Palomino-Ochoa at TASS 2020: Transformer-based Data Augmentation for Overcoming Few-Shot Learning

Daniel Palomino[a], José Ochoa-Luna[a]

[a]*Universidad Católica San Pablo, Quinta Vivanco s/n St, Arequipa, Arequipa, Peru*

Abstract

This paper describes the participation of the Department of Computer Science at Universidad Católica San Pablo (UCSP) for the TASS 2020 Workshop. We have developed sentiment analysis algorithms for the monovariant and multivariant challenges. In both cases, our approach is based on transfer learning using BERT language modeling. We also propose a procedure based on this language model to generate contextual data augmentation aimed to increase the training dataset and prevent overfitting. Our design choices allow us to achieve comparable state-of-the-art results regarding the TASS benchmark datasets provided.

Keywords

Transformer, Few–Shot Learning, Data Augmentation, Sentiment Analysis

1. Introduction

Deriving an effective algorithm for Spanish Twitter sentiment analysis has been long pursued since the first TASS challenge in 2012 [1]. Nowadays, despite recent advances in algorithms (Deep Learning [2]) and word encoding (embeddings [3, 4, 5]), the basic polarity detection task has not been completely solved. Moreover, whereas it is usually claimed that a transfer learning approach can solve any classification tasks in NLP, smoothly [6]; this is not usually the case when we applied it to Spanish. Hence, the task becomes harder when several language variants are provided. In fact, low Macro-F1 values were reported in a previous TASS workshop [7].

In this paper we still rely on transfer learning to solve the classification task, but our approach has been carefully designed bearing in mind that the Spanish language has several variants.

In NLP it is common to use text input encoded as word embeddings. Those embeddings, which allow us to encode semantic similarities among words, can be defined using several approaches such as Word2vec [4], Glove [5] and FastText [3], to name a few. When we reuse pre-trained word embeddings in several tasks, we are indirectly employing a transfer learning scheme.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: daniel.palomino.paucar@ucsp.edu.pe (D. Palomino); jeochoa@ucsp.edu.pe (J. Ochoa-Luna)

ORCID: 0000-0003-2075-3379 (D. Palomino); 0000-0002-8979-3785 (J. Ochoa-Luna)

Nowadays, this kind of encoding has evolved to a language model encoding. The idea is to use language context in order to better encode words [8]. Overall the aim is to transfer the knowledge encoded in the language model to a specific task, in this case polarity detection in sentiment analysis.

Our proposed classifier relies on three components: a multivariant Spanish corpus, a general language model and a data augmentation step. Thus, we start by training a general language model based on BERT [9] using a multivariant Spanish corpora. Such language model allows us to learn general features of the Spanish language. The final prediction is enhanced using an unsupervised data augmentation process.

Those simple design choices allow us to obtain comparable state-of-the-art results regarding the two subtasks of polarity classification task presented in the 2020 TASS challenge [10] (This work is not covering the emotion detection task [11])

The rest of the paper is structured as follows. In Section 2, we describe the task at hand. In Section 3, the system is explained. The experimental setup is described in Section 4. Results and conclusions are presented, accordingly, in Section 5 and 6.

2. Task description

The aim of the task 1 is the correct classification of sentiments (N:Negative, P: Positive, NEU: Neutral) in tweets written in Spanish and variants. This year, the task has been sub-divided in 2 subtasks[1]:

- Subtask-1: Monovariant. In this challenge, we have to predict the sentiment in tweets of 5 Spanish variants included: Spain, Peru, Costa Rica, Uruguay and Mexico, and for each one we have three datasets: Training, Validation and Test. Moreover and even when we could use any additional corpora or linguistic resource, we have to submit the predictions of every country Test dataset in a different file to the system evaluation.
- Subtask-2: Multivariant. Complementary to the subtask-1, we have been provided with one additional Test dataset extracted from the different datasets of countries commented before. Again, it is possible to use any corpora or linguistic resource.

Furthermore, informal language used in tweets and lack of context because the limited characters permitted in those add up several challenges to this competition.

Finally, the metrics used in the evaluation system are the macro-averaged versions of Macro-Precision, Macro-Recall and Macro-F1. However, Macro-F1 was used to rank the different submits to the system.

[1] http://tass.sepln.org/2020/#tasks

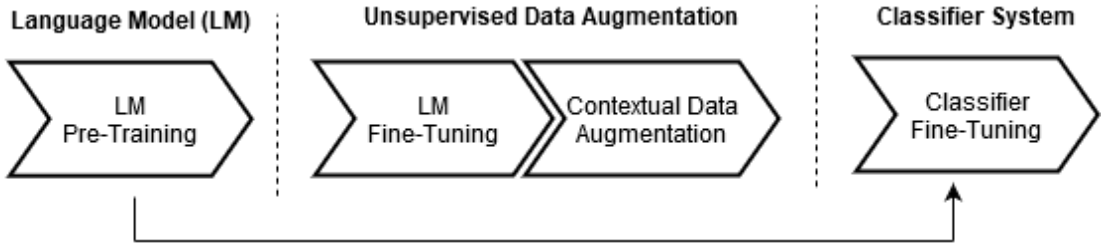

Figure 1: Stages of our system pipeline. From left to right, we start pre-training a general LM in order to use it later in Unsupervised Data Augmentation and Classifier System steps.

3. System

We propose a single system for improving polarity classification in small datasets based on the high performance Language Model (LM) referred to as BERT developed by Devlin et al [9]. In addition we also use a clever BERT variant to produce contextual data augmentation [12]. Overall, our aim is to develop a robust approach able to be used indistinctly in different variants of Spanish and get competitive results in several NLP tasks.

In order to do so, we start by pre-training a general LM on a multivariant Spanish corpus related to the target task. Subsequently, to enhance our small target dataset and prevent overfitting, we use the LM to generate Unsupervised Data Augmentation using a novel technique called Conditional BERT Contextual Augmentation [12]. Finally, this augmented dataset is used to fine-tune a classifier on top of previous LM. A general view of these stages is shown in Figure 1 and will be detailed as follows.

3.1. Language Model

Due to the huge impact of BERT in NLP, we decided to use it as our initial LM in its base form. By doing so, we have fewer parameters to train and it is easier to fine-tune with less samples. According the guidelines presented by Devil et al. [9], that describes how to train a Transformer Encoder for successful results, we use 2 objective tasks to pre-train our LM: Masked Language Model and Next Sentence Prediction.

- **Masked Language Model.** Mask some percentage of the input at random and then predict those masked tokens.
- **Next Sentence Prediction.** Given a first sentence, the model must predict the one that would follow.

Also, this process is made using a free source dataset described in spanish-corpora repository[2].

[2]https://github.com/josecannete/spanish-corpora

3.2. Unsupervised Data Augmentation

Based on the Conditional BERT Contextual Augmentation algorithm [12], we fine-tune the previous pre-trained LM using conditional MLM task on labeled target dataset. Once the LM is fine-tuned, we take one tweet in our train dataset and randomly mask k words for later predicting the masked words using the LM fine-tuned. This process is repeated for every tweet in our dataset n times, where k and n are numbers chosen heuristically.

Finally, we add up the new formed sentences into the original training dataset and perform downstream task on it.

3.3. Classifier System

In this final step we build a classifier, a simple fully connected layer, on top of the initial LM and a Softmax which allow us to predict the correct class. However, in order to prevent overfitting, we use the new enhanced dataset created using the aforementioned Semi-supervised Data Augmentation process.

4. Experimental Setup

A detailed description of hardware and software requirements for replicating our research results. In addition, we describe the hyper-parameters tuned during experimentation that allow us to train and converge without overfitting and regularization.

4.1. Technical Resources

The experiments were executed on Jupyter Notebooks running Python 3.7 kernel and PyTorch 1.4.0. Moreover, all models were trained on a GPU 2080 Ti with 11 GB GDDR6.

For a complete description about dependencies we refer to our public project repository[3].

4.2. Datasets

The datasets used in this work are freely available to everyone to use.

4.2.1. Pre-Training

The dataset used to pre-train the LM is a compilation of Large Spanish Unannotated Corpora [13] including Spanish Wikipedia, Spanish portion of TED conferences and other resources.

4.2.2. Target

The dataset used to fine-tune the LM of the unsupervised Data Augmentation and the final classifier is provided by the competition committee and can be accessed via the web of the challenge[4].

[3] https://github.com/dpalominop/atlas
[4] http://tass.sepln.org/2020

4.3. Pre-processing

All datasets were pre-processed regarding the following rules:

1. The text was converted to lowercase and every accent mark was removed.
2. Repeated characters were replaced by single characters.
3. User references were transformed to a specific token.
4. Useless spaces were removed.

4.4. Pre-Training Language Model

As will be shown in Table 3, using a multilingual LM decreases the performance of the system. In contrast, if we use a monolingual LM, results are noticeably improved. Furthermore, we have to use a LM trained on a similar dataset w.r.t. the target task and the best option was training one using the general corpora described in section 4.2. The main hyper-parameters used through this process are:

1. Backpropagation Trough Time (BPTT): 70
2. Weight Decay (WD): $1e-2$
3. The batch size (BS): 64

4.5. Fine-Tunning Language Model

The main hyper-parameters used through this fine-tuning LM process are:

1. Backpropagation Trough Time (BPTT): 70
2. Weight Decay (WD): $1e-2$
3. The batch size (BS): 32
4. Number of randomly masked words (k): 4
5. Repetitions of the process (n): 2

4.6. Fine-Tunning Classifier

Similar to the sub-section before, the main hyper-parameters used through this fine-tuning process are:

1. Backpropagation Trough Time (BPTT): 70
2. Weight Decay (WD): $1e-2$
3. The batch size (BS): 16

5. Results

Results for **TASS 2020 Task 1 - Monovariant** are shown in Table 1. Our submission (referred to as **daniel.palomino.paucar**) was ranked 1st or 2nd in all variants of Spanish presented in the competition among all competitors (M-F1 score).

Table 1
Top 3 results on TASS 2020 Task 1 - Monovariant test dataset (M-F1 Score).

Country	Team	M-F1
CR	jogonba2elirf	0.6463
	daniel.palomino.paucar	0.6462
	elchudi	0.4580
ES	jogonba2elirf	0.6711
	daniel.palomino.paucar	0.6645
	joseantonio.garcia8	0.5033
MX	jogonba2elirf	0.6344
	daniel.palomino.paucar	0.6332
	elchudi	0.5127
PE	jogonba2elirf	0.6355
	daniel.palomino.paucar	0.6335
	elchudi	0.4477
UY	**daniel.palomino.paucar**	0.6690
	jogonba2elirf	0.6547
	joseantonio.garcia8	0.5201

Table 2
Top 3 results on TASS 2020 Task 1 - Multivariant test dataset (M-F1 Score).

Team	M-F1	Precision	Recall
daniel.palomino.paucar	0.4979	0.4878	0.5095
elchudi	0.3302	0.3203	0.3408
joseantonio.garcia8	0.3578	0.3586	0.3570

Furthermore, results for **TASS 2020 Task 1 - Multivariant** are shown in Table 2. Our submission (referred to as **daniel.palomino.paucar**) was ranked 1st among all competitors (M-F1 score).

On the other hand, given the possibility to send 3 different submissions to the competition, we decided to use that to test three variations of our system and perform ablation experiments in order to get a better understanding of the competitive results obtained.

The output of these experiments is shown in Table 3. Thus, we can observe a significant positive impact when using a monolingual LM instead of a multilingual LM. Regarding the M-F1 metric used, there is 10 percent difference between the two methods. Moreover, if we include the additional step of Unsupervised Data Augmentation, the results further improve.

Table 3
Ablation using different pre-trained language models according to M-F1 metric. "w/" denotes "with".

	CR	ES	MX	PE	UY
Using multilingual LM					
Classifier Fine-Tuning	0.5235	0.5618	0.5457	0.5233	0.5556
Using monolingual LM					
Classifier Fine-Tuning LM	0.6378	0.6569	0.6332	0.6286	0.6595
w/Unsupervised Data Augmentation	**0.6462**	**0.6645**	**0.6198**	**0.6335**	**0.6690**

6. Conclusions

We have presented a novel sentiment classification system based on BERT that includes an additional step of Unsupervised Data Augmentation. The system has been applied to sentiment analysis on Spanish tweets and its variants. Despite its simplicity, this approach allowed us to be ranked 1st or 2nd on the **TASS 2020 Task 1 - Multivariant** and 1st on the **TASS 2020 Task 1 - Monovariant**.

Furthermore, the ablation experiments have shown that a careful choice of the language model can improve the results drastically. Thus, we have chosen to pre-train the language model using a similar dataset to the target task. Moreover, the use of data augmentation allowed us to further improve our previous results for most variants of the Spanish language. However, the performance on the Mexican variant decreased after using this technique —probably due to overfitting during the fine-tuning process.

Acknowledgments

This work was funded by CONCYTEC-FONDECYT under the call E041-01 [contract number 34-2018-FONDECYT-BM-IADT-SE].

References

[1] J. Villena-Román, J. García-Morera, C. Moreno-García, L. Ferrer-Ureña, S. Serrano, J. Gonzalez-Cristobal, A. Westerski, E. Martínez-Cámara, M. García-Cumbreras, M. Martín-Valdivia, L. López, Tass-workshop on sentiment analysis at sepln, 2012.

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

[3] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing

Systems 26, Curran Associates, Inc., 2013, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[5] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[6] 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. URL: https://www.aclweb.org/anthology/P18-1031.

[7] M. C. Díaz-Galiano, M. G. Vega, E. Casasola, L. Chiruzzo, M. Á. G. Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. M. Ráez, M. A. S. Cabezudo, E. S. Tellez, M. Graff, S. Miranda-Jiménez, Overview of tass 2019: One more further for the global spanish sentiment analysis corpus, in: IberLEF@SEPLN, 2019.

[8] D. Palomino, J. Ochoa-Luna, Advanced transfer learning approach for improving spanish sentiment analysis, in: L. Martínez-Villaseñor, I. Batyrshin, A. Marín-Hernández (Eds.), Advances in Soft Computing, Springer International Publishing, Cham, 2019, pp. 112–123.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018).

[10] M. García-Vega, M. C. Díaz-Galiano, M. A. García-Cumbreras, A. Montejo Ráez, S. M. Jiménez Zafra, E. Martínez-Cámara, C. A. Murillo, E. Casasola Murillo, L. Chiruzzo, D. Moctezuma, Sobrevilla, Overview of tass 2020: Introduction emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[11] F. M. Plaza del Arco, C. Strapparava, L. A. Urena Lopez, M. Martin, EmoEvent: A multilingual emotion corpus based on different events, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1492–1498. URL: https://www.aclweb.org/anthology/2020.lrec-1.186.

[12] X. Wu, S. Lv, L. Zang, J. Han, S. Hu, Conditional BERT contextual augmentation, CoRR abs/1812.06705 (2018). URL: http://arxiv.org/abs/1812.06705. arXiv:1812.06705.

[13] J. Cañete, Compilation of large spanish unannotated corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731. doi:10.5281/zenodo.3247731.