URJC-Team at FinancES 2023: Financial Targeted Sentiment Analysis in Spanish Combining Transformers

Miguel Ángel Rodríguez-García[1,*], Adrián Riaño-Martínez[1], David Roldán-Álvarez[1] and Soto Montalvo-Herranz[1]

[1]*Universidad Rey Juan Carlos, Spain*

Abstract

Financial and economic news is continuously monitored due to their impact on future stock prices. Thus, the polarity extraction from the news is a very relevant task for investment decision-making by traders. In this sense, Sentiment Analysis models can provide accurate methods to extract signals that influence this decision-making. In this work, we describe the contribution to IberLEF 2023 Challenge - FinancES, where we proposed a hybrid approach that addresses the targeted sentiment analysis by creating a pipeline with different phases. First, a phase for cleaning texts, followed by an entity recognition phase and, finally, the polarity extraction. The hybrid approach combines different models to proceed with each phase. Thus, RoBERTa transformer architecture is employed as a NER, BETO transformer model is employed for polarity analysis, and, finally, a Spanish Spacy model is used for the part-of-speech tagging process. Although the proposed approach has still scope for improvement, since it reached mid-table positions in the leaderboard, it put forwards a different method to carry out the proposed classification tasks.

Keywords

Deep Learning, Transformers, Natural Language Processing, Name Entity Recognition, Targeted Sentiment Analysis

1. Introduction

Financial markets are changeable and sensitive to global events and phenomena [1]. Factors like political news to users' opinions can produce an immediate effect directly on the market, making it grow positively if the news is good or, conversely, pull downwards if it is bad [2]. Therefore, understanding the emotions embedded in these resources can assist financial professionals and economics in predicting stock market fluctuations [3].

In this sense, inside the Natural Language Processing (NLP) research area, Sentiment Analysis is a sub-field that includes a range of computational methods to extract the subjectivity

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

†These authors contributed equally.

miguel.rodriguez@urjc.es (M. Á. Rodríguez-García); a.riano.2016@alumnos.urjc.es (A. Riaño-Martínez); david.roldan@urjc.es (D. Roldán-Álvarez); soto.montalvo@urjc.es (S. Montalvo-Herranz)

0000-0001-6244-6532 (M. Á. Rodríguez-García); 0009-0004-8755-255X (A. Riaño-Martínez); 0000-0001-8158-7939 (S. Montalvo-Herranz)

information by recognising the emotion, opinion or polarity in human language [4]. This type of analysis is becoming an essential tool in various domains to transform emotions and attitudes into actionable knowledge that assists in decision-making processes [5, 6]. In this work, the proposed challenge is focused on the financial domain [7]. The IberLEF FinancES shared task go beyond extracting the polarity of text and require recognising the financial target entity affected by this opinion [8]. To address these tasks, we propose a Deep Learning based system that combines three different models to carry out polarity analysis, entity recognition, and grammar tagging. After analysing various language models, we utilised RoBERTa (https://huggingface.co/MMG/xlm-roberta-large-ner-spanish) for named entity recognition, a Spacy model for Spanish is employed for identifying the grammatical category and, finally, we selected BETO (https://huggingface.co/finiteautomata/beto-sentiment-analysis) as a polarity extractor.

The rest of the paper is organised as follows. Section 2 presents related work, where various approaches that address Opinion Mining in the financial domain are analysed. Section 3 details the distribution of the dataset delivered for each task and describes the systems' architecture proposed. Section 4 analyses the results achieved in the challenge. Finally, Section 5 summarises the findings harvested facing the challenge, and point out various future research line to explore.

2. Related work

Sentiment Analysis is growing and taking an important role in better understanding users' opinions in several domains [9]. This growing importance is being applied to finance, since several studies have demonstrated to find strong correlations between textual sentiment and other financial measures, such as stock returns and volatilities [10]. Following this line of work, Xiang et al., in [11], address the sentiment analysis in the financial domain trying to predict the sentiment intensity of a determined target in a text. They proposed a semantic and syntactic enhanced neural model, called (SSENM), that constructs sentiment-aware representations considering target information, semantic features and syntactic knowledge for modelling more precisely the correlation between target mentions in the text and sentiment-relevant keywords. Addressing the same working task, Shang et al., in [3] propose LECN, a novel Lexicon Enhanced Collaborative Network to capture associations between financial targets and sentiment signals. The model is mainly based on three components: the shared Encoder Layer, which receives input sentences encoded into word embeddings by using BERT and is based on a BiLSTM architecture; the Task-specific Attention layer responsible for driving the model to focus on text segments linked to sentiments and targets; and finally, the message selective-passing mechanism in charge of adding features information gathered from previous interactions from the target extraction and sentiment analysis tasks to control the information shared between both tasks and enhance the collaborative effect. A different approach that faces the same classification problem is the work proposed by Shijia et al. [12], where they proposed a neural network architecture composed of a stack of LSTM layers and it receives as inputs word vectors encoded by the word2vec model.

In this analysis, we have selected a set of proposals that face the same classification problem thrown in the IberLEF challenge. Various architectures have been analysed, from neural networks to transformers architectures. Given the outcomes achieved by the proposals, we

addressed the challenge by using transformers, specifically, language models, since they seem to reach more promising results.

3. Material and methods

This section describes the datasets proposed by the organizers and the model developed to deal with the tasks delivered in the challenge.

3.1. Data

The FinancES dataset is constituted of Spanish news headlines harvested from digital newspapers specialized in the targeted domain, such as Expansión, El Economista, Modaes and El Financiero [13]. The resulting dataset was about 14k headlines, manually labelled by three individuals, selecting the target entity and the sentiment polarity on three dimensions: target, companies, and consumers. During this selection, some headlines were discarded cause of their short length and controversy during the labelling process. As a result, the dataset was reduced to 6k-8k. Table 1 shows the distribution of the datasets released for the practising and evaluation phase.

		Practice		Evaluation	
	label	Train	Test	Train	Test
target_sentiment	positive	305	109	2815	816
	neutral	60	9	606	205
	negative	370	51	2935	600
Total		735	169	6356	1621
companies_sentiment	positive	252	39	645	523
	neutral	349	81	3841	822
	negative	134	49	1870	276
Total		735	169	6356	1621
consumers_sentiment	positive	167	42	895	553
	neutral	392	89	4170	803
	negative	173	38	1289	265
Total		732	169	6354	1621

Table 1
Distribution of the released datasets.

As we can see in Table 1, each label is separated into the three common values assigned in sentiment analysing tasks: positive, neutral and negative, to depict a more detailed picture of the dataset. This organization shows some irregularities in the datasets of practice and evaluation, where the total count of the 'targets_sentiment' and 'companies_sentiment' does not match the 'consumers_sentiment', differing in two and three examples. This inconsistency is because these samples were labelled differently, with a label not included in these values. Apart from

this, it is noteworthy that the clear unbalance of neutral cases in 'target_sentiment' for practice and evaluation datasets has a highly unequal distribution of examples.

Another characteristic that draws attention is the similar distribution in 'companies_sentiment', where the neutral label differs to others in a large number of samples.

3.2. Method

The FinancES challenge of the IberLEF 2023 evaluation campaign consisted of two main complex tasks detecting the main economic target and its linked sentiment and characterizing the polarity at the document level for companies and consumers. Figure 1 shows the architecture that was proposed to address both these two main tasks.

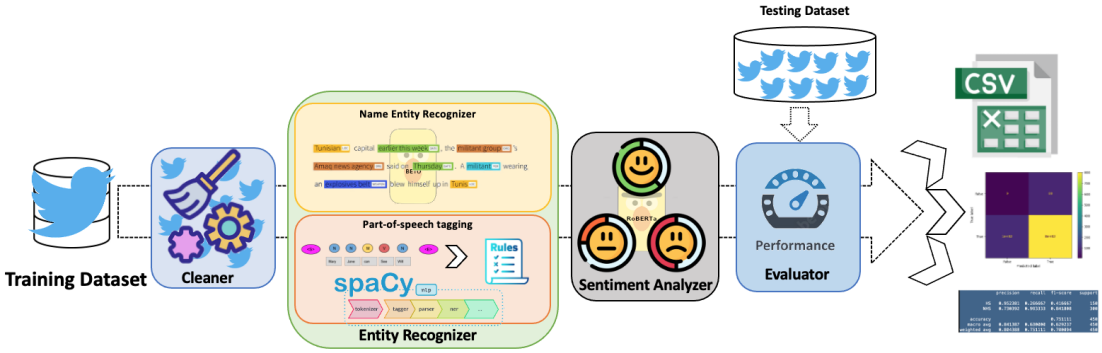


Figure 1: Architecture of the proposed system.

As we can see in Figure 1, the proposed system is configured as a pipeline, where the inputs are the news headlines. First, the cleaner module is responsible for unifying the format of the sentences, removing links, emojis and lowercase words. Next, the target extractor aims at identifying potential financial targets. It receives the pre-processed text and employs a pre-trained model to identify named entities specifically, it operates the transformer-based language model RoBERTa. If the model does not identify entities, we came up with a second extractor method, which was based on the premise that the sentence structure of news headlines is not too complex and follows a simple composition like DET+NOUN+VERB. In this sense, we use grammatical tagging to identify potential targets. Concretely, we employ the spaCy pipeline for NER to mark up each word of each sentence to a particular part of speech. Then, we created a set of 14 simple grammar rules to extract potential targets by combining in different ways grammatical tags like DET, NOUN, ADJ, and VERB, among others. To decide how to define these grammar rules, we accomplished a statistical analysis of the headlines given to analyse what were the most used grammar structures by the writers. These rules are employed specifically from the start of the sentence until the word tagged as a VERB is located since we assume the target entity will appear in the first part of the sentence. These are the 14 grammar rules designed are shown in Table 2. When a news headline follows this structure, the words that match the rule are extracted systematically. For instance, if we have the following headline: "Las empresas chinas piden menos burocracia para invertir en España", the set of words that

will be extracted is: “Las empresas chinas”. On the Sentiment Analysis task, the sentiment analyser module tackles each task differently. The polarity of “target_sentiment” in the sentence is obtained by employing the transformer model BETO, which, for each sentence, it returns two values, the label corresponding to the sentiment that could be NEU or POS and NEG referencing the three existing types of emotional tones, and its score. Thus, to assign this tone, we have defined a threshold, which was established based on several experiments conducted. Therefore, depending on the label and score inferred by the model, and if this score defeats the threshold established, the polarity is assigned in this task, otherwise, it is left empty.

Index	Rules	Index	Rules
#1	DET+NOUN	#2	DET+NOUN+ADJ
#3	DET+NOUN+NOUN	#4	DET+NOUN+NOUN+ADJ
#5	NOUN+ADP+NOUN	#6	NOUN+ADP+NOUN+ADJ
#7	NOUN+VERB	#8	NOUN+ADJ+VERB
#9	NOUN+ADP+ADJ+VERB	#10	NOUN+NOUN+VERB
#11	NOUN+AUX	#12	NOUN+NOUN+AUX
#13	PROPN+VERB	#14	PROPN+ADJ

Table 2
Grammar rules.

For companies and consumers, the sentiment analyzer module needs first that entities are classified into a person or organization. Two techniques are used in this classification: using the labels inferred by the pre-trained Name Entity Recognition model or two dictionaries of words that group nouns related to persons and organizations domains. When the entities are classified, the sentiment analyzer module analyses the context where the entity is placed and extracts the polarity of text equally to the ‘target_sentiment’ task described above.

4. Results and discussion

In this section, we analyse the performance of the architecture proposed in the evaluation dataset delivered. It is worth stressing that, to facilitate the experiments’ reproducibility, we utilised the default parameters in pre-trained transformed models in the experiments carried out. The standardised metrics selected to assess the performance were Precision, Recall and F1-measure. Table 3 collects the results achieved on each challenge’s classification task.

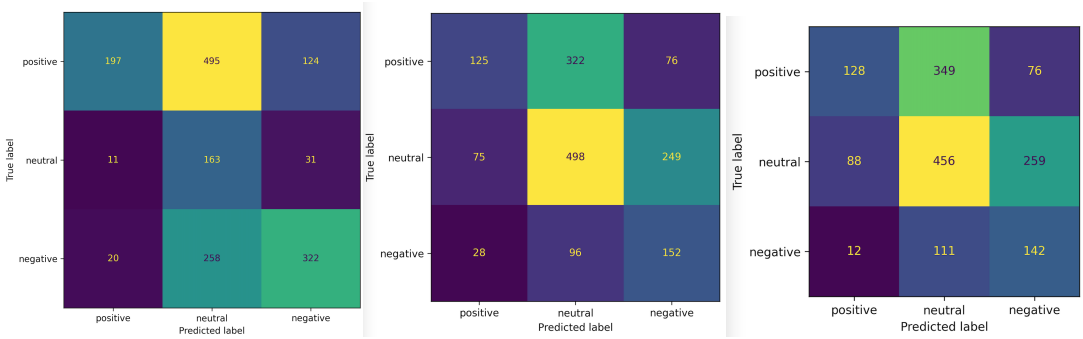
The complete classification report of the target extraction was not included since it would need a large table to add all the terms extracted during the evaluation. However, the accuracy reached by the system was 0.61 (see in Table 4), indicating that the combination of the RoBERTa model as NER with the designed grammatical rules obtained admissible outcomes. However, despite this reasonable result, the remaining subtasks got weak scores such as 0.37, 0.47 and 0.48 in ‘target_sentiment’, ‘companies_sentiment’ and ‘consumer_sentiment’, respectively. We think this low score in extracting the target is due to there are specific expressions in the news headlines that grammar rules do not cover, making the extractor loses some entities during the evaluation. Concerning the polarity detection, the task where the proposed system accomplished the best outcome was in ‘target_sentiment’, harvesting the highest value on precision and recall,

		Evaluation		
Task	Label	Precision	Recall	F1-score
target_sentiment	negative	0.62	0.69	0.65
	neutral	0.1	1	0.17
	positive	1	0.17	0.3
	Macro AVG	0.57	0.62	0.37
companies_sentiment	negative	0.53	0.61	0.57
	neutral	0.54	0.63	0.58
	positive	0.42	0.2	0.27
	Macro AVG	0.5	0.48	0.47
consumer_sentiment	negative	0.44	0.66	0.53
	neutral	0.55	0.58	0.57
	positive	0.53	0.24	0.33
	Macro AVG	0.51	0.49	0.48

Table 3
Results obtained on the official test set.

on positive and neutral labelling, respectively. The worst results coincided with the task and label, obtaining 0.1 and 0.17 on precision and recall. In spite of the overall results displaying balanced values, which vary between 0.5 and 0.4, if we focus on the positive labelling results, it is easy to recognize that they have a negative impact on the system's performance since recall outcomes drop below 0.3. However, if we look at the dataset's distribution, the number of samples for the positives across the tasks does not reflect the outcomes obtained. In some tasks, the number of positive examples is higher, but the system's performance is quite low. For instance, in the "target_sentiment" task, in the positive labelling the system reached a recall of 0.17, but it has assigned the highest amount of samples in the dataset. Consequently, this situation reflects that the Transformer model selected for addressing the sentiment analysis classification problem does not work precisely. We think this behaviour might be reduced by using augmentation techniques, which sampling generation could teach better the model to differentiate between the three types of classes, positive, negative and neutral. For a more detailed study of these low results, Figure 2 shows the confusion matrix of each classification task.

Confusion matrices allow an analysis deeply of the results. In this case, it is easy quantifiable the mistakes conducted by the system during the classification process. As we can see, the noticeable mistakes were positive to neutral and neutral to negative classifications. For instance, on the task 'targets_sentiment', the system misclassified 495 samples, and 259 on the 'consumers_sentiment'. From here, it can be inferred that the system had grave problems distinguishing positive and neutral polarities since it hesitates and makes significant mistakes. The mistakes in predicting neutral cases are elevated if we examine and compare the three confusion matrices. We believe that this behaviour is due to the value assigned to the threshold for classifying neutral sentences was not precise enough since it seems that a high percentage of the cases are classified as neutral, but they are not. It is worth mentioning that in positive versus



(a) Target Sentiment Task. (b) Companies Sentiment Task. (c) Consumers Sentiment Task

Figure 2: Confusion matrix computed on each task.

negative classifications in the pictures' corner, the number of errors is less inflated, reducing misclassified cases. Thus, despite the problems in classification-neutral samples, we think the system knows how to differentiate acceptably between positive and negative polarity. Despite the several experiments conducted to find the best system's performance, we can infer from these results that the pre-established value of the threshold does not work correctly for any case. Thus, we think each polarity tone has to be addressed differently, assigning an individual limit for each type. These analysed issues have made the proposed system only achieve the 7^{th} place in the leaderboard, depicted in Table 4.

Ranking	Group Name	Target	F1-score Target Sentiment	F1-score Companies Sentiment	F1-score Consumers Sentiment
1	abc111	0.88	0.71	0.53	0.62
2	lli-uam	0.85	0.73	0.59	0.69
3	thindang	0.85	0.71	0.59	0.63
	...	...	...	...	...
7	NLP_URJC	0.61	0.42	0.44	0.41

Table 4
A list of approaches ordered by the results achieved in the four subtasks.

5. Conclusions

In this work we describe the contribution to the FinancES challenge, allocated in the IberLEF 2023 shared evaluation campaign of Natural Language Processing systems. The proposal combines several Deep Learning models to address the main subtasks in this challenge, target extraction and sentiment analysis. Thus, the system combined different models: BETO for classifying the polarity, RoBERTa for extracting the target, and a Spacy model for grammar labelling.

As we can see in the evaluation section, the gathered results show that there is still a range

of improvement. In the study of the resulting confusion matrices, it can be easily observed that our system can "easily" detect negative labels in any of the proposed tasks but works worse in detecting positive labels in any or neutral sentences on the 'targets_sentiment' issue. To improve this situation, we suggest the addition of new grammar rules for detecting more structures in sentences. On the other hand, as a future work line, we would like to try cutting-edge Deep Learning strategies that do not require extensive weight training processes, like prompt engineering.

6. Acknowledgments

This work has been partially supported by projects DOTT-HEALTH (PID2019-106942RB-C32, MCI/AEI/FEDER, UE), grant "Programa para la Recualificación del Sistema Universitario Español 2021-2023", and the project M2297 from call 2022 for impulse projects funded by Rey Juan Carlos University.

References

[1] I. Almalis, E. Kouloumpris, I. Vlahavas, Sector-level sentiment analysis with deep learning, Knowledge-Based Systems 258 (2022) 109954.

[2] B. Agarwal, Financial sentiment analysis model utilizing knowledge-base and domain-specific representation, Multimedia Tools and Applications 82 (2023) 8899–8920.

[3] L. Shang, H. Xi, J. Hua, H. Tang, J. Zhou, A lexicon enhanced collaborative network for targeted financial sentiment analysis, Information Processing & Management 60 (2023) 103187.

[4] K. Naithani, Y. P. Raiwani, Realization of natural language processing and machine learning approaches for text-based sentiment analysis, Expert Systems (2022) e13114.

[5] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, F. Amenta, Text mining with sentiment analysis on seafarers' medical documents, International Journal of Information Management Data Insights 1 (2021) 100005.

[6] S. Žitnik, N. Blagus, M. Bajec, Target-level sentiment analysis for news articles, Knowledge-Based Systems 249 (2022) 108939.

[7] J. A. García-Díaz, A. Almela, F. García-Sánchez, G. Alcaráz Mármol, M. J. Marín-Pérez, R. Valencia-García, Overview of FinancES 2023: Financial Targeted Sentiment Analysis in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[8] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[9] S. Minaee, E. Azimi, A. Abdolrashidi, Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models, arXiv preprint arXiv:1904.04206 (2019).

[10] T. Renault, Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages, Digital Finance 2 (2020) 1–13.

[11] C. Xiang, J. Zhang, F. Li, H. Fei, D. Ji, A semantic and syntactic enhanced neural model for financial sentiment analysis, Information Processing & Management 59 (2022) 102943.
[12] E. Shijia, L. Yang, M. Zhang, Y. Xiang, Aspect-based financial sentiment analysis with deep neural networks., in: WWW (Companion Volume), 2018, pp. 1951–1954.
[13] P. Ronghao, J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in spanish, PeerJ Computer Science 9 (2023) e1377. URL: https://doi.org/10.7717/peerj-cs.1377. doi:10.7717/peerj-cs.1377.