From Recurrency to Attention in Opinion Analysis

Comparing RNN vs Transformer Models

Rosa María Montañés-Salas[1], Rafael del-Hoyo-Alonso[1], and Rocío Aznar-Gimeno[1]

Technological Institute of Aragón (ITAINNOVA), María de Luna, 7–8, Zaragoza, Spain
{rmontanes,rdelhoyo,raznar}@itainnova.es

Abstract. This paper describes the participation of ITAINNOVA at the Sentiment Analysis in Twitter task (TASS) framed within the new evaluation forum IberLEF (Iberian Languages Evaluation Forum). This work explores two different Deep Learning approaches, validating their performance on both subtasks (Monolingual and Crosslingual Sentiment Analysis). The first one is an embedding-based strategy combined with bidirectional recurrent neural networks, which receives the name Char Bi-LSTM network, and the second one, a recent language representation model, called BERT (Bidirectional Encoder Representations from Transformers). Although the performance of the second approach is not recognized in the official results of the task, we also present this approach, which performance has been reasonably remarkable and greater than the first approach.

Keywords: Sentiment analysis · Twitter · Deep learning

1 Introduction

The Workshop on Sentiment Analysis, framed within the new evaluation forum IberLEF (Iberian Languages Evaluation Forum) and celebrated under the umbrella of the International Conference of the Spanish Society for Natural Language Processing (SEPLN), known as *TASS*, has become one of the most important events in the field of semantic analysis over Spanish written texts [6]. The workshop is an ideal meeting point for the exchange of ideas between professionals and researchers in the field of Natural Language Processing (NLP) in general and sentiment analysis in particular. The aim of the proposed task is to promote the current state of development of polarity classification systems at tweet level in Spanish. As of TASS 2018 edition [14] the challenges of multilinguality and generalization capacity of the systems arised in the form of new subtasks.

Subtask 1 (Monolingual Sentiment Analysis) is focused on single-language analysis using the same variant for training, validation and testing, while subtask 2 (Cross-lingual Sentiment Analysis) aims to evaluate the dependency of systems on a language.

In this context we have explored two different approaches based on advanced deep learning techniques. The first, and the official one, applies a traditional *feature-based* strategy, which commonly uses pre-trained data representations as feature inputs of the task-oriented model architecture. The second approach is based on the transfer learning method, where a model is trained with one specific learning objective and then is reused as the starting point to learn how to solve a different problem. One of the most recent, effective and adaptable works based on deep bidirectional transformers is the BERT (*Bidirectional Encoder Representations from Transformers*) implementation, which has shown considerable improvements over a wide range of NLP tasks [5]. We have participated on both subtasks in order to validate our proposed solutions.

The paper is organised as follows: after this introduction, we will briefly describe the set of works which has inspired our approaches. In section 3 detailed architectures of the solutions are presented, followed by the details of the experiments carried out in section 4. Results of those experiments will be presented as well. Finally, in section 5 we will summarize the main conclusions drawn during the experimentation and future working directions.

2 Related work

Language modeling is one of the most difficult problems yet to resolve in the NLP field. In recent years this problem has been tackled by the generation of dense semantic representations obtained through training unsupervised algorithms on large text corpora. In the case of the Spanish language, the most commonly used resources are Wikipedia and the Spanish Billion Word Corpus compiled by Cardellino [3]. There have been developed multiple approaches in order to obtain pretrained word embeddings such as: word2vec [15], GloVe [17] and FastText [2]. From a character level perspective, some authors have reported performance improvements by using char embeddings on language modeling [10] and fine-grained sentiment analysis [9].

Moreover, the effectiveness of recurrent neural networks has been widely demonstrated over several NLP tasks, in particular the use of Long Short-Term Memory networks (LSTMs) and its bidirectional version (Bi-LSTMs) (for example in [13] and [12]). Unlike traditional recurrent networks, these networks have the characteristic of learning long-term dependencies, allowing a greater window of context information, thus improving performance on language related tasks, where the context significantly influences semantic analysis.

Combination of these techniques has lead to the development of complex but increasingly powerful architectures such as the *Char Bi-LSTM networks* which are used mainly for sequence tagging tasks as Named Entity Recognition and Classification (NERC) problem, as shown in [4] and [11].

While the previous models use word embeddings in order to introduce the word concept and RNN (LSTM) models that do model word order directly and explicitly track states across the sentence. Our second approach uses BERT, a transformers based architecture, which in contrast to LSTM, where order is important, BERT does not have an explicit notion of word order beyond marking each word with its absolute-position embedding, the language modelling relies mainly on attention mechanisms [19,5,8]. BERT is a recent natural language processing model that has shown groundbreaking results in many tasks such as question answering, natural language inference and paraphrase detection [5], but has been scarcely tested on Spanish language until recently ([18], [1]).

3 Proposed approaches

Inspired by the general conclusions derived from the workshop on Semantic Analysis in 2018 [14] and by our previous contributions on the Sentiment Analysis tasks [16], we have followed the line of deep learning based solutions. Firstly, we have explored an embedding-based strategy combined with bidirectional recurrent neural networks, which receives the name *Char Bi-LSTM network*. Secondly, motivated by the reported improvements of BERT in English language modeling, we decided to study its efficiency on this challenging Spanish Tweet classification task.

3.1 Char Bi-LSTM network

As stated before, joining the strengths of embedding-based language models with neural architectures focused on temporal sequence learning, has shown promising results on some NLP tasks. Consequently, our first approach relies on the architecture shown in the following figure (Fig. 1), in order to solve the polarity classification problem over Spanish written texts.

Proposed architecture learns a representation of input documents as a concatenation of self-learned char-embeddings with sequence word embeddings (loaded from Spanish pretrained word embedding models). This representation feeds the bidirectional LSTM module, which could be composed of various layers. The output class is obtained through a softmax cross entropy layer which returns the probability for each label.

3.2 BERT classifier

Currently, research community is studying a new typology of language architectures that goes beyond the traditional vector representations, such as ELMO (Embedded from Language Model), GPT (Generative Pre-trained Transformer), GPT-2 and BERT. The goal of these architectures is to develop models, increasingly complex, for language understanding. Both Open AI GPT and BERT [5] use the transformer architecture to learn text representations. The main difference between them is that BERT uses a bidirectional transformer (from left to

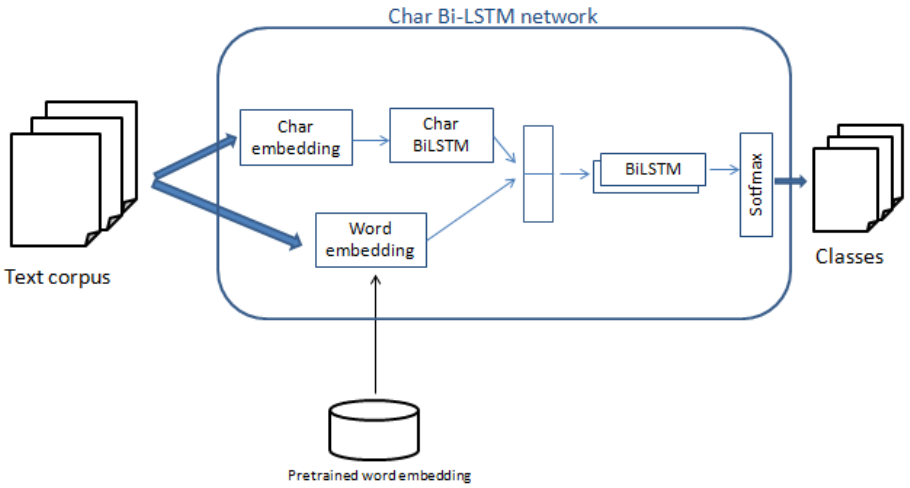

Fig. 1: Char Bi-LSTM architecture.

right and from right to left) instead of a directional transformer (from left to right). Regarding ELMo, it uses a shallow concatenation layer, while BERT uses a deep neural network.

In its basic form, BERT includes three separate mechanisms: an encoder that reads text input, a group of pooler layer and a decoder that produces a prediction or a classification layer for the task. When learning linguistic models, it is difficult to define a prediction objective. Many models predict the next word in a sequence (for example, "The man traveled to its work by ___"), a directional approach that inherently limits contextual learning. To overcome this challenge, BERT uses two training strategies. In the first method, named *"masked LM"* due to the masking procedure applied to train the Language Model, before entering sequences of words in BERT, 15 % of the words in each sequence are replaced by a token [MASK]. Next, the model attempts to predict the original value of the masked words, based on the context provided by the other unmasked words in the sequence. This method tries to obtain relationships between the existing words in a sentence. The second method is the prediction of the following sentence, so try to offer continuity in the discourse. In this training process, the model receives pairs of sentences as input and learns to predict whether the second sentence of the pair is the next sentence of the original document. During training, 50% of the entries are a pair in which the second sentence is the next sentence of the original document, while in the other 50% a random sentence of the corpus is chosen as the second sentence.

Released pretrained language models, build by these methodologies, include a variety of options: English and multilingual models (including Spanish), cased and uncased models, and the possibility to choose between a base or large version.

A detailed list of released models can be found at Google research Github[1]. On

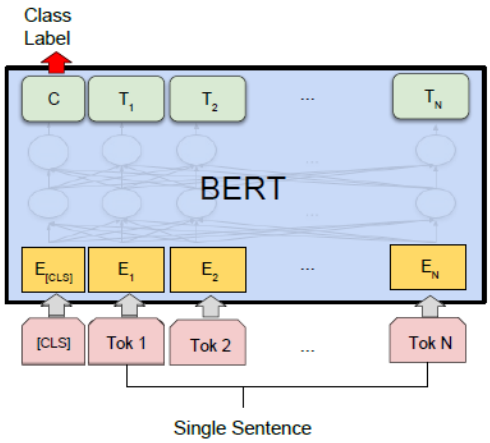

Fig. 2: BERT fine-tuned architecture for single sentence classification ([5])

top of pretrained language models, it also provides the functionality for fine tuning them to update learnt weights, by re-training the language model using our own text corpus.

4 Experiments

With the goal of providing an experimental setup for reproducing the results exposed in this section, we include a description of the datasets, model configuration parameters and execution environment used in the course of our trials.

Datasets for training, evaluation and test have been provided by TASS organization [7] . InterTASS dataset collects five Spanish variants in this edition: ES (Spain), PE (Peru), CR (Costa Rica), UR (Uruguay) and MX (Mexico). All of them are annotated with 4 different levels of opinion intensity: P (positive), NEU (neutral), NONE (no opinion), N (negative).

Model configuration parameters have been established through an exhaustive searching process mainly focused on the Spanish model. It has been decided to use this parameterization for the training and evaluation of the rest of the languages considered. The following subsections indicate the parameters and values used for each of the approximations studied in order to allow the reproducibility of the results obtained. All the experiments have been configured to use Nvidia GPUs (Tesla V100 and TITAN Xp).

[1] https://github.com/google-research/bert

4.1 Char Bi-LSTM network

Table 1 shows experimental settings for both Char Bi-LSTM models studied (*Char1* and *Char2*):

Table 1: Char Bi-LSTM Hyperparameters

Parameter	CHAR1	CHAR2
dim_word	300	300
dim_char	50	100
type_pretrained	Word2Vec	FastText
train_embeddings	True	True
nepochs	100	100
dropout	0.5	0.5
batch_size	24	100
opt_method	Adam	Rmsprop
lr	0.0001	0.0009
lr_decay	0.9	0.9
layers	2	2
hidden_size_char	140	10
hidden_size_lstm	50	30
hidden_size_lsmt2	5	5

Word2Vec and FastText pretrained models have been trained over our own corpus built from the SBWC corpus merged with a set of 10000 tweets approximately, retrieved from the Twitter public API over the past year.

During the first trials, we observed a fast overfitting of the network, which caused an slightly accuracy improvement due to fact that the model was discarding NEU and NONE classes and getting right extreme opinions (positive and negative).

4.2 BERT classifier

Experimental settings of BERT model are listed below (non mentioned parameters, available on the original implementation, has been left at their default values):

- **bert_model** : bert-base-multilingual-uncased
- **train_batch_size** : 32
- **gradient_accumulation_steps** : 1
- **num_train_epochs** : 5
- **learning_rate** : 1e-5
- **warmup_proportion** : 0.1
- **max_seq_length** : 70

4.3 Results

This section collects the set of official and unofficial results retrieved from the experiments previously described. It also includes results from our contribution on TASS previous edition [16] (stated as *Model2018*) for comparison purposes.

Just Char Bi-LSTM models results are considered official. Non official results, as BERT monolingual and crosslingual metrics, have been calculated making use of the evaluation scripts provided by the organization at the beginning of the competition, so that metrics are obtained under the same conditions as the official ones.

Subtask 1 (Monolingual Sentiment Analysis) is focused on single-language analysis using the same variant for training, validation and testing, while subtask 2 (Cross-lingual Sentiment Analysis) aims to evaluate the dependency of systems on a language.

Monolingual Training, validation and test using each InterTASS dataset independently.

Table 2: F1 score for each language.

Model	ES	CR	MX	PE	UY
Char1	0.3729	0.3654	0.3977	0.3168	0.3658
Char2	0.3659	n/a	0.4073	n/a	n/a
Bert	0.5145	0.4932	0.5056	0.4580	0.5380
Model2018	0.3830	n/a	n/a	n/a	n/a

Crosslingual Training a selection of any dataset and use a different one to test. In our case we have trained independently on ES and MX datasets, choosing finally the MX model to be tested on the rest of languages, given its superior results.

Table 3: F1 score for MX model crossed with rest of languages

Model MX	CR	ES	PE	UY
Char1	0.2968	0.2343	0.2819	0.2855
Bert	0.4500	0.4890	0.4341	0.5045

In the case of *Model2018*, Spanish (ES) variant was selected to be tested against available variants in that edition, obtaining a F1 score of 0.4090 for CR and 0.3670 for PE.

5 Conclusions and future work

Within the previous edition of TASS [14] we obtained some rather discouraging official results, so that we decided to explore a more complete deep learning approach based on recurrent neural networks and embedding representations. Unfortunately, results obtained in this editions are very close to the previous ones, probably due to the excessive computational complexity of the joint embedding model for a sentence level classification task on short texts. This circumstance led us to consider a less traditional approach such as BERT, based on attention mechanisms which has shown a good generalization capability in a great variety of English-NLP tasks. We have been able to confirm that its Spanish language model works surprisingly well on the sentiment analysis task, and furthermore it adapts seamlessly against different variants of the same language. Therefore, we can conclude that models based on deep learning continue to be one of the most successful approaches from a computational point of view.

Nevertheless, studied approaches have certain limitations, such as the ability to distinguish between NEU and NONE labels. It has been observed systematically the difficulty of the algorithms to learn this classification due to their semantic proximity. Furthermore, the multilingual challenge on Twitter publications analysis remains open and gives much room for improvement.

As future work lines we expect to explore further the re-training of the Spanish language model with a larger corpus and searching for optimal parameters pursuing a significant improvement in this model performance, as well as research and get deep insights in the use of attention models on natural language analysis.

Acknowledgements

This work has been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon. We also thank the support of the FSE Operative Programme for Aragon 2014–2020 (IODIDE research group).

References

1. Benballa, M., Collet, S., Picot-Clemente, R.: Saagie at semeval-2019 task 5: From universal text embeddings and classical features to domain-specific text classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 469–475 (2019)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), https://crscardellino.github.io/SBWCE/
4. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016)

5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
6. Díaz-Galiano, M.C., et al.: Overview of tass 2019. CEUR-WS, Bilbao, Spain (2019)
7. García Cumbreras, M.Á., Martínez Cámara, E., Villena Román, J., García Morera, J.: Tass 2015 the evolution of the spanish opinion mining systems (2016)
8. Jawahar, G., Sagot, B., Seddah, D.: What does bert learn about the structure of language? In: 57th Annual Meeting of the Association for Computational Linguistics (ACL) (July 2019)
9. Jebbara, S., Cimiano, P.: Improving opinion-target extraction with character-level word embeddings. arXiv preprint arXiv:1709.06317 (2017)
10. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
11. Limsopatham, N., Collier, N.H.: Bidirectional lstm for named entity recognition in twitter messages. Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (2016)
12. Lin, B.Y., Xu, F., Luo, Z., Zhu, K.: Multi-channel bilstm-crf model for emerging named entity recognition in social media. In: Proceedings of the 3rd Workshop on Noisy User-generated Text. pp. 160–165 (2017)
13. Liu, P., Joty, S., Meng, H.: Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1433–1443 (2015)
14. Martínez Cámara, E., Almeida Cruz, Y., Díaz Galiano, M.C., Estévez-Velarde, S., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Montejo Ráez, A., Montoyo, A., Muñoz, R., et al.: Overview of tass 2018: Opinions, health and emotions. CEUR Workshop Proceedings **2172** (2018)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Montañés, R., Aznar, R., Hoyo, R.D.: Aplicación de un modelo híbrido de aprendizaje profundo para el análisis de sentimiento en twitter(application of a hybrid deep learning model for sentiment analysis in twitter). In: TASS@SEPLN (2018-09)
17. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
18. Siddiqua, U.A., Chy, A.N., Aono, M.: Kdehateval at semeval-2019 task 5: A neural network model for detecting hate speech in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 365–370 (2019)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)