SINAI at FinancES@IberLEF2023: Evaluating Popular Tools and Transformers Models for Financial Target Detection and Sentiment Analysis

Salud María Jiménez-Zafra[1,*], Daniel García-Baena[1], Miguel Ángel García-Cumbreras[1] and Manuel García-Vega[1]

[1]*Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain*

Abstract

This work presents the participation of the SINAI team at FinancES@IberLEF2023 shared task, Financial Targeted Sentiment Analysis in Spanish. We have addressed the two proposed tasks, consisting on identifying the main economic target from headlines of financial news for determining their sentiment polarity and identifying the sentiment polarity of each news headline towards both companies and consumers. For target detection, we have explored some popular tools as Stanza and spaCy, and different transformers models from Hugging Face and ChatGPT4. For sentiment analysis, we have evaluated some of the most popular transformers models and specific financial transformers. In total, 11 systems have participated (including the baseline provided by the organizers). The best run sent by our team have been placed in position 4th for Task1 and position 2nd for Task 2 with an F1-score of 0.7780 and 0.6349, respectively, being 0.7922 and 0.6423 the best results obtained in the competition for both tasks.

Keywords

financial target detection, sentiment analysis, financial multi-dimensional sentiment classification, machine translation, transformers, natural language processing

1. Introduction

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages [1]. In an annual cycle that starts in December (with the call for task proposals) and ends in September (with an IberLEF meeting collocated with SEPLN), several challenges are run with large international participation from research groups in academia and industry. Specifically, this shared task was titled FinancES: Financial Targeted Sentiment Analysis in Spanish [2], and aims to explore targeted sentiment analysis in the financial domain for target detection and multi-dimensional sentiment classification.

This is an important shared task because being able to manage financial data has come into the spotlight [3]. While in the past this type of data was stored in big warehouses from banks and financial companies, after the invention of the Web, which facilitated access to all types of

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

sjzafra@ujaen.es (S. M. Jiménez-Zafra); daniel.gbaena@gmail.com (D. García-Baena); magc@ujaen.es (M. García-Cumbreras); mgarcia@ujaen.es (M. García-Vega)

0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0002-3334-8447 (D. García-Baena); 0000-0003-1867-9587 (M. García-Cumbreras); 0000-0003-2850-4940 (M. García-Vega)

information worldwide, including, of course, financial literacy, more people started to manifest their interest in economics. Nowadays, it is much easier to find financial information posted online and, therefore, it is possible to monitor public information, receive early warnings and perform positive and negative impact analysis. The effects of emotions on financial markets have been demonstrated in several studies [4, 5]. In any case, there are many different factors to take into account when we try to evaluate the effectiveness of sentiment analysis when we work in a financial context. In the first place, some complex vocabulary frequently populates economic texts, underlying social, and legal context [6]. On the other hand, in this domain, language is more related to circumstances and every word or expression may have either positive or negative connotations depending on the context and subjectivity is always present because texts are written according to the point of view, and/or the interests, of the author.

Specifically, our team has participated in both of the tasks of this shared competition. For Task 1, and with the objective of identifying the main economic target from texts, we evaluated some strategies based on different transformers, ChatGPT4 [7], the Python natural language analysis package Stanza [8], and the popular NLP tool spaCy [9]. In relation to Task 2, our team classified texts from headlines determining their sentiment polarity with the aid of, again, transformers-based models from Hugging Face.

Finally, this paper is divided into six different sections. Immediately after this introduction, we will talk about everything related with the task description, detailing its purpose and reviewing the dataset. Methodology section will explain what we did for generating our results and in Experimental setup, we discuss about the tools and how they were set up for conducting the experiments. On the other hand, Results and discussion summarizes the results that we obtained in each experiment and reviews them in a comparative way. The last section of this work is Conclusions and future work, and there we discern about our outcomes and, consequently, point to future works that would offer better results.

2. Task description

This shared task comprises two subtasks. The first one is for target detection and here, participants have to identify the economic target in newspaper headlines. After identifying the main economic target from headlines of financial news, teams have to classify the sentiment polarity (positive, neutral or negative) towards such target in the processed text. For the evaluation, the systems are ranked using the arithmetic mean of the target F1-score and sentiment classification macro-F1. Regarding the second task, participants are expected to conduct a multi-dimensional sentiment classification. As opposed to traditional multi-target tasks, in which multiple targets are identified within the scope of each individual processed text, here each news headline refers to a single target entity, but the stances of other economic agents, companies (com) and consumers (con), are also considered. For Task 2, the systems are ranked using the arithmetic mean of the macro-F1-com and macro-F1-con.

In relation to the dataset [10], it is an extension of the dataset published in [11]. It is composed of news headlines written in Spanish collected from digital newspapers specialized in economic, financial and political news as: Expansión, El Economista, Modaes or El Financiero. It is important to highlight that no all the newspapers are from the same Spanish-speaking

Table 1
Dataset distribution

Set	Sentiment	target	companies	consumer	Total headlines
dev-train	positive	370	134	173	737
	negative	305	252	167	
	neutral	60	349	392	
dev-test	positive	109	39	42	170
	negative	51	49	38	
	neutral	9	81	89	
train	positive	2,231	519	730	5,087
	negative	2,373	1,518	1,039	
	neutral	483	3,050	3,318	
test	positive	816	523	553	1,624
	negative	600	822	265	
	neutral	205	276	803	

country. The creators of the dataset reviewed all headlines removing the irrelevant ones and manually labelled each headline with the target entity and the sentiment polarity on three dimensions: target, companies, and consumers. Three options are available for sentiment analysis: positive, neutral, o negative. The final dataset comprises of 7,618 news headlines. For the shared task, at a first stage, development-training and development-test sets were made available for participants to develop their systems. Later, training set (including the data of the development set) and test sets were released to participate in the shared task. In Table 1, it is presented the dataset distribution. Finally, it is worth mentioning that the competition was organized through CodaLab and can be accessed at the following link: https://codalab.lisn.upsaclay.fr/competitions/10052#learn_the_details.

3. Methodology

We evaluated several technologies for tasks 1 and 2. For the first part of Task 1, consisting in identifying the main target of each headline in the dataset, we tested three different approaches, one based on some popular tools, such as Stanza [8] and spaCy [9], in this last case using the Spanish pipeline optimized for CPU model: es_core_news_lg, other on different transformer models from Hugging Face and the last one using the popular large language model ChatGPT4 [12]. ChatGPT4 was tested with different prompts and the one that worked best for extracting the target entities was: "Dime cuál es la entidad objetivo en la siguiente oración, sólo la entidad, sin punto final, manteniendo su forma de aparición en el texto"/*Tell me what is the target entity in the following sentence, just the entity, without a period, keeping its form of appearance in the text.* All the results obtained during the development phase are shown in Table 2. As can be seen, transformers-based models from Hugging Face obtained the best results over the rest of the options. Concretely, Babelscape/wikineural-multilingual-ner model obtained the best F1-score with a value of 0.8005.

Table 2
Rank list for target detection in the development phase

Model	F1-score
Babelscape/wikineural-multilingual-ner	0.8005
mrm8488/bert-spanish-cased-finetuned-ner	0.7983
ChatGPT4	0.5979
Stanza	0.5104
spaCy (es_core_news_lg model)	0.4629

Table 3
Sentiment analysis results for main targets, consumers and companies in the development phase

Model	Pretrained language	F1 target sentiment	F1 companies sentiment	F1 consumers sentiment
BERT	English	0.7530	0.5495	0.6422
BERTweet	English	0.7484	0.5876	0.6515
BETO	Spanish	0.7472	0.5362	0.6596
distilrobertafinancial	English	0.7408	0.5631	0.6315
dunnbc22	English	0.7272	0.5164	0.6210
FinancialBERT	English	0.7136	0.5102	0.5858
MarIA	Spanish	0.7557	**0.6050**	**0.6968**
mDeBERTa	Spanish	**0.7576**	0.5844	0.6820
ROBERTA	English	0.7539	0.5895	0.6793
RoBERTuito	Spanish	**0.7576**	0.5844	0.6820
Sigma	English	0.7263	0.4886	0.5725

In relation to sentiment analysis, we tested different transformer-based models with the original Spanish dataset and, for the English models, with an English-translated version of the original Spanish corpus using Google Translator from Python's deep_translator library [13]. After checking more than ten alternative configurations, based on finance-related and popular transformers models, for the second part of Task 1, related to sentiment analysis on the main target of news headlines, we obtained the best results with mDeBERTa and RoBERTuito and the original, non translated, Spanish dataset. On the other hand, for Task 2, on identifying sentiments for companies and consumers, the best results were obtained using MarIA and the original Spanish dataset. All F1 scores obtained in the development phase can be consulted in Table 3.

4. Experimental setup

Regarding to the software we used to translate the dataset from Spanish to English, it was Google Translator from Python's deep_translator library [13]. In addition, it is important to note that we did not perform any prior data pre-processing on the dataset to perform the experiments.

With respect to the models, all were downloaded from their public profiles in Hugging

Table 4
Best hyperparameter selection for sentiment analysis for main targets

Model	learning rate	num train epochs	per device train batch size	warmup steps	weight decay
BERT	3.8e-05	3	16	0	0.29
BERTweet	1.5e-05	1	8	1000	0.21
BETO	2.1e-05	1	16	0	0.24
distilrobertafinancial	2.5e-05	1	8	1000	0.29
dunnbc22	2.1e-05	3	16	0	0.085
FinancialBERT	1.5e-05	2	8	0	0.081
MarIA	1.3e-05	4	16	250	0.029
mDeBERTa	2.3e-05	5	16	500	0.15
ROBERTA	3.8e-05	5	16	0	0.12
RoBERTuito	4.6e-05	3	8	0	0.12
Sigma	4.5e-05	2	16	250	0.15

Table 5
Best hyperparameter selection for sentiment analysis for companies

Model	learning rate	num train epochs	per device train batch size	warmup steps	weight decay
BERT	1.6e-05	2	8	0	0.25
BERTweet	3.9e-05	4	16	0	0.022
BETO	4.5e-05	2	8	0	0.12
distilrobertafinancial	3.8e-05	4	8	500	0.27
dunnbc22	2.9e-05	4	16	0	0.26
FinancialBERT	4.2e-05	3	8	250	0.17
MarIA	3.8e-05	5	8	250	0.27
mDeBERTa	3.1e-05	4	8	250	0.2
ROBERTA	1.2e-05	3	8	0	0.039
RoBERTuito	2.3e-05	4	16	0	0.15
Sigma	4.8e-05	4	16	500	0.064

Face. During the finetuning process we always used Google Colab for coding under a Pro configuration for being able to use their GPU based hardware options.

Finally, concerning the hyperparameters, Table 4, Table 5 and Table 6 show the configurations that provided the best results for each of the models in the tasks sentiment analysis for main targets, sentiment analysis for companies and sentiment analysis for consumers, respectively. For target detection task, we used default parameters.

5. Results and discussion

This section presents the results obtained in the evaluation phase of the shared task FinancES [2], Financial Targeted Sentiment Analysis in Spanish, at IberLEF 2023. The organizers selected the arithmetic mean of the target F1-score and the target sentiment F1-score for ranking the systems

Table 6
Best hyperparameter selection for sentiment analysis for consumers

Model	learning rate	num train epochs	per device train batch size	warmup steps	weight decay
BERT	1.8e-05	3	500	0.21	0.25
BERTweet	1e-05	2	250	0.22	0.022
BETO	2.3e-05	3	8	0	0.16
distilrobertafinancial	2.8e-05	1	250	0.033	0.27
dunnbc22	4.4e-05	4	250	0.067	0.26
FinancialBERT	4.3e-05	2	250	0.014	0.17
MarIA	2.4e-05	5	8	250	0.22
mDeBERTa	2.7e-05	3	16	0	0.058
ROBERTA	2.1e-05	3	500	0.23	0.039
RoBERTuito	4e-05	1	8	0	0.29
Sigma	2.5e-05	2	0	0.11	0.064

Table 7
Results for Task 1 and Task 2 in the evaluation phase

run	avg. macro f1	f1 task 1	f1 target	f1 target sentiment	f1 task 2	f1 com sentiment	f1 con sentiment
run_1	0.5652	0.5973	0.5485	0.6461	0.5331	0.5017	0.5645
run_2	0.5479	0.5558	0.4465	0.6650	0.5400	0.5105	0.5695
run_3	0.6298	0.6369	0.5485	**0.7253**	0.6226	**0.5864**	0.6588
run_4	0.5902	0.6314	0.5485	0.7143	0.5491	0.5321	0.5660
run_5	0.6464	0.6579	0.5979	0.7178	**0.6349**	0.5835	0.6863
run_6	0.7061	0.7773	0.8368	0.7178	**0.6349**	0.5835	0.6863
run_7	**0.7065**	**0.7780**	**0.8382**	0.7178	**0.6349**	0.5835	**0.6863**
run_8	0.6756	0.7672	**0.8382**	0.6963	0.5841	0.5212	0.6469
run_9	0.6924	0.7729	**0.8382**	0.7077	0.6118	0.5637	0.6598
run_10	0.6611	0.7549	**0.8382**	0.6717	0.5672	0.5625	0.5719

in *Task 1: Financial targeted sentiment analysis*, and the arithmetic mean of the macro-F1-com and macro-F1-con for ranking the systems in *Task 2: Financial Sentiment Analysis at document level for companies and consumers*. Each participating team could submit a maximum of 10 runs through CodaLab, from which each team had to select the best one for ranking. We selected our 10 runs based on the experiments carried out on the training phase. The results for each of the runs are shown in Table 7 and the models used for target detection and sentiment classification in each of them are displayed in Table 8. The best results for each of the measures are marked in bold and the run that provided the best performance is highlighted with a gray background.

For the first part of Task 1, concerning the identification of the main economic target from financial news headlines, the transformer-based models (Babelscape/wikineural-multilingual-ner and mrm8488/bert-spanish-cased-finetuned-ner) outperformed ChatGPT4, Stanza and the es_core_news_lg model of spaCy. It is in this task where we appreciate the biggest differences between our systems and the best results are always for NER specific models that were developed

Table 8
Model target and model sentiment combination for each run

run	model target	model sentiment
run_1	Stanza	FinancialBERT
run_2	spaCy (es_core_news_sm model)	Sigma
run_3	Stanza	MarIA
run_4	Stanza	RoBERTuito
run_5	ChatGPT4	mDeBERTa
run_6	Babelscape/wikineural-multilingual-ner	mDeBERTa
run_7	**mrm8488/bert-spanish-cased-finetuned-ner**	**mDeBERTa**
run_8	mrm8488/bert-spanish-cased-finetuned-ner	BETO
run_9	mrm8488/bert-spanish-cased-finetuned-ner	ROBERTA
run_10	mrm8488/bert-spanish-cased-finetuned-ner	distilrobertafinancial

using some transformer approach.

Continuing with the first task, but now in relation to the second part consisting of determining the sentiment (positive, neutral or negative) towards the main target in the news headlines, we can see that the models using the original Spanish dataset performed better than those using the translated corpus, with MarIA being the best performing model. It should be noted that the English model ROBERTA works slightly better than the Spanish model BETO.

Regarding the second task, on determining the sentiment polarity of each news headline towards both companies and consumers, again the Spanish models overall obtain better results than the English models. However, it is worth noting that the English models ROBERTA and distilrobertafinancial work better than the Spanish models RoBERTuito and BETO. On this occasion, MarIA and mDeBERTa are the transformers that provide the best results for consumers and companies sentiment classification, respectively.

Finally, highlight that the finance-specific transformers have performed worse than the general transformers in the subtasks related to sentiment analysis.

For the competition, we selected run 7, which uses the transformers *mrm8488/bert-spanish-cased-finetuned-ner* and *mDeBERTa* for target detection and sentiment classification, respectively. With this approach we reached 4th position for *Task 1: Financial targeted sentiment analysis* and 2nd position for *Task 2: Financial Sentiment Analysis at document level for companies and consumers*. The official leaderboards for both tasks can be consulted in Table 9 and Table 10.

6. Conclusions and future work

In this paper we have presented the participation of the SINAI team in the shared task FinancES, Financial Targeted Sentiment Analysis in Spanish, at IberLEF 2023. The objective of our experiments, for the target detection task, was to test the performance of the most popular NER tools against transformer-based models. The main conclusion is that transformers-based solutions outperformed others as Stanza or spaCy related approaches. On the other hand, for the sentiment analysis tasks, the aim of our experiments was to test how some of the most popular transformers models behave compared to specific financial transformers models. In

Table 9
Official leaderboard for *Task 1: Financial targeted sentiment analysis*

ranking	team	f1 task1	f1 target	f1 target sentiment
1	abc111	**0.792244 (1)**	**0.877137 (1)**	0.707350 (4)
2	LLI-UAM	0.792172 (2)	0.852179 (3)	**0.732164 (1)**
3	ABCD Team	0.782175 (3)	0.854511 (2)	0.709838 (3)
4	SINAI	0.778002 (4)	0.838174 (4)	0.717829 (2)
5	AnkitSinghRaikuni	0.554211 (5)	0.575360 (6)	0.533062 (7)
6	UTB-NLP	0.529229 (6)	0.410079 (8)	0.648379 (5)
7	NLP_URJC	0.514414 (7)	0.606773 (5)	0.422055 (9)
8	BASELINE	0.498107 (8)	0.428393 (7)	0.567822 (6)
9	mario.pv	0.276926 (9)	0.106326 (9)	0.447526 (8)
10	UNAM Text Mining	0.134680 (10)	0.086643 (10)	0.182717 (10)
11	fanchuyi	0.000000 (11)	0.000000 (11)	0.000000 (11)

Table 10
Official leaderboard for *Task 2: Financial Sentiment Analysis at document level for companies and consumers*

ranking	team	f1 task 2	f1 com sentiment	f1 con sentiment
1	LLI-UAM	**0.642349 (1)**	**0.592590 (1)**	**0.692109 (1)**
2	SINAI	0.634901 (2)	0.583483 (3)	0.686320 (2)
3	ABCD Team	0.610373 (3)	0.588635 (2)	0.632111 (3)
4	abc111	0.575015 (4)	0.530284 (4)	0.619746 (4)
5	fanchuyi	0.472685 (5)	0.414230 (7)	0.531139 (5)
6	AnkitSinghRaikuni	0.457632 (6)	0.419755 (6)	0.495509 (6)
7	BASELINE	0.433783 (7)	0.384268 (8)	0.483298 (7)
8	NLP_URJC	0.425126 (8)	0.436560 (5)	0.413692 (8)
9	UNAM Text Mining	0.370396 (9)	0.345686 (9)	0.395107 (9)
10	mario.pv	0.248196 (10)	0.269267 (10)	0.227125 (10)
11	UTB-NLP	0.000000 (11)	0.000000 (11)	0.000000 (11)

this case we conclude that generic transformer models perform better than existing financial transformers models.

In the future, we plan to evaluate why financial transformers perform worse. In addition, we want to continue evaluating external resources to further improve the training phase of the system by analyzing the contribution of each model, testing different transfer learning systems as well as models trained on general topics, and using different machine translation systems to generate new datasets and/or augment existing ones.

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, Project FedDAP (PID2020-116118GA-I00) supported by MICINN/AEI/10.13039/501100011033 and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

References

[1] 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.

[2] J. A. García-Díaz, 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).

[3] M. M. Hasan, J. Popp, J. Oláh, Current landscape and influence of big data on finance, Journal of Big Data 7 (2020) 1–17.

[4] J. W. Goodell, S. Kumar, P. Rao, S. Verma, Emotions and stock market anomalies: A systematic review, Journal of Behavioral and Experimental Finance 37 (2023). URL: https://ideas.repec.org/a/eee/beexfi/v37y2023ics2214635022000557.html. doi:10.1016/j.jbef.2022.10072.

[5] L. Nemes, A. Kiss, Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication 5 (2021) 375 – 394.

[6] A. Milne, M. Chisholm, The Prospects for Common Financial Language in Wholesale Financial Services, SWIFT Institute Working Paper, SSRN, 2013. URL: https://books.google.es/books?id=ZZEhzwEACAAJ.

[7] OpenAI. (2023). ChatGPT (May version) [Large language model], https://chat.openai.com, 2023.

[8] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, 2020. arXiv:2003.07082.

[9] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[10] 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.

[11] J. García-Díaz, M. Salas Zarate, M. Hernández-Alcaraz, R. Valencia-García, J. Gómez Berbis, Machine Learning Based Sentiment Analysis on Spanish Financial Tweets, 2018, pp. 305–311. doi:10.1007/978-3-319-77703-0_31.

[12] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu,

D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models, 2023. arXiv:2304.01852.
[13] Baccouri, Nidhal, A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators https://deep-translator.readthedocs.io/, 2020.