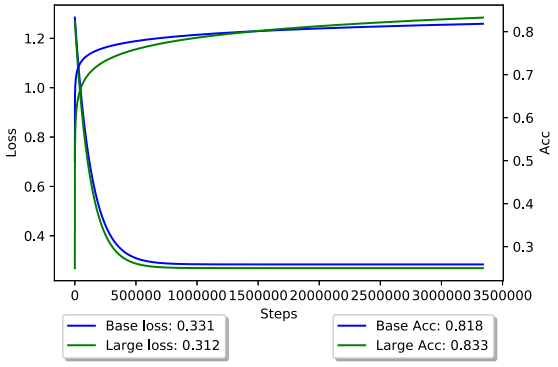

Fig. 2. Loss and accuracy of the ROP signal.

Table 1 summarizes the differences among both TWilBERT models and M-BERT. Figs. 1 and 2 show the results, at each training step, for the MLM and ROP signals respectively at each training step for both TWilBERT models. It can be seen that TW-Large outperforms TW-Base for both the MLM and ROP tasks.

5. Experimental work

In order to evaluate the performance of our proposal, we selected a broad set of text classification tasks for Spanish language

on the Twitter domain. Specifically, we are interested in addressing tasks related with social media analysis such as sentiment analysis, emotion detection, stance detection, hate speech detection and topic detection. Furthermore, to make a fair comparison with the state of the art, we only considered reference corpora provided in three international competitions that are highly relevant in the field: Evaluation of Human Language Technologies for Iberian languages (IberEval), Iberian Languages Evaluation Forum (IberLEF), and International Workshop on Semantic Evaluation (SemEval). Additionally, several tasks that we address in our experimentation provide the corpora in different Spanish variants from Spain, Mexico, Uruguay, etc. We addressed also these cases in order to observe the behavior of our models in specific low-resources variants of the Spanish language.

According to the requirements of BERT-based models to operate on a given input and its respective output, all the text classification tasks considered in this paper can be divided in the following categories:

- **Single-input single-label**: given as input a sample $X \in \mathcal{V}^T$, a degenerate $(X, \varnothing)$ pair is generated. The pooled token representations are used as input for a softmax output layer that computes a probability distribution over the set of classes $\mathcal{C}$.
- **Single-input multi-label**: in this case, the input is identical to the previous one, however, the last output layer is a sigmoid layer that computes the probability of each class $c \in \mathcal{C}$ as a Bernoulli distribution.
- **Multi-input single-label**: the input is composed by k different text segments. To handle it, all the text segments are concatenated by means of the [SEP] token in order to compose the input for TWilBERT. The output layer is a softmax layer to compute a probability distribution over the set of classes $\mathcal{C}$.

The evaluation metrics used in this experimentation are those considered in the competitions to rank the systems. Specifically, for the single-label tasks, the metrics considered were: Accuracy (Acc), Macro-Precision (MP, Eq. (2), which is defined in terms of Eq. (3)), Macro-Recall (MR, Eq. (4), which is defined in terms of Eq. (5)), Macro-F_1 (MF_1, Eq. (6)), and Binary F_1 (Eq. (7) when $c = 1$). For the multi-label case, Jaccard Accuracy (JAcc, Eq. (8)) is considered.

$$\mathrm{MP} = \frac{1}{|\mathcal{C}|}\sum_{c=0}^{|\mathcal{C}|}\mathrm{P}_c \tag{2}$$

$$\mathrm{P}_c = \frac{\sum_{i=1}^{N}[y_i = \hat{y}_i = c]}{\sum_{i=1}^{N}[\hat{y}_i = c]} \tag{3}$$

$$\mathrm{MR} = \frac{1}{|\mathcal{C}|}\sum_{c=0}^{|\mathcal{C}|}\mathrm{R}_c \tag{4}$$

$$\mathrm{R}_c = \frac{\sum_{i=1}^{N}[y_i = \hat{y}_i = c]}{\sum_{i=1}^{N}[y_i = c]} \tag{5}$$

$$\mathrm{MF}_1 = \frac{1}{|\mathcal{C}|}\sum_{c=0}^{|\mathcal{C}|}\mathrm{F}_1^c \tag{6}$$

$$\mathrm{F}_1^c = 2 \cdot \frac{\mathrm{P}_c \cdot \mathrm{R}_c}{\mathrm{P}_c + \mathrm{R}_c} \tag{7}$$

$$\mathrm{JAcc} = \frac{1}{N}\sum_{i=0}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \tag{8}$$

To be able to compare our results with those of the first-ranked system in each task, the training, development, and test partitions provided by the organizers of each task were used. In some tasks the organization did not provide the development partition. In these cases, we have generated them by splitting the train set using a random sampling process. The sampling process selects 20% of the training set as development set, maintaining the original class distribution in both sets. We fine-tuned the TWilBERT models by using a grid search over batch size ([16,32]), learning rate ([1e−5, 5e−5, 1e−4, 5e−4]), and pooling strategy (averaging the contextualized embeddings or using the embedding of the [CLS] token). Furthermore, weighted cross-entropy was used to tackle with the class imbalance. Each experiment was repeated 3 times, and the best model on the development set was selected to be evaluated on the test set. In all the following subsections, we perform a comparison among M-BERT, the TWilBERT models and the best system of each competition. Additionally, we consider TW-Large without ROP to observe the behavior of the proposed loss signal, and a BERT model, with the same hyper-parameters than TW-Large, trained only with the Spanish Wikipedia (S-BERT) to observe how much the domain inconsistency between pre-training and fine-tuning affects the performance on downstream tasks. For the sake of simplicity, we added the results of these two systems in all the tables of the following subsections although their results are discussed in Section 6.

5.1. Topic classification

For topic classification, we used the dataset of the Classification of Spanish Election Tweets (COSET) task [33]. This task is intended to classify the topic discussed in a tweet into one of five topics related with the Spanish 2015 electoral cycle. The five topics are: Political Issues, Policy Issues, Campaign Issues, Personal Issues, and Other Issues. It is a single-input single-label task, where the MF_1 is used to evaluate and rank the systems. Table 2 show the results of M-BERT, TWilBERT models and the best system of the competition.

The TW-Base system outperforms the best system of the competition by +1.76 MF_1. Also, a large difference of +5.11 MF_1 can be observed between the results of TW-Base and TW-Large. M-BERT is competitive in this task, outperforming also the best system of the competition by +0.43 MF_1. However, TW-Base shows a better behavior than M-BERT, outperforming it by +1.33 MF_1.

5.2. Stance detection

For stance detection, we considered two different tasks on the same fact. On the one hand, the Stance Detection in Tweets on Catalan Independence (SDTC) task [35]. This dataset was collected by the organizers during the Catalan elections in September 2015, which have been interpreted by many political actors and citizens as a de facto referendum on the independence of Catalonia from Spain. On the other hand, the Multimodal Stance Detection in Tweets on Catalan 1Oct Referendum (MSDTC) task [36]. In this

Table 2
Results for COSET task.

System	MP	MR	MF_1	Acc
M-BERT	67.65	64.30	65.25	70.35
S-BERT	63.49	62.07	61.73	64.58
TW-Base	72.03	63.80	**66.58**	71.00
TW-Large	67.84	59.51	61.47	73.20
TW-Large (w/o ROP)	64.77	60.88	62.22	66.18
Best [34]	–	–	64.82	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

case, the dataset was collected by the organizers during the Catalan Referendum in October 2017. Along with the tweets, a context composed by the previous and the following tweet to each tweet is also provided.

The two competitions were proposed with the aim of detecting the stance of tweets (in favor, against or neutral) towards the target independence of Catalonia in Twitter messages written in Spanish. The SDTC task is a single-input single-label task whereas the MSDTC task can be addressed both as a single-input single-label task (if the context is discarded) or as a multiple-input single-label task. In order to evaluate and rank the systems for the SDTC task, MF_1 discarding the neutral class is used. For MSDTC, the evaluation metric is MF_1 considering the three classes.

Table 3 shows the results for the SDTC task. Neither M-BERT nor TWilBERT models outperform the best system of the competition, that is based on a combination of stylistic, structural and contextual features based on n-grams. This can be related with the low performance, observed in this task, obtained by systems based on distributed features in comparison to systems based on categorical features [35]. M-BERT and TW-Base obtained similar results, being both outperformed by TW-Large by +2.01 MF_1 and +1.86 MF_1 respectively.

Table 4 shows the results for the MSDTC task. In this case, we considered the task both as single-input (*sgl*) and multiple-input (*mpl*). For the *mpl* experiments, the central tweet and the next tweet are joined by means of a [SEP] token to compose the input.

In the *sgl* experiments, both TW-Base and TW-Large outperform the M-BERT system by +4.40 MF_1 in the best case. In the *mpl* experiments, the ranking of these systems is the same than for *sgl* experiments, being again the TW-Large the system that obtains the best results, by +4.05 MF_1 in comparison with M-BERT and by +2.43 MF_1 in comparison to TW-Base. It can be observed how the addition of context improves the results of all the systems. This could be favored by the NSP and ROP signals used during the pre-training of the models to learn coherence relationships among pairs of inputs. However, in the case of M-BERT, adding the context do not improve the results of TW-Large even without considering the context. This suggests that the ROP signal is better suited for the Twitter domain than the NSP signal. Both M-BERT and TWilBERT models with *sgl* and *mpl* input configurations, clearly outperform the results of the best system in the competition.

It is interesting to observe that the TW-Large without the ROP signal obtains similar results than M-BERT for the *sgl* experiments, however, when the context is included, M-BERT outperforms it by +1.57 MF_1. This shows that including a coherence signal in the training process, even it is not well suited for the Twitter domain, improves the capability of the models for reasoning with multiple inputs. Additionally, the improvement obtained when the context is considered for TW-Large without ROP is smaller compared with the improvements on the other models (1.52 vs 3.42 MF_1 for M-BERT, 1.52 vs 1.92 MF_1 for TW-Base and 1.52 vs 3.07 MF_1 for TW-Large). TW-Large without ROP signal is outperformed, in both *sgl* and *mpl* experiments, by both versions of TWilBERT that consider the ROP signal.

Table 3
Results for SDTC task.

System	MP	MR	MF_1	Acc
M-BERT	57.11	54.61	43.73	69.80
S-BERT	54.07	53.75	43.36	65.77
TW-Base	52.05	55.40	43.88	61.79
TW-Large	54.86	56.27	45.74	66.51
TW-Large (w/o ROP)	52.22	54.10	43.16	64.57
Best [37]	–	–	**48.88**	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 4
Results for MSDTC with *sgl* and *mpl* input configurations.

System	MP	MR	MF_1	Acc
M-BERT (*sgl*)	48.71	47.63	47.06	52.53
S-BERT (*sgl*)	50.87	50.06	49.53	55.14
TW-Base (*sgl*)	51.08	50.40	50.18	54.96
TW-Large (*sgl*)	52.62	51.48	51.46	55.23
TW-Large (w/o ROP) (*sgl*)	50.81	48.53	47.39	54.33
M-BERT (*mpl*)	54.12	51.14	50.48	56.68
S-BERT (*mpl*)	54.29	52.84	52.70	57.49
TW-Base (*mpl*)	56.23	52.62	52.10	58.03
TW-Large (*mpl*)	57.48	54.51	**54.53**	59.30
TW-Large (w/o ROP) (*mpl*)	55.72	49.71	48.91	54.06
Best [38]	–	–	28.02	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

5.3. Irony detection

The Irony Detection in Spanish Variants (IroSVA) task [39] was used to evaluate the behavior of our proposal for Irony Detection. The main objective of the IroSVA task is to identify the presence of irony in short messages (tweets and news comments) written in three different Spanish variants from Spain, Mexico, and Cuba. It is a single-input single-label binary classification task where the evaluation metric is the MF_1.

Tables 5–7 show respectively the results for the Spain, Mexico, and Cuba variants of the IroSVA task.

It can be seen that, for all the Spanish variants, the TWilBERT models outperform M-BERT and the best systems of the

Table 5
Results for the Spain variant of IroSVA task.

System	MP	MR	MF_1	Acc
M-BERT	69.21	69.63	69.40	72.50
S-BERT	69.02	70.88	69.32	71.17
TW-Base	73.17	72.75	**73.00**	76.17
TW-Large	71.89	70.00	70.70	75.00
TW-Large (w/o ROP)	69.43	67.50	68.16	73.00
Best [28]	–	–	71.67	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 6
Results for the Mexico variant of IroSVA task.

System	MP	MR	MF_1	Acc
M-BERT	61.92	62.42	62.10	65.66
S-BERT	69.35	67.14	67.86	73.00
TW-Base	68.79	67.91	68.27	72.50
TW-Large	69.30	70.01	**69.61**	72.50
TW-Large (w/o ROP)	64.99	66.06	65.28	68.00
Best [28]	–	–	68.03	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 7
Results for the Cuba variant of IroSVA task.

System	MP	MR	MF_1	Acc
M-BERT	69.72	65.87	66.75	73.00
S-BERT	63.94	64.75	64.19	67.17
TW-Base	67.17	67.50	67.32	70.67
TW-Large	70.00	66.88	**67.73**	73.33
TW-Large (w/o ROP)	67.04	66.00	66.40	71.00
Best [40]	–	–	65.96	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 8
Results for SemEval-Ec task.

System	JAcc
M-BERT	44.85
S-BERT	40.40
TW-Base	46.11
TW-Large	**48.60**
TW-Large (w/o ROP)	47.48
Best [42]	46.90

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

competition by a margin between +0.98 and +7.51 MF_1. For the Spain variant, TW-Base outperforms TW-Large by +2.3 MF_1, however, for the Mexico and Cuba variants, TW-Large outperforms TW-Base by +1.34 MF_1 and +0.41 MF_1, respectively.

5.4. Emotion detection

For emotion detection, we used the dataset provided in the SemEval-2018 Task 1: Affect in Tweets task [41]. This task includes an array of subtasks for inferring the affectual state of a person from a given tweet. In our case, we only focused on the subtask E-c (SemEval-Ec). It is a single-input multi-label task with 11 different classes $\mathcal{C} = \{$anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust $\}$ where the evaluation metric is JAcc.

Table 8 shows the results for the SemEval-Ec task. M-BERT obtained worse results than the best system of the competition, being outperformed by +2.05 JAcc. TW-Base outperforms M-BERT by +1.26 JAcc, but it showed a lower performance in comparison to the best system. TW-Large is the system that obtained the most competitive results, outperforming the best approach in the competition by +1.70 JAcc.

5.5. Hate speech detection

The dataset provided in the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter task (HatEval) [43] was used to evaluate our proposal. Specifically, we focused on the Subtask A, that is a single-input single-label binary classification, where the systems have to predict whether a tweet in Spanish with a given target (women or immigrants) contains hate speech. The evaluation metric used in this task is MF_1.

Table 9 shows the results for the HatEval task. It can be seen that, the M-BERT model and the two TWilBERT models outperform the best system of the competition. Specifically, M-BERT showed the best behavior, with an improvement of +2.98 MF_1 compared to the best system. M-BERT also outperformed TW-Large by +2.83 MF_1. The results of TW-Base and M-BERT are similar, being +0.68 MF_1 higher for the M-BERT model.

Table 9
Results for HatEval task.

System	MP	MR	MF_1	Acc
M-BERT	75.82	76.25	**75.98**	76.50
S-BERT	71.48	71.23	71.34	72.38
TW-Base	74.68	75.45	75.30	74.44
TW-Large	73.19	73.90	73.15	73.44
TW-Large (w/o ROP)	70.40	70.99	70.39	70.75
Best [44]	–	–	73.00	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

5.6. Sentiment analysis

For evaluating the performance in sentiment analysis, we used the datasets provided in the 2019 edition of the Workshop on Semantic Analysis at SEPLN (TASS). It is a single-input single-label task on four classes $\mathcal{C} = \{Negative, Neutral, None, Positive\}$ where the *None* class refers to tweets that do not express sentiment and the *Neutral* class refers to tweets where both *Positive* and *Negative* sentiments are expressed with the same intensity. The organizers of the task provided five different corpora, considering five different variants of the Spanish language from Spain, Mexico, Peru, Costa Rica, and Uruguay. The evaluation metric used for evaluating and ranking the systems is the MF_1.

Tables 10–14 show the results on the Spain, Costa Rica, Uruguay, Peru, and Mexico variants respectively. Except the case of Costa Rica variant, always there is a TWilBERT model that obtains better results than the best system of the competition. For all the variants, both TWilBERT models outperformed M-BERT, obtaining results up to +11.07 MF1.

These results show the lack of specialization of M-BERT in almost all the Spanish variants as those from Uruguay, Peru, Costa Rica, or Mexico. Besides that, the results of M-BERT in the Spain variant are more competitive than in the other variants. This may be due to the Spanish Wikipedia dataset used for training M-BERT does not include expressions from the Latin American variants.

Table 10
Results for the Spain variant of TASS task.

System	MP	MR	MF_1	Acc
M-BERT	49.17	49.36	48.89	59.38
S-BERT	43.08	41.98	42.82	51.82
TW-Base	51.96	50.75	50.84	59.14
TW-Large	52.10	51.94	**51.64**	59.50
TW-Large (w/o ROP)	50.47	48.00	48.55	55.51
Best [29]	50.50	50.80	50.70	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 11
Results for the Costa Rica variant of TASS task.

System	MP	MR	MF_1	Acc
M-BERT	46.85	46.49	46.20	50.52
S-BERT	45.33	43.16	43.37	52.06
TW-Base	49.40	50.51	49.46	57.20
TW-Large	50.05	51.06	50.24	59.52
TW-Large (w/o ROP)	46.30	46.95	46.21	50.85
Best [45]	58.88	45.40	**51.20**	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 12
Results for the Uruguay variant of TASS task.

System	MP	MR	MF_1	Acc
M-BERT	46.80	46.01	45.14	56.58
S-BERT	46.95	47.81	46.95	53.29
TW-Base	53.76	56.40	54.56	63.00
TW-Large	55.49	60.12	**56.21**	62.88
TW-Large (w/o ROP)	49.74	48.25	48.35	54.41
Best [29]	49.70	53.60	51.50	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 13
Results for the Peru variant of TASS task.

System	MP	MR	MF_1	Acc
M-BERT	46.58	40.60	37.90	37.43
S-BERT	38.75	39.51	38.48	42.14
TW-Base	45.83	45.36	**45.49**	48.22
TW-Large	48.40	46.28	45.01	44.06
TW-Large (w/o ROP)	47.45	41.46	39.30	39.48
Best [46]	46.20	44.60	45.40	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

Table 14
Results for the Mexico variant of TASS task.

System	MP	MR	MF_1	Acc
M-BERT	49.78	47.97	46.71	64.80
S-BERT	45.60	46.66	45.57	58.26
TW-Base	47.37	52.13	47.75	62.73
TW-Large	51.39	50.64	**50.38**	63.93
TW-Large (w/o ROP)	47.80	48.57	48.13	63.67
Best [29]	49.00	51.20	50.10	–

Bold value represents the best value of the official metric used in the competitions for ranking the systems.

6. Analysis of the model

In the previous section, we have studied the behavior of several TWilBERT and BERT models on a set of 14 different text classification datasets. The average results obtained in these 14 datasets are shown in Table 15. It can be seen how the TWilBERT models that consider the ROP signal outperform the M-BERT model by +3 points on average. By contrast, if the ROP signal is not used during pre-training, the results are very similar to those obtained by M-BERT. Also, if BERT is pre-trained only with the Spanish Wikipedia, the results obtained are 1 point lower on average than those obtained by M-BERT. These results suggest that the language, the domain and the coherence are relevant for obtaining better results on downstream tasks. According to the results (S-BERT < TW-Large (w/o ROP) < TW-Base < TW-Large) the language seems to be the less relevant aspect, followed by the domain consistency and the coherence. Regarding M-BERT, the multilingual pre-training shows a great capability for generalizing both for the language and the domain, however, the TW-Large (w/o ROP), which is pre-trained with substantially less data and does not consider inter-sentence coherence, obtains the same results. As shown in 15, is the combination of the specific language, domain and coherence signal that makes the difference.

We hypothesized that these improvements are possibly due to three main factors, related with the Twitter domain: the performance of the language model on tweets, the coherence between tweets learned by means of ROP (especially useful in multi-input tasks) and a lower redundancy among the patterns captured by the attention heads of TWilBERT in comparison to those of M-BERT. The aim of the next subsections is to analyze these three factors.

Table 15
Averaged results on all the text classification datasets.

	M-BERT	TW-Base	TW-Large	TW-Large (w/o ROP)	S-BERT
Avg	53.60	56.49	**56.89**	53.57	52.68

Bold value represents the best value of the table.

Table 16
$\gamma(\mathcal{D})$ results for each model.

	M-BERT	TW-Base	TW-Large	TW-Large (w/o ROP)	S-BERT
$\gamma(\mathcal{D})$	−4.16	−1.26	**−1.19**	−1.21	−3.19

Bold value represents the best value of the table.

6.1. Language model specialization

In this subsection, we study the specialization of the language models of M-BERT and TWilBERT to the Twitter domain. To do this, we built a dataset, $\mathcal{D}$, that contains all the tweets of the 14 datasets used in the previous section. This dataset is composed by 86,542 tweets. The aim of this analysis is to compute the probability that each language model assigns to $\mathcal{D}$, because the more probability a model assigns to the elements of $\mathcal{D}$, the more specialized this model is in $\mathcal{D}$. However, it is not easy to compute a probability for a text sequence using BERT-based language models due to they are bidirectional. Nevertheless, as shown in [47], BERT-based models can be interpreted as Markov Random Field language models. This interpretation can be used to compute unnormalized log-probabilities, which allow us to find the model that assigns a higher score to the tweets of $\mathcal{D}$. Eqs. from (9)–(12) show the process to compute the averaged unnormalized log-probabilities assigned by a model f_θ to the dataset $\mathcal{D}$.

$$\gamma(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\alpha(X_i) \tag{9}$$

$$\alpha(X) = \frac{1}{|X|}\sum_{t=1}^{|X|} log\ \phi_t(X) \tag{10}$$

$$\phi_t(X) = f_\theta(X_{\setminus t})_{x_t} \tag{11}$$

$$X_{\setminus t} = \{x_1, \ldots, [\text{MASK}], x_{t+1}, \ldots, x_{|X|}\} \tag{12}$$

where N is the number of samples in $\mathcal{D}$, $\phi_t(X)$ is the probability assigned by the model f_θ to the token t in the sample $X \in \mathcal{D}$, $\alpha(X)$ is the average of unnormalized log-probabilities for all the tokens in X, $\gamma(X)$ is the average of α for all the samples X, and $X_{\setminus t}$ is the tweet X where the token t is masked using the token [MASK], used as input for the model f_θ. The higher the probability assigned to the sample X, the closer to zero is $\alpha(X)$. Therefore, the more fitted a model f_θ is to the dataset $\mathcal{D}$, the closer to zero is $\gamma(\mathcal{D})$. Table 16 shows $\gamma(\mathcal{D})$ for M-BERT, S-BERT and the TWilBERT models.

It can be observed a correspondence between the results shown in Tables 15 and 16 for the M-BERT, TW-Base and TW-Large models, where the ranking among them is the same in both tables. However, in spite of $\gamma(\mathcal{D})$ for TW-Large without ROP is very similar to TW-Large, it obtains similar results to M-BERT. These results suggest that, although the performance of the language model is relevant for improving the performance on downstream tasks, there are other aspects that affect to the performance. A deeper study will be necessary in order to explain these results. The results of TW-Base and TW-Large are also very similar (difference of 0.07 in terms of $\gamma(\mathcal{D})$), being also similar their results averaged for the downstream tasks (difference of 0.40 points in average). It is interesting to see that $\gamma(\mathcal{D})$ is significantly higher for those mod-

Table 17
Accuracy of M-BERT and TWilBERT models for the two levels of coherence.

	$\mathcal{D}'_1$	$\mathcal{D}'_2$	Avg
M-BERT	47.68%	49.41%	48.55%
TW-Base	**55.43%**	86.15%	70.79%
TW-Large	54.57%	**91.27%**	**72.92%**

Bold values represent the best value of each column.

els trained with tweets, in comparison to those trained with a general domain. This suggests that the domain inconsistency between pre-training and fine-tuning affects negatively to the language modeling task on the downstream domain. TW-Large is the language model which best fits the dataset $\mathcal{D}$.

6.2. Coherence analysis

In this subsection, we carry out a study to analyze the coherence between tweet pairs captured by ROP (TWilBERT models) and NSP (M-BERT). To do this, we crawled a new dataset $\mathcal{D}'$ that contains 15,000 (tweet, reply) pairs unseen during the training phase. Following [14], we considered two different levels of coherence: topic prediction and inter-sentence coherence. From $\mathcal{D}'$, we generated two new datasets, $\mathcal{D}'_1$, $\mathcal{D}'_2$. Both datasets are composed of 15,000 positive pairs and 15,000 negative pairs. The purpose of both datasets is to evaluate M-BERT and the TWilBERT models in two binary classification tasks to classify positive and negative pairs with respect to the aforementioned levels of coherence. The positive instances of $\mathcal{D}'_1$ and $\mathcal{D}'_2$ are the samples of the dataset $\mathcal{D}'$, while the negative samples are built following the coherence level to study. On the one hand, the negative samples of $\mathcal{D}'_1$ are (tweet, reply) pairs where the reply of a tweet is randomly sampled among the replies of all the other tweets in $\mathcal{D}$. Thus, this coherence level is focused on topic relationships among tweets and their replies [14]. On the other hand, the negative instances of $\mathcal{D}'_2$ are (reply, tweet) shifted pairs, thus breaking sequentiality of the conversations to force models to detect inter-sentence coherence. Table 17 shows the Accuracy for M-BERT and for the TWilBERT models in both datasets.

As it can be seen in Table 17, M-BERT behaves like a random system in both datasets, thus showing a lack of specialization in the two levels of coherence when it is applied to the Twitter domain. The TWilBERT models better capture the coherence between pairs of tweets, obtaining statistically significant improvements both for topic prediction ($\mathcal{D}'_1$) and for inter-sentence coherence ($\mathcal{D}'_2$) in comparison to M-BERT. The same behavior was also observed in [14]. It is interesting to highlight that, in spite of M-BERT was trained over pairs of sentences by using the NSP signal,[2] this system obtained up to -7.75% of accuracy less than the TWilBERT models on $\mathcal{D}'_1$ dataset. Both TWilBERT models obtained similar results in $\mathcal{D}'_1$, without significant differences. However, TW-Large obtained significant improvements on $\mathcal{D}'_2$ in comparison to TW-Base. TW-large is the system which better captures the coherence, in average, for the two coherence levels.

6.3. Redundancy in attention heads

In this subsection we study the redundancy of the attention heads of each model. The aim of this study is to determine if some attention heads detect similar patterns than other attention heads in the same model. This way, a low redundant system must be specialized in detecting a wide variety of patterns in each abstraction level, thus improving the performance in downstream tasks [48].

To do this analysis, we computed the Jensen-Shannon Divergence (JSD) among all the pairs of attention heads in all the Transformer layers, in the same way that [49]. We computed the distance between the attention distributions of two heads, H_i and H_j, as shown in Eq. (13).

$$J = \sum_{X \in \mathcal{D}'} \sum_{t \in X} \mathrm{JSD}(H_i(t), H_j(t)) \tag{13}$$

[2] These pairs were built by the authors of [6] in the same way that we built the $\mathcal{D}'_1$ dataset. However, they used sentences from Wikipedia.

We applied multidimensional scaling to project the JSD among the attention heads in two dimensions. This projection is shown in Figs. 3–5 for M-BERT, TW-Base and TW-Large, respectively; where L indicates the layer to which each head belongs.

As it can be observed, for all the models, there are several clusters of heads that behave similarly. Specifically, attention heads in the same layers tends to get closer, which was also observed by the authors of [49]. Furthermore, they mentioned that "one possibility for the apparent redundancy in BERT's attention heads is the use of attention dropout, which causes some attention weights to be zeroed-out during training". However, the TWilBERT models, that

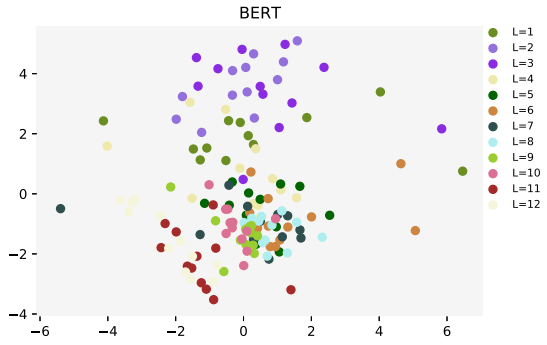

Fig. 3. Visualization of JSD divergences among M-BERT attention heads embedded in two dimensions.

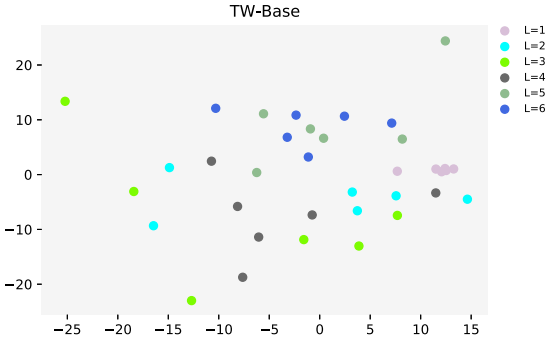

Fig. 4. Visualization of JSD divergences among TW-Base attention heads embedded in two dimensions.

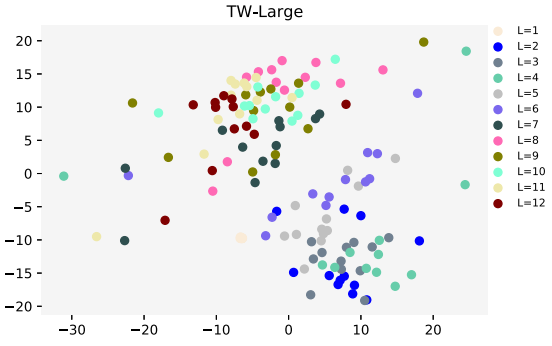

Fig. 5. Visualization of JSD divergences among TW-Large attention heads embedded in two dimensions.

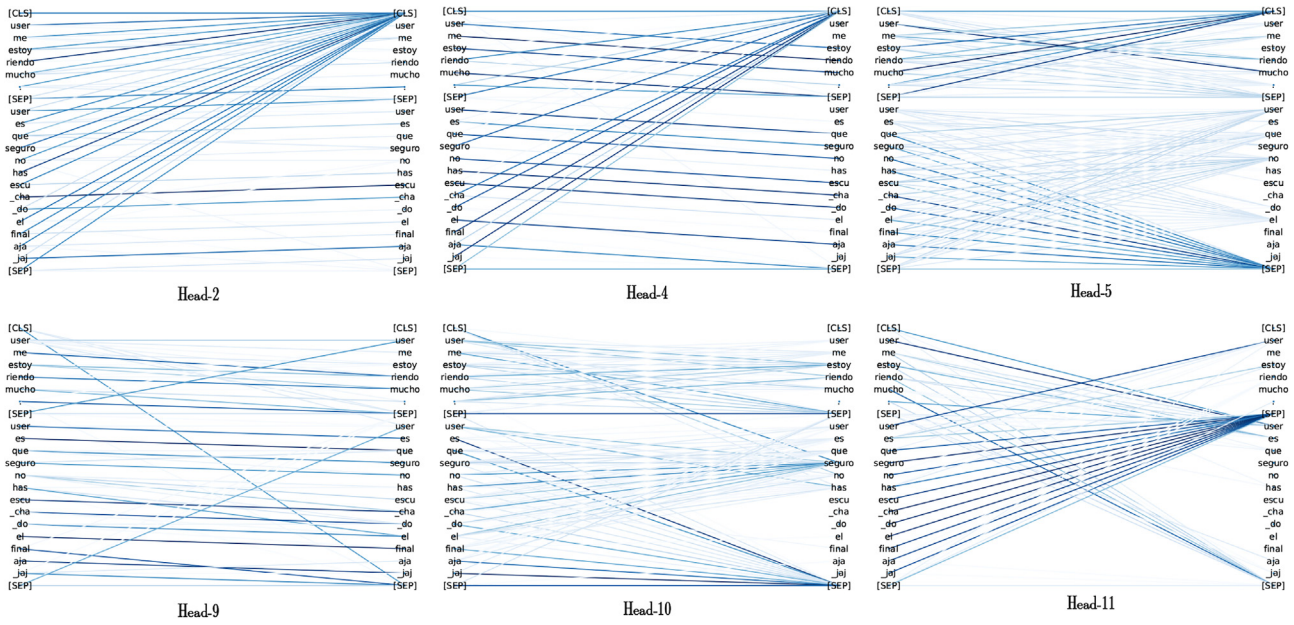

Fig. 6. Attention weights for the sample "*[CLS] @user me estoy riendo mucho. [SEP] @user es que seguro no has escu _cha _do el final aja _jaj [SEP]*" (tweet and reply are separated by the intermediate [SEP] token). The English translation of this pair is: "*[CLS] @user I'm laughing a lot. [SEP] @user is that surely you have not heard the end ahaha [SEP]*".

do not use any kind of dropout, also shown this inner-layer redundancy.

The system that shows less redundancy is TW-Base, possibly because the reduced number of attention heads forces a higher specialization of these heads. M-BERT and TW-Large show a similar behavior. However, the JSD among the M-BERT heads is more concentrated in a reduced space ($x_1 \in [-6,6], x_2 \in [-6,6]$) than in TW-Large ($x_1 \in [-20,20], x_2 \in [-30,20]$). It can be also observed that the attention heads are grouped in two different clusters. The first cluster is composed by the first half of the heads ($L \in \{1,6\}$) and the second cluster is composed by the second half ($L \in \{7,12\}$). The inter-class and the intra-class distances between the two clusters are higher in TW-Large than in M-BERT, which suggests that TW-Large is less redundant than M-BERT and, thus, more specialized in computing different patterns at each abstraction level.

From the two aforementioned clusters, we randomly selected three attention heads to observe what patterns they capture. Specifically, we selected heads 2, 4, and 5 (from the first cluster) and heads 9, 10, and 11 (from the second cluster). Fig. 6 shows the attention weights of these heads for a given sample. The first row refers to the heads 2, 4, and 5, and the second row refers to the heads 9, 10, and 11.

Several surface-level patterns can be observed in Fig. 6. Heads 2, 9, and 4, attend to the previous, next, and 2 position next token, respectively. Heads 5 and 10 are focused on the separation between the tweet and its reply. This separation is clearer in head 5, where the tweet attends to the [CLS] token and the reply attends to the last [SEP] token. In head 10, the attentions are more scattered than in the head 5. In general, it is also observed a large amount of attention to the tokens [CLS] and [SEP], especially in the head 11, where all the tokens attend to the intermediate [SEP] token.

7. Conclusions and future work

In this work, we have presented TWilBERT, a specialization of BERT both for the Spanish language and the Twitter domain. Two TWilBERT models were trained on a dataset of 47 million of (tweet, reply) pairs. To our knowledge, this is the first work that proposes a full specialization of BERT for the Twitter domain, taking coherence into account by means of a novel Reply Order Prediction signal on Twitter conversations.

We performed an extensive evaluation and analysis of the TWilBERT models in comparison to Multilingual BERT. TWilBERT models outperformed Multilingual BERT on 14 different datasets of text classification tasks such as irony detection, sentiment analysis, emotion detection, hate speech detection, stance detection, and topic detection.

The proposed models seem to capture better the topic and inter-sentence coherence between tweets, they are a better language models on the Twitter domain, and their attention heads shown lower redundancy, capturing a greater diversity of patterns, compared to Multilingual BERT.

TWilBERT is provided as a framework to train, evaluate, and fine-tune models, and it allows the researchers to use some recent techniques proposed in the literature such as cross-sharing parameter layers, factorized embedding parameterization [14] or Product Key Memory layers [21]. In addition, we provided the weights of the pre-trained models used in this paper, that are freely available in.[1]

As future work, we plan to increase the size of the training dataset in order to release more competitive TWilBERT models, along with further improvements on the architecture. Another interesting line of research is to improve the learning of the coherence. In this respect, we plan to propose new training signals to learn coherence in Twitter timelines and threads. Also, it is interesting to explore how the language and the domain inconsistency between pre-training and fine-tuning affects to the results of BERT models on other languages and domains such as clinical or economical ones. We plan to continue maintaining and updating the framework to include state-of-the-art techniques for BERT models.

CRediT authorship contribution statement

José Ángel González: Data curation, Formal analysis, Investigation, Software, Writing - original draft. **Lluís-F. Hurtado:** Conceptualization, Formal analysis, Data curation, Investigation,

Methodology, Writing - review & editing, Funding acquisition. **Ferran Pla:** Data curation, Formal analysis, Investigation, Methodology, Validation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades and FEDER founds under project AMIC (TIN2017-85854-C4-2-R), and the Generalitat Valenciana under GiSPRO (PROMETEU/2018/176) and GUAITA (INNVA1/2020/61) projects. Work of José Ángel González is financed by Universitat Politècnica de València under grant PAID-01-17.

References

[1] 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, URL:http://dl.acm.org/citation.cfm?id=2999792.2999959.

[2] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606.

[3] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237, https://doi.org/10.18653/v1/N18-1202, URL:https://www.aclweb.org/anthology/N18-1202.

[4] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339, https://doi.org/10.18653/v1/P18-1031, URL:https://www.aclweb.org/anthology/P18-1031.

[5] B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in translation: Contextualized word vectors., in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), NIPS, 2017, pp. 6297–6308. URL:http://dblp.uni-trier.de/db/conf/nips/nips2017.html#McCannBXS17.

[6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805. arXiv:1810.04805. URL: http://arxiv.org/abs/1810.04805.

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

[8] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, CoRR abs/1906.01502. arXiv:1906.01502. URL:http://arxiv.org/abs/1906.01502.

[9] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, Bertje: A dutch bert model (2019). arXiv:1912.09582.

[10] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french (2019). arXiv:1912.05372.

[11] L. Martin, B. Muller, P.J. Ortiz Suárez, Y. Dupont, L. Romary, É. Villemonte de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a Tasty French Language Model, arXiv e-prints (2019) arXiv:1911.03894 arXiv:1911.03894.

[12] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: Bert for finnish (2019). arXiv:1912.07076.

[13] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), Vol. 2481, CEUR, 2019. URL:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[14] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020. URL:https://openreview.net/forum?id=H1eA7AEtvS.

[15] F. Chollet, et al., Keras, URL: https://keras.io (2015).

[16] 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. arXiv:1907.11692. URL: http://arxiv.org/abs/1907.11692.

[17] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, CoRR abs/1907.10529. arXiv:1907.10529. URL: http://arxiv.org/abs/1907.10529.

[18] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 785–794, https://doi.org/10.18653/v1/D17-1082, URL:https://www.aclweb.org/anthology/D17-1082.

[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet (2019). URL: http://arxiv.org/abs/1906.08237.

[20] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, C.-J. Hsieh, Large batch optimization for deep learning: Training bert in 76 minutes, in: International Conference on Learning Representations, 2020. URL:https://openreview.net/forum?id=Syx4wnEtvH.

[21] G. Lample, A. Sablayrolles, M. Ranzato, L. Denoyer, H. Jégou, Large memory layers with product keys, in: Advances in Neural Information Processing Systems 32 Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 2019, pp. 8546–8557, URL:http://papers.nips.cc/paper/9061-large-memory-layers-with-product-keys.

[22] I. Beltagy, K. Lo, A. Cohan, Scibert: Pretrained language model for scientific text, in: EMNLP, 2019, arXiv:arXiv:1903.10676.

[23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics doi:10.1093/bioinformatics/btz682.

[24] J. Mao, W. Liu, Factuality classification using the pre-trained language representation model BERT, 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. 126–131. URL:http://ceur-ws.org/Vol-2421/FACT_paper_3.pdf.

[25] J. Iranzo-Sánchez, R. Ruiz-Dolz, VRAIN at irosva 2019: Exploring classical and transfer learning approaches to short message irony detection, 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.

[26] J. Mao, W. Liu, A bert-based approach for automatic humor detection and scoring, 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. 197–202. URL:http://ceur-ws.org/Vol-2421/HAHA_paper_8.pdf.

[27] M. Pastorini, M. Pereira, N. Zeballos, L. Chiruzzo, A. Rosá, M. Etcheverry, Retuyt-inco at TASS 2019: Sentiment analysis in spanish tweets, 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. 605–610. URL:http://ceur-ws.org/Vol-2421/TASS_paper_6.pdf.

[28] J. González, L. Hurtado, F. Pla, Elirf-upv at irosva: Transformer encoders for spanish irony detection, 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.

[29] J. González, L. Hurtado, F. Pla, Elirf-upv at TASS 2019: Transformer encoders for twitter sentiment analysis in spanish, 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. 571–578. URL:http://ceur-ws.org/Vol-2421/TASS_paper_2.pdf.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (56) (2014) 1929–1958, URL:http://jmlr.org/papers/v15/srivastava14a.html.

[31] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71, https://doi.org/10.18653/v1/D18-2012, URL:https://www.aclweb.org/anthology/D18-2012.

[32] 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. URL: http://arxiv.org/abs/1412.6980.

[33] M.G. Fayos, T. Baviera, G. Llorca, J. Gámir, D. Calvo, P. Rosso, F.M.R. Pardo, Overview of the 1st classification of spanish election tweets task at ibereval 2017, in: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing

(SEPLN 2017), Murcia, Spain, September 19, 2017, 2017, pp. 1–14. URL:http://ceur-ws.org/Vol-1881/Overview2.pdf.

[34] J. González, F. Pla, L. Hurtado, Elirf-upv at ibereval 2017: Classification of spanish election tweets (COSET), in: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017, 2017, pp. 55–60. URL:http://ceur-ws.org/Vol-1881/COSET_paper_7.pdf.

[35] P. Rosso, F.M.R. Pardo, Author profiling in social media: The impact of emotions on discourse analysis, in: Statistical Language and Speech Processing - 5th International Conference, SLSP 2017, Le Mans, France, October 23–25, 2017, Proceedings, 2017, pp. 3–18. doi:10.1007/978-3-319-68456-7_1. URL: https://doi.org/10.1007/978-3-319-68456-7_1.

[36] M. Taulé, F.M.R. Pardo, M.A. Martí, P. Rosso, Overview of the task on multimodal stance detection in tweets on catalan #1oct referendum, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, 2018, pp. 149–166. URL:http://ceur-ws.org/Vol-2150/overview-Multistance18.pdf.

[37] M. Lai, A.T. Cignarella, D.I.H. Farías, itacos at ibereval2017: Detecting stance in catalan and spanish tweets, in: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017, 2017, pp. 185–192. URL:http://ceur-ws.org/Vol-1881/StanceCat2017_paper_2.pdf.

[38] I. Segura-Bedmar, Labda's early steps toward multimodal stance detection, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, 2018, pp. 180–186. URL:http://ceur-ws.org/Vol-2150/MultiStanceCat_paper3.pdf.

[39] R. Ortega, F. Rangel, I. Hernández, P. Rosso, M. Montes, M. Pagola, Overview of the task on irony detection in spanish variants, in: Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019), CEUR Workshop Proceedings, 2019.

[40] 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.

[41] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1–17, https://doi.org/10.18653/v1/S18-1001, URL:https://www.aclweb.org/anthology/S18-1001.

[42] Y. Kim, H. Lee, K. Jung, AttnConvnet at SemEval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 141–145, https://doi.org/10.18653/v1/S18-1019, URL:https://www.aclweb.org/anthology/S18-1019.

[43] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63, https://doi.org/10.18653/v1/S19-2007, URL:https://www.aclweb.org/anthology/S19-2007.

[44] J.M. Pérez, F.M. Luque, Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 64–69, https://doi.org/10.18653/v1/S19-2008, URL: https://www.aclweb.org/anthology/S19-2008.

[45] M. Pastorini, M. Pereira, N. Zeballos, L. Chiruzzo, A. Rosá, M. Etcheverry, Retuyt-inco at TASS 2019: Sentiment analysis in spanish tweets, 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. 605–610. URL:http://ceur-ws.org/Vol-2421/TASS_paper_6.pdf.

[46] F.M. Luque, Atalaya at TASS 2019: Data augmentation and robust embeddings for sentiment analysis, 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. 561–570. URL:http://ceur-ws.org/Vol-2421/TASS_paper_1.pdf.

[47] A. Wang, K. Cho, BERT has a mouth, and it must speak: BERT as a Markov random field language model, in: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 30–36. doi:10.18653/v1/W19-2304. URL:https://www.aclweb.org/anthology/W19-2304.

[48] J. Li, Z. Tu, B. Yang, M.R. Lyu, T. Zhang, Multi-head attention with disagreement regularization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2897–2903, https://doi.org/10.18653/v1/D18-1317, URL:https://www.aclweb.org/anthology/D18-1317.

[49] K. Clark, U. Khandelwal, O. Levy, C.D. Manning, What does BERT look at? an analysis of bert's attention, CoRR abs/1906.04341. arXiv:1906.04341. URL: http://arxiv.org/abs/1906.04341.

José Ángel González is a Computer Science Ph.D student, currently working at Natural Language Engineering and Pattern Recognition Group (ELiRF) at the Universidad Politécnica de Valencia. His main interest is the Natural Language Processing field, and, specifically, text classification for social media analysis and automatic summarization.

Lluís-F. Hurtado received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 2004. He is currently a Associate Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He has published over 80 research papers being involved in many research projects. His research interests cover many areas within speech processing and natural language processing, including spoken dialog systems, voice-activated question answering, spoken language understanding, and sentiment analysis.

Ferran Pla received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 2000. He is currently an Associate Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He has published over 50 papers being involved in many research projects. His research interests cover many areas within natural language processing, including: POS tagging, parsing, name entity recognition, word sense disambiguation, and sentiment analysis in social media.