FD. A Platform for Monitoring Financial and Economic Information towards 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

For efficient financial asset management, it is necessary to select, process and analyze specific information on the Internet. However, the large volumes of information available and the fact that most of this data is stored in an unstructured way, hinder this task. In this demo, we present Financial Dashboard (FD), a global platform to monitor financial assets from the Internet using Natural Language Processing and Semantic Web technologies focused on the Spanish language. The objective is to allow users to monitor financial data from a set of keywords, accounts and digital newspapers. FD compiles data periodically and annotates semantic information such as financial entities or sentiments. All the information is made available to users from a web dashboard composed by configurable and independent KPIs and a REST API.

Keywords

Alternative Investment Funds, Sentiment Analysis, Semantic Web, Natural Language Processing

1. Introduction

To boost financial management and to improve the efficiency in the use of public and private economy-related resources, it is necessary to monitor the Internet in search of financial data. This process involves selecting, processing and analyzing the global and local financial activity. A proper financial management helps companies and public authorities to identify risks and opportunities. However, this task is not an easy one. First, there is a huge amount of information on the Internet, which makes it challenging to handle, especially with real time requirements. Second, most information can be found stored in an unstructured or semi-structured way, so it is hard to take advantage of all such information. Third, state-of-the-art technologies for Natural Language Processing (NLP) are mainly focused on the English language and have not been tested properly in Spanish texts.

FD (Financial Dashboard) is a global platform that eases the monitoring of financial assets from the Internet using NLP tools and Semantic Web technologies. In a nutshell, this tool allows managers of companies and technicians of public organizations to establish a set of keywords, social networks accounts and digital newspapers to monitor. The system compiles the data from those sources periodically and extract semantic information including financial entities or sentiments. Also, the system scores each piece of information in order to determine their relationship with a set of objectives that can be previously defined by end-users. Finally, all information is made accessible through a web dashboard composed by configurable and independent Key Performance Indicators (KPIs) and deployed in the form of a REST API.

At the technological level, this platform makes use of NLP techniques based on state-of-the-art Large Language Models (LLMs) for extracting objective and subjective information from textual sources and Semantic Web technologies to map those concepts to a domain ontology.

2. Background information

In this section we briefly analyze similar tools and approaches for monitoring financial data on the Internet. In [1] the authors describe a system procedure for measuring explicit and implicit linkages between large U. S. bank holding companies. In their methodology, the authors propose the usage of mixed-frequency regression techniques. This component provides bank supervisors

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)

CEUR Workshop Proceedings (CEUR-WS.org)

with knowledge about when new assets need to be monitored. Besides, the authors demonstrate how variables concerning outcome can be applied to measure the extent to which firms are interconnected. Another related study is [2], in which the authors introduce a framework for quantitative investments and trading in financial markets. This framework is subdivided into four components to monitor global variables, including (1) quantitative investment trading, (2) financial risk monitoring, (3) economic situation, and (4) environmental risk monitoring.

These previous studies do not take into account social data. However, in [3] the authors monitor fine-grain housing rental prices in order to bring insights for fair housing policies. For this, they focus on housing rental websites for their studies in China. They consider features concerning the location, the neighborhood, the home structure or accessibility, among other features. They use time data between 2017 and 2018 and evaluate several classical machine-learning regression algorithms such as random forest, gradient-boosting or support vector machines. Their results suggest that most of the generated models have good performance and that the two most influential features are related to job opportunity and accessibility to health care services.

As far as our knowledge goes, there are no studies focused on monitoring financial data from social networks such as Twitter considering texts written in Spanish.

3. System architecture

Figure 1 depicts the overall architecture of the system. As it can be seen, the system architecture is divided into layers. The first layer comprises all the models responsible for data compilation. We distinguish among two data sources. On the one hand, we compile news from digital newspapers and, on the other, posts from social networks. The second layer focuses on the annotation and information extraction. Each piece of information compiled is ranked, semantically annotated with a domain ontology and their sentiment towards several entities is calculated. The last layer of the architecture is the dashboard, which is composed of a set of configurable and autonomous KPIs and a configurable alerts system. In the next subsections, these layers are described with some detail.

The platform will be available freely to the users but with some limitations. For example, users will be only allowed to create one dashboard and the total number of sources to monitor will be also limited. Premium users will be able to create several dashboards and consider an unlimited number of sources.

Finally, it worth noting that all services and software artifacts will be deployed using docker containers. This strategy allows to scale the software platform according to the needs of the final users.

3.1. Layer 1. Data acquisition module

This project has two main data sources, namely, news from digital newspapers and publications from social media sites. On the one hand, the news are extracted using a custom web crawler. This crawler can filter news sites based on two strategies. The first strategy is to filter by URL using regular expressions. For example, it is possible to restrict the system to consider only pages whose URL contains `/economia/`. The second strategy is to filter using CSS filters, as some of the news sites include certain rules in the style to denote financial content. The content is then stored in a markdown format, as we keep some structural information of the news. On the other hand, the social media items are extracted from Twitter. We use the Twitter API together with the UMUCorpusClassifier tool [4] to filter certain Twitter accounts of digital newspapers.

The stored data is also pre-processed in order to remove hyperlinks, mentions, and languages that are not Spanish. Finally, every piece of information is geolocated with a latitude, longitude and a radius. The radius allows to set very specific items located to specific regions or cities or to be more generic, spotting autonomous communities or even countries. To calculate this position, we use different heuristics such as looking for locations in the headline or the main text and then use a reverse geolocation utility, to set the current position in the map. If no information is found, we search for meta-data and author information.

3.2. Layer 2. Semantic data annotation

This layer comprises a set of modules that extract semantic information from the news items.

First, we extract the sentiment polarity of the news item. However, determining the polarity of financial news is complex as an event can be deemed positive or negative depending on the target. For instance, news can report facts that are positive for banks but negative for the society. Besides, the language employed in financial news is highly dependent on the context since there are expressions that may be either positive or negative depending on the context. For example, is not the same that the stock market shares *rise* than the debt *rises*. These facts has lead to unsatisfactory results of current sentiment analysis solutions within the financial domain. Besides, novel LLMs have yet to be evaluated in the financial domain.

To solve these limitations different strategies have been explored. On the one hand, we have compiled a corpus of financial news to test the effectiveness and performance of state-of-the-art LLMs including MarIA [5] or BETO [6]. The results of this analysis have been published in [7]. On the other hand, we have compiled and annotated

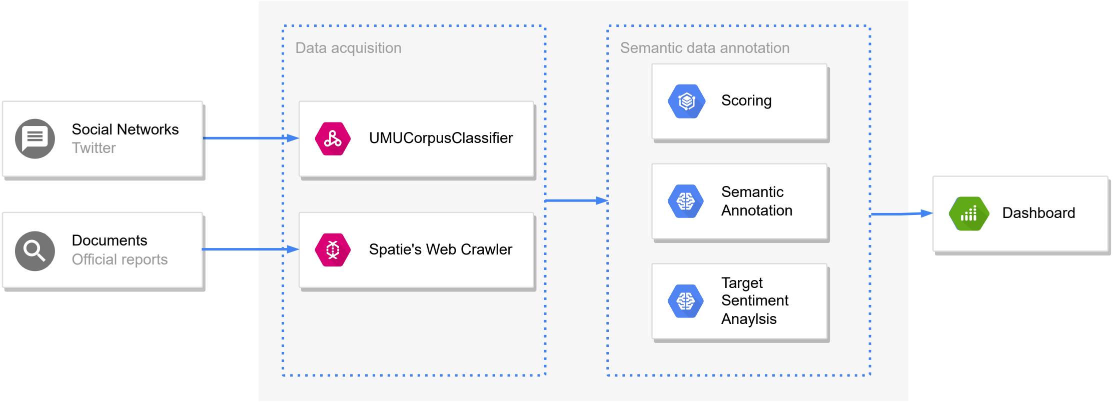


Figure 1: System architecture

a new corpora of financial headlines and sites to conduct a targeted sentiment analysis. The goal was to extract the main economic target of the document and then the sentiment towards this target as well as the sentiments towards other companies and society in general. For this, we use two neural network models. The first one is trained with a Named Entity Recognition (NER) task, and it is capable of identifying the target. The second neural network is a multi-label document classification model that it is able to capture the sentiments towards the three targets (main economic target, companies, and society) at once. For the latest model, different LLMs have been evaluated, including large and base models of BETO and MarIA, and also lightweight models based on distillation and multilingual LLMs.

The next module is focused on extracting entities and mapping them to a domain ontology. For this, we created a novel ontology that contains concepts related to different financial sectors including tourism, technologies, health, industry or energy, among others. Each concept in the ontology allows to define a set of named entities to identify relevant companies and related actors. Each compiled piece of information is mapped to the ontology using semantic annotation based on an extended version of the Term-Frequency Inverse Document score (TF–IDF–e) [8]. This strategy is based on the TF–IDF measure, that calculates the frequency of different terms (TF) and weights this information concerning how informative is the term in the rest of the documents (IDF). Once we have obtained the TF–IDF for each of the terms of the ontology that appear explicitly in the texts, we calculate the weight of the terms that appear implicitly. For this, the extended TF–IDF takes into account the distance between each identified entity with the rest of the concepts in the ontology.

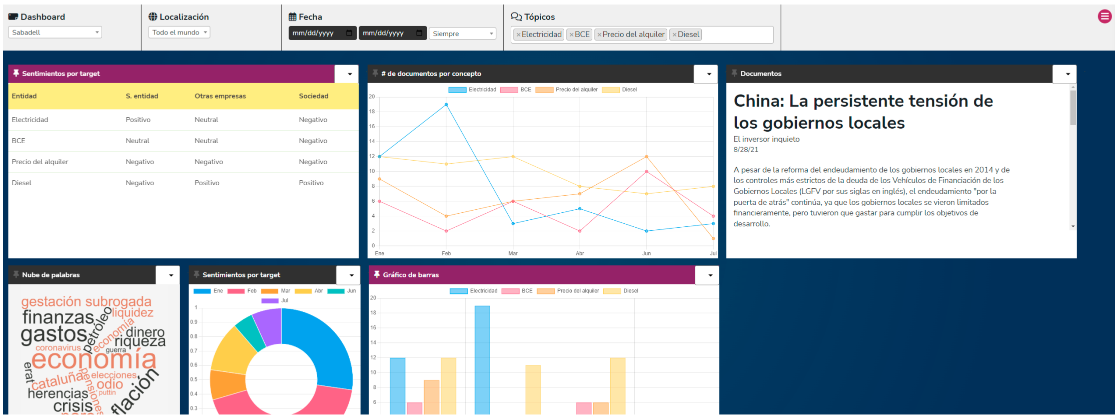
Once the sentiments and the semantic annotations have been obtained, the relevance of each piece of information is ranked according to the users' interests. This process allows to prioritize some items over others.

3.3. Dashboard

The last module of the FD platform is a dashboard. This dashboard is built using progressive web technologies. It enables users to create multiple projects and dashboards. Each dashboard is composed of several configurable KPIs, with each KPI associated to a group of customs filters. These filters allow to set concepts from the ontology, time series, keywords, or the way in which data is to be visualized (e.g., word clouds, timelines, heatmaps, tables, etc.). Besides, each KPI can be attached to specific filters and facilitates the comparison of trends to final users. It is also worth mentioning that the data from each KPI is accessible using a REST API too, so that the platform can be interconnected to external systems and tools.

Figure 2 presents a screenshot of the dashboard. As it can be seen, the dashboard contains a generic filter for all KPIs. In the capture, this filter is configured to show data from the last six months for four topics, including electricity, the European Central Bank, the rental price, and diesel oil. Below, the main KPIs are shown. Some of them are configured to show the sentiments and targets for the selected topics. Another KPI panel contains the number of documents per topic, and there are also KPI panels to show relevant documents, pie charts and a word cloud.

It worth noting that the KPIs that are organized per target are based on the dataset published in the shared task FinancES 2023 [9], which consists in determining the main entity that appears in economic headlines and the sentiments towards this target, other companies and society in general. Targeted sentiment analysis can determine



Dashboard: Sabadell

Localización: Todo el mundo

Fecha: mm/dd/yyyy mm/dd/yyyy Siempre

Tópicos: ×Electricidad ×BCE ×Precio del alquiler ×Diesel

Sentimientos por target

Entidad	S. entidad	Otras empresas	Sociedad
Electricidad	Positivo	Neutral	Negativo
BCE	Neutral	Neutral	Negativo
Precio del alquiler	Negativo	Negativo	Negativo
Diesel	Negativo	Positivo	Positivo

Documentos

China: La persistente tensión de los gobiernos locales

El inversor inquieto

8/28/21

A pesar de la reforma del endeudamiento de los gobiernos locales en 2014 y de los controles más estrictos de la deuda de los Vehículos de Financiación de los Gobiernos Locales (LGFV por sus siglas en inglés), el endeudamiento "por la puerta de atrás" continúa, ya que los gobiernos locales se vieron limitados financieramente, pero tuvieron que gastar para cumplir los objetivos de desarrollo.

Figure 2: Illustrative screenshot of the dashboard including the filter and several of the KPIs

the polarity of certain texts to different economical and social groups. This strategy distinguishes among three types of targets: (1) the main economic entity (MET), (2) the rest of the companies, and (3) the society and consumers. Besides, this approach can extract the main entity using a NER system.

4. Further work

In this work we have described the FD platform for monitoring economic and financial data on the Internet. This platform relies on Semantic Web technologies and NLP techniques for extracting, annotating and classifying financial data from several data sources, including web sites and social networks. The data is presented to the end-users in a web platform that allows them to configure a personalized dashboard with a set of configurable KPIs.

We are currently on the last stages of the development of the platform and we are preparing several case studies for its validation. The further work is focused on improving the explainability of the neural network models. In particular, we plan to create a module based on linguistic features [10] and define KPIs that highlight the relevant parts of the text that contributed the most to the predictions of the sentiments. Another idea for improving the platform is to add more data filters.

Currently, we are working on incorporating information from video platforms such as YouTube. We will also focus on the definition of KPIs that cluster results per data-source [11] and improving the number of filters. Finally, we will evaluate the feasibility of incorporating KPIs for the detection of fake news.

5. 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.

References

[1] G. Hale, J. A. Lopez, Monitoring banking system connectedness with big data, Journal of Econometrics 212 (2019) 203–220. URL: https://www.sciencedirect.com/science/article/pii/S030440761930082X. doi:https://doi.org/10.1016/j.jeconom.2019.04.027.

[2] H. Pan, Intelligent finance global monitoring and observatory : A new perspective for global macro beyond big data, in: 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019, pp. 623–628. doi:10.1109/ICPHYS.2019.8780156.

[3] L. Hu, S. He, Z. Han, H. Xiao, S. Su, M. Weng, Z. Cai, Monitoring housing rental prices based on social media:an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies, Land Use Policy 82 (2019) 657–673. URL: https://www.sciencedirect.com/science/article/pii/S0264837718316429. doi:https://doi.org/10.1016/j.landusepol.2018.12.030.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. Á. Rodríguez-García, R. Valencia-Garcí, F. García-Sánchez, J. J. Samper-Zapater, Creating a semantically-enhanced cloud services environment through ontology evolution, Future Generation Computer Systems 32 (2014) 295–306. URL: https://www.sciencedirect.com/science/article/pii/S0167739X13001684. doi:https://doi.org/10.1016/j.future.2013.08.003.

[9] 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).

[10] 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.

[11] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74. doi:10.1016/j.future.2021.12.011.