Classic and recent (neural) approaches to automatic text classification: a comparative study with e-mails in the Spanish language

Juan M. Fernandez[1,2], Nicolás Cavasin[3], and Marcelo Errecalde[4]

[1] Master Student at Computer Science School, La Plata National University
[2] Professor and Researcher at Luján National University
[3] Luján National University
[4] Professor and Researcher at LIDIC, San Luis National University
{jmfernandez, ncavasin}@unlu.edu.ar, merreca@unsl.edu.ar

Abstract. Currently, millions of data are generated daily and its exploitation and interpretation has become essential at every scope. However, most of this information is in textual format, lacking the structure and organisation of traditional databases, which represents an enormous challenge to overcome.
Over the course of time, different approaches have been proposed for text representation attempting to better capture the semantic of documents. They included classic information retrieval approaches (like Bag of Words) to new approaches based on neural networks such as basic word embeddings, deep learning architectures (LSTMs and CNNs), and contextualized embeddings based on attention mechanisms (Transformers). Unfortunately, most of the available resources supporting those technologies are English-centered.
In this work, using an e-mail-based study case, we measure the performance of the three most important machine learning approaches applied to the text classification, in order to verify if new arrivals enhance the results from the Spanish language classification models.

Keywords: Text Classification · SVM · Word2Vec · LSTM · BERT

1 Introduction

As a result of the massive access to the internet, millions and millions of data are daily generated and its exploitation and interpretation has become essential at every scope. Information retrieval and text mining became, along the years, the most popular investigation fields, specially in the text classification field [5]. Following this direction, papers about text classification can be found since 1957, where the research work only proposed text classification using the words frequency method [9]. Since then, diverse approaches have been developed for text representation and the knowledge creation using them as a data source. Nevertheless, most of the resources available are English-centered, leaving a reduced group of alternatives for the remaining languages. At the same

time, there are not many works with empirical comparisons measuring those new approaches' performance in languages like Spanish, where reliable resources are not frequently available for the implementation of those new text classification approaches. This work presents experiments comparing the performance of the three most relevant approaches of machine learning applied to text classification in order to measure how beneficial their contributions to non-English languages are.

2 Related work

As stated before, even though in the last 60 years several approaches for automatic text classification proliferated, there are not enough researches about the performance of those different strategies on non-Engligh languages. Below is a brief review of the three strategies used in the frame of this research for the study comparison.

#1: BoW+SVM One of the simplest methods for document representation, and also one of the oldest, is called Bag of Words *(BoW)* or vector space model [11]. This technique generates a vector that represents a document using the frequency count of each term inside the document [6] and is called that way because words are taken as features and documents like collections of unordered words [8]. This representation strategy has simplicity as advantage and also the possibility of applying any classification technique to the final representation. One of the most frequently used is the Support Vector Machine (SVM), created in the mid 1990s, which won popularity due to some attractive features and its empirical performance. SVM is based on the statistical learning theory principle Structural Risk Minimization (SRM), which consists in finding the optimal hyperplane that guarantees the smallest real error [7]. For the distances calculus and hyperplanes pursuit, SVM uses functions called kernels [12].

#2: Word2Vec+LSTM A fairly current line of research includes the usage of contextual information in conjunction with simple neural-network models to obtain words and phrases representations in the vectorial space [14]. One of the most popular models is Word2Vec, which has two different architectures namely CBow and Skip-gram [10]. These models of word embeddings are usually complemented with recurrent neural-networks as Long Short-Term Memory (*LSTM*). These neural-networks provide two new features that drastically improve the performance against conventional neural-networks for text processing: they are able to identify the order of text sequences in documents and process different length documents. [1].

#3: BERT As an evolution of the previous strategy, in 2017, a new neural-network architecture, called *Transformer* [13], arose simpler and parallelizable and is only based on attention mechanisms that completely avoid recurrencies and convolutions. They can be described as the assignment of a query and a set of key-value pairs to an output where the query, the keys and the values are all vectors. From this logic rises what in nowadays literature is known as the text representation models' actual state-of-art, called Bidirectional Encoder

Representations from Transformers *(BERT)* [4]. Briefly, this framework has two steps: initial pre-training and posterior fine-tuning. During the pre-training step, the model is trained with unlabelled data in different tasks. Then, during the fine-tuning step, the BERT model is first initialized with the pre-trained model's parameters and are finally adjusted using labelled data from posterior tasks.

3 Research methodology

This research work uses a dataset generated from academic questions made by e-mail by students of the National University of Luján to the administrative staff on topics related to academic activities. From a total of 24700 e-mails, 1000 interactions have been selected, labelled by a domain expert and assigned based on four classes (public transport discount ticket, admission to the university, admission requirements, other topics). For the experiments, the original e-mails were used without any human supervision on semantic nor syntactic mistakes. For the approach based on **BoW+SVM** the text was normalized removing stop-words, adding static attributes (such as question's length and punctuation marks usage) and using variations of *n-grams* and characters. On the other hand, for the approaches based on **Word2Vec+LSTM** and **BERT**, the text sequences related to the selected interactions were just normalized. Only for **Word2Vec+LSTM** the stop-words were removed due to the fact that **BERT** was experiencing a decrease in its performance when they were not there. Additionaly, Spanish pre-trained word embeddings[2] were used for **Word2Vec+LSTM**. Regarding **BERT**, two pre-trained models were used. One of them is Spanish-native [3] and the other, called *Multilingual* [4], was developed for several languages. For the evaluations, a cross validation approach was adopted with a *5-fold* on the 80% of the training instances while the models were *tested* with the remaining 20% of the instances using *accuracy*, *precision*, *recall* and *f1-score*.

4 Experimental results

In every approach a search for the best hyper-parameters was applied to each strategy, getting the following results[5]:

Table 1: Results of the different learning strategies.

Strategy	**Accuracy**	**Precision**	**Recall**	**F1-score**
BoW+SVM	0.870	0.862	0.830	0.840
Word2Vec+LSTM	0.835	0.814	0.841	0.820
BERT (Multilingual)	0.860	0.838	0.842	0.840
BERT (BETO)	**0.890**	**0.870**	**0.885**	**0.876**

[5] Experiments available at github.com/jumafernandez/clasificacion_correos

Results show **BERT** as the most effective approach for classifying this dataset, using the previously mentioned Spanish pre-trained model. Nevertheless, the difference with the resulting metrics regarding **BOW+SVM** were 0.03 or smaller.

5 Conclusions and future work

Based on the previous results, and considering the 30 years of evolution that this discipline has experienced since **BOW+SVM** first appearance and **BERT**'s presentation, we have verified that the traditional representation and classification methods are still a very competitive option.

However, it is important to keep in mind that e-mails in general, and this dataset in particular, have some features that do not help these cutting-edge models due to its informal manners and syntactic mistakes which are frequently seen in this type of communication model. That is why the precision gap between cutting-edge models and traditional ones is expected to maximize when datasets with cleaner texts are used. At the same time, and for our collection, this could be solved using spell-checkers to purge the documents during the pre-processing step.

References

1. Aggarwal, C.C., et al.: Neural networks and deep learning. Springer **10**, 978–3 (2018)
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (August 2019), https://crscardellino.github.io/SBWCE/
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Fanny, F., Muliono, Y., Tanzil, F.: A comparison of text classification methods k-nn, naïve bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT **3**(2), 157–160 (2018)
6. Harish, B.S., Guru, D.S., Manjunath, S.: Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2) pp. 110–119 (2010)
7. Islam, M.R., Chowdhury, M.U., Zhou, W.: An innovative spam filtering model based on support vector machine. In: CIMCA-IAWTIC'06. vol. 2, pp. 348–353. IEEE (2005)
8. Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K.: Fast text categorization using concise semantic analysis. Pattern Recognition Letters **32**(3), 441–448 (2011)
9. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development **1**(4), 309–317 (1957)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
12. Skiena, S.S.: The data science design manual. Springer (2017)

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
14. Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P., Witbrock, M.J.: Word mover's embedding: From word2vec to document embedding. arXiv preprint arXiv:1811.01713 (2018)

Acknowledgement

Authors are grateful to the Center of Research, Teaching and TIC Extension from the National University of Luján (CIDETIC) for providing the computational resources that allowed this project's experiments to be executed.