AIInFunds. Intelligent Services for Monitoring Tendencies of Alternative Investment Funds

José Antonio García-Díaz[1], José Antonio Miñarro-Giménez[1], Ángela Almela[2], Gema Alcaraz-Mármol[3], María José Marín-Pérez[2], Francisco García-Sánchez[1] and Rafael Valencia-García[1]

[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, España
[2] Facultad de Letras, Universidad de Murcia, Campus de la Merced, 30001, Murcia, España
[3] Departamento de Filología Moderna, Universidad de Castilla La Mancha, 45071, España

Abstract

Alternative investment funds have high strategic value but high uncertainty associated with them because they are relatively new and difficult to monitor using traditional approaches. In addition, being such specialized products there is little public information. The main objective of the AIInFunds project is to use Natural Language Processing and Semantic Web technologies to extract information about alternative assets and apply entity extraction techniques and sentiment and emotion analysis to measure social perception towards these alternative assets. AIInFunds is being developed by the TECNOMOD group of the University of Murcia and is financed by the Spanish National Research Agency and by the European Union NextGenerationEU/PRTR through the Prueba de Concepto 2021 projects.

Keywords

Large Language Models, Sentiment Analysis, Emotion Analysis, Semantic Web, Natural Language Processing

1. Introduction and main objective

This project is funded by the Prueba de Concepto 2021 call[1] and is based on a previous project KBS4FIA [1], whose objective was the development of Natural Language Processing and knowledge extraction technologies within the financial domain to enable a better management of unstructured information for decision-making within this domain. Knowledge extraction technologies based on Spanish texts were developed including named entity recognition, text classification, semantic annotation and ontology population. Deep-learning technologies were also developed to process subjective language in Spanish, which involved sentiment analysis in different domains, including the financial domain. In addition, technologies for the efficient extraction of information from news and social networks were also implemented and a large number of financial corpora were collected, which were only partially annotated.

Alternative investment funds have a high strategic value and have become fashionable in this sector. Some examples of these assets are venture capital and certain physical assets such as infrastructure or real estate. One of the reasons for the success of alternative assets is that the foreseeable future of traditional investment options is worse, which is why investors have opted for assets with higher return expectations, although riskier. However, for the efficient use of these assets certain challenges should be considered. First, since they have not been used traditionally, investors are not familiar with the risks involved. Second, these assets require a high degree of specialization, as well as extensive knowledge of the existing legislation on them. Third, these assets are often illiquid, which makes it difficult both to determine their market price and sell them quickly. Fourth, they require a long-term investment, which, in some cases, can be as long as 10 to 15 years. Additionally, these assets have high management costs, including higher fees than traditional assets, which may also put off some potential investors. Finally, these assets raise doubts among investors who are wary due to the little historical and analytical information available. The scarce number of people interested in investing in such specialized products has caused a relative lack of information, which, to top it all off, is often not made public. In fact, in Spain there are very few companies dedicated to managing or advising on this type of assets.

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

joseantonio.garcia8@um.es (J. A. García-Díaz); jose.minyarro@um.es (J. A. Miñarro-Giménez); angelalm@um.es (: Almela); Gema.Alcaraz@uclm.es (G. Alcaraz-Mármol); mariajose.marin1@um.es (M. J. Marín-Pérez); frgarcia@um.es (F. García-Sánchez); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-7221-9843 (J. A. Miñarro-Giménez); 0000-0002-1327-8410 (: Almela); 0000-0001-7703-3829 (G. Alcaraz-Mármol); 0000-0003-0177-4860 (M. J. Marín-Pérez); 0000-0003-2667-5359 (F. García-Sánchez); 0000-0003-2457-1791 (R. Valencia-García)

[1] https://www.aei.gob.es/convocatorias/buscador-convocatorias/proyectos-idi-pruebas-concepto-2021

To assess the opinion and experience of Spanish investors with alternative assets, AFI Trust conducted a survey among a group of entities with different profiles and investment objectives: (1) insurance companies and pension funds, (2) banks private, (3) foundations or associations, and (4) non-financial corporations. The report [2] highlights liquidity restrictions and the difficulty of analysis among the main obstacles when it comes to investing in this type of assets. Due to the motivation described above, this project consists in the creation of a platform that allows investors to create a series of alternative assets and facilitates the monitoring of information found in documents on the Internet, as well as in publications on social networks or specialized news. In this way, we aim to fill the non-specialized knowledge gap that currently exists about alternative assets.

We intend to use Natural Language Processing technologies to, first, identify concepts related to these assets and, second, apply sentiment analysis techniques focused on measuring people's perception of these issues. This will be available through a customizable control panel that will allow to filter information geographically using time intervals. At a technological level, this platform is integrating the solutions developed during the execution of the KBS4FIA project, which are being optimized to be efficient and scalable through their implementation on a real salable platform. In addition, a microservices architecture is being built to commercialize the platform and be able to integrate and adapt these technologies to software development companies.

2. Project status

The AIInFunds project has the following specific objectives: (OB1) Development and annotation of linguistic resources in the alternative investment funds domain, (OB2) Development and optimization of deep learning-based models for conceptual aspect modelling and emotion analysis in the alternative investment funds domain, (OB3) Integration and optimization of modules in a global platform, (OB4) Deployment of the global platform in a Software-as-a-Service commercial and scalable environment, and (OB5) Customization and validation of the global platform in various scenarios.

This project was scheduled in 24 months starting the 1st of December 2021 and was divided into two management and five development work packages. The five development-related work packages are described next. The first work package consists of compiling and annotating linguistic resources in the domain of alternative investment funds. Here, different resources and lexicons from the financial domain, including news corpus, social networks and domain ontologies, are being compiled for the semantic exploitation of the results. The compiled data is being recorded in different categories to be able to carry out different studies, such as feelings or emotions. Tools such as web crawlers and the UMUCorpusClassifier [3] are being used for the compilation and annotation processes. The second work package consists of training and optimizing deep learning-based models for conceptual aspect modeling and emotion analysis in the domain of alternative investment funds. The last two work packages consist of the development of a global platform for the monitoring of alternative assets based on a microservices architecture. The final work package is concerning the customization and validation of the global platform in two scenarios related to alternative investment funds.

Some of the most important milestones reached so far in the project are described next.

2.1. FINA

One of the objectives of this project is the development of a language model, namely FINA, focused on the domain of economics and finances. To do this, we trained a model based on the RoBERTa [4] architecture with a corpus compiled from 5 gigabytes of financial news. To compile the corpus we used the Spatie crawler[2]. We selected more than 100 newspapers with financial content. Some examples of these newspapers are: Expansión[3], Invertia[4], ABC [5] or ModaEs[6]. It is worth noting that not all newspapers are from Spain. Some of them are from other Spanish-spearking countries such as Diario Financiero[7] from Chile, or El Financiero[8] from Mexico. As not all the newspapers are focused solely on providing financial information, the news were extracted by applying different filters. The first filter was based on regular expressions within the news URLs. Thus, it was possible to focus on news items that were located within sections such as `/finanzas/` or `/economia/` of the processed web portals. The second filter was based on CSS rules, so that we could keep the relevant content, discarding from the HTML code advertisements, links to related news or unrelated content.

The HTML tags were removed from the compiled news items, which were then converted into Markdown format. We chose this format because we wanted to keep the basic structural elements of each news item, such as headlines or sections. The next step was cleaning the corpus. To do this, a script was developed that removes irrelevant data from the news items (e.g., dates or information from the authors who have written the journalistic piece).

[2] https://github.com/spatie/crawler
[3] https://www.expansion.com
[4] https://www.elespanol.com/invertia
[5] https://www.abc.es
[6] https://www.modaes.com/
[7] https://www.df.cl/
[8] https://www.elfinanciero.com.mx/

Table 1
Example of the FinancES 2023 dataset

Text	MET	S. MET	S. Companies	S. Society
Acuerdos comerciales, sinónimo de oportunidades para República Dominicana	Acuerdos comerciales	POS	POS	POS
EDP Renováveis vende unos activos eólicos en Portugal a China Three Gorges por 242 millones	EDP Renováveis	POS	POS	NEU
El petróleo avista los 82 dólares: la demanda gana a la normalidad en Kazajistán y Libia	Petróleo	POS	NEG	NEG

We are currently training two configurations based on RoBERTa. The first model has 12 attention heads and 12 hidden layers. The second model also has 12 attention heads, but only 6 hidden layers. The tokenizer model is `Byte-Level BPE` with the following parameters: a maximum of 512 tokens, a vocabulary size of 52,262 (the RoBERTa default), and a minimum frequency of 2. The dataset is processed paragraph by paragraph, that is, we extract multiple texts from each news. Finally, it is worth mentioning that the model is trained with the *Masked Language Modeling* task, with a probability of 15% and for 5 epochs.

To validate the suitability of the model and its applicability to different tasks, we are testing its performance on a sentiment analysis task with a financial corpus. We are comparing the results with more general Spanish pretrained models (e.g., BETO [5], MarIA [6], BERTIN [7]) and other multilingual models.

2.2. Sentiment analysis and evaluation of language models

In [8] we explore the impact of combining different feature sets for sentiment analysis in financial texts in Spanish. To do this, we compiled a corpus of 15,915 tweets and annotated them as either positive, negative, or neutral. Then, features based on word embeddings were evaluated along with linguistic features [9]. These features were evaluated both individually and combined using ensemble learning and knowledge integration. The ensemble learning strategy consists into generate a new output based on the outputs of the rest of the models. For this, we evaluate the mode of the predictions, averaging probabilities, selecting the label with the highest probability along all the models and a weighted mode based on custom validation results. The second evaluated strategy, knowledge integration, is the retrain of a new multi-input neural network model that combines all the feature sets and outpus a unique result. The best approach achieved an F1 of 73.15880%, combining the features evaluated using the knowledge integration strategy. The non-contextual word and sentence embeddings considered in our study are as follows: GloVe [10], Word2Vec [11] and fastText [12]. Also different pre-trained transformer-based models were evaluated, namely, (i) BETO [5], (ii) ALBETO and (iii) DistillBETO, which are light variants of BETO [13], (iv) MarIA [6] and (v) Bertin [14], both based on the RoBERTa architecture, and (vi) multilingual BERT [15] and (vii) XLM [14], two multilingual models.

2.3. Targeted Sentiment Analysis

Another of the tasks involved in this project is the development of a targeted sentiment analysis approach in the financial domain. To do this, we compiled a corpus with close to 80,000 texts, coming from social networks and news headlines. The idea behind this targeted classification method is to predict the sentiment polarities of different targets in the same text. In classic sentiment analysis system, only the sentiment towards the main entity or topic is calculated. In this approach, however, the model is trained to obtain the sentiments towards three types of entities: (1) towards the main economic target (MET), (2) towards other companies, and (3) towards society in general. In addition, this model is also capable of recognizing the entity (i.e., the MET) using a sequence classification model, in the style of Named Entity Recognition.

For the purposes of this experiment, in the first place, a subset of the corpus of news and tweets was selected and the MET and sentiment towards these three entities (i.e., the MET, other companies, and society in general) were manually annotated. Then, different classifiers were trained evaluating various language models, both specific to Spanish and multilingual. The results of this study can be found at [16].

Another contribution in this field is the organization of the FinancES 2023 [17] shared-task, which is held within the IberLEF 2023 workshop. This shared task uses a subset of the compiled corpora and proposes two tasks to the participants. On the one hand, to identify the main entity that appears in the text and its sentiment and, on the other hand, the sentiments towards other companies and society as a whole. The FinancES 2023 dataset contains 6359 documents for training and 1621 documents for testing. Table 1 contains some examples of the dataset, the MET and the sentiments towards this target (S. MET), other companies (S. Companies), and society (S. Society). This competition is hosted in the Codalab platform[9].

In order to facilitate the participation we have allowed the participants to send their results for both subtasks independently. Besides, we prepared two notebooks to show to the participants how to train a baseline model based on TF–IDF features trained with logistic regression for the sentiment polarity detection task, and a model based on Spacy [18] for detecting the MET. These notebooks also contain instructions about how to prepare the submission files required to participate in the competition.

At the end of the evaluation stage, a total of 10 teams have participated, achieving competitive results in both subtasks.

3. Further work

As future work, we are working on developing a global platform in a commercial environment with the aim to deliver the functionality through a *Software-as-a-Service* model. Once developed, we will validate it based on common use cases to assess its usefulness in real scenarios. One major line of improvement is to incorporate open-data sources to the platform. As these data are usually available through different APIs, we are exploring the reliability of using code generation tools based on LLMs models to automate this process [19].

Once we finish with the evaluation of the FINA model in several NLP tasks, we will release two versions (a large one with 12 hidden layers and a lightweight one with only 6 hidden layers) for the research community and industry in Huggingface[10] together with a manuscript with the details of its compilation and the analysis and benchmark of this model applied to the different NLP tasks.

Concerning the organization of shared tasks, next year we will focus on NLP tasks dealing with financial data in Spanish. However, we want to incorporate texts that include comments and opinions from blogs in order to introduce a more informal speech and the presence of figurative language [20].

4. Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

[9] https://codalab.lisn.upsaclay.fr/competitions/10052
[10] https://huggingface.co

References

[1] F. García-Sánchez, M. A. Paredes-Valverde, R. Valencia-García, G. Alcaraz-Mármol, Á. Almela, KBS4FIA: leveraging advanced knowledge-based systems for financial information analysis, Procesamiento del Lenguaje Natural 59 (2017) 145–148. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5507.

[2] AFI: Analistas Financieros Internacionales, Inversión en activos alternativos, https://media.afi.es/webcorporativa/2022/07/Estudio-activos-alternativos-Afi-Aberdeen_DEF_Actualizacion-JUN22.pdf, 2022. Accessed: 2023-05-25.

[3] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6292.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC at ICLR 2020, 2020, pp. 1–10.

[6] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[7] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, M. Grandury, BERTIN: efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[8] J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Smart analysis of economics sentiment in spanish based on linguistic features and transformers, IEEE Access 11 (2023) 14211–14224. URL: https://doi.org/10.1109/ACCESS.2023.3244065. doi:10.1109/ACCESS.2023.3244065.

[9] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, UMUTextStats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation

Conference, 2022, pp. 6035–6044.

[10] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[12] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893 (2018).

[13] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, arXiv preprint arXiv:2204.09145 (2022).

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

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

[16] R. Pan, J. A. García-Díaz, F. Garcia-Sanchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in spanish, PeerJ Computer Science 9 (2023) e1377.

[17] J. A. García-Díaz, F. García-Sánchez, R. Valencia García, Overview of FinancES 2023: Financial targeted sentiment analysis in spanish (to appear), Procesamiento del Lenguaje Natural (2023).

[18] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, Y. LeTraon, A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2019, pp. 338–343.

[19] B. Yetiştiren, I. Özsoy, M. Ayerdem, E. Tüzün, Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt, arXiv preprint arXiv:2304.10778 (2023).

[20] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of English literature on figurative language applied to social networks, Knowledge and Information Systems 62 (2020) 2105–2137. doi:10.1007/s10115-019-01425-3.