SIGAMER: A Decision Support System for Demand Management in the Retail Industry Based on the Intelligent Analysis of Structured Data and Social Networks

Diego Roldán[1], Ronghao Pan[2], Camilo Caparrós-Laiz[2], José Antonio García-Díaz[2] and Rafael Valencia-García[2]

[1]DANTIA Tecnología S.L., Parque Empresarial de Jerez 10, Calle de la Agricultura, 11407, Jerez de la Frontera, Cádiz, Espa~na
[2]Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

This project aims to develop a state-of-the-art decision support platform for retail management. The project, called SIGAMER, integrates Natural Language Processing technologies and big data capabilities. The main objective is to use historical data from companies' ERPs and extract insights from different sources such as structured data, open data, social media and news. The platform, accessible through a centralized interface and various REST APIs, follows a Software as a Service approach for seamless integration with existing Enterprise Resource Planning systems. The platform's technologies include semantic extraction, knowledge graphs, aspect-based multimodal sentiment analysis, topic modeling, author profiling, and regression and classification models for demand forecasting. The final system is accessible through a web technology-based dashboard composed of configurable key performance indicators to provide decision support to retailers who can use this insight to improve asset management, optimize deployment and identify lucrative market sectors.

Keywords

Retail Management, Multi-modal, Aspect-based Emotion Analysis, Knowledge Graphs, Natural Language Processing

1. Introduction and main objective

This project is funded by the Spanish Government and the Ministry of Digital Transformation and by the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia", under the 2021 call for research projects in Artificial Intelligence and other digital technologies and their integration in value chains.

DANTIA Tecnología S.L. is a software consulting and development company specialized in providing Enterprise Resource Planning (ERP) solutions for procurement, sales, production and warehouse management. In recent years, DANTIA has focused on incorporating Natural Language Processing (NLP) technologies into its pipeline and big data capabilities.

The main objective of this project, called SIGAMER, is to develop a decision support platform to improve retail management by leveraging companies' historical data from their ERPs and extracting insights from structured sources, open data and natural language documents such as social media and news. This platform will be accessible to end users through a centralized interface and an ecosystem of services through various REST APIs, enabling integration with existing services such as ERP under a Software as a Service (SaaS) approach.

SIGAMER collects data from open data sources and social media about trends, potential customer opinions, and short to medium term risk and opportunity analysis. Advanced NLP and deep learning technologies are used to organize this information and provide decision-support KPIs. As a result, retailers can gain deeper insights into market trends to improve asset management and provisioning. It also facilitates the analysis of new products and potentially lucrative market sectors. This main goal is divided into 5 objectives.

- (OB1) Development of an intelligent information crawler for news for official reports and social media for subjective information and news sources for official reports. The crawlers will compile multimodal data from text, images or PDF documents.
- (OB2) Implementation of a semantic extraction and representation system based on ontologies. The goal is to associate information extracted from indexed documents with semantic structures based on knowledge graphs for data reasoning.

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain
droldan@dantia.es (D. Roldán); ronghao.pan@um.es (R. Pan); camilo.caparrosl@um.es (C. Caparrós-Laiz); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)
https://www.dantia.es/ (D. Roldán); https://github.com/Smolky (J. A. García-Díaz); https://webs.um.es/valencia (R. Valencia-García)
0009-0008-7317-7145 (R. Pan); 0000-0002-5191-7500 (C. Caparrós-Laiz); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

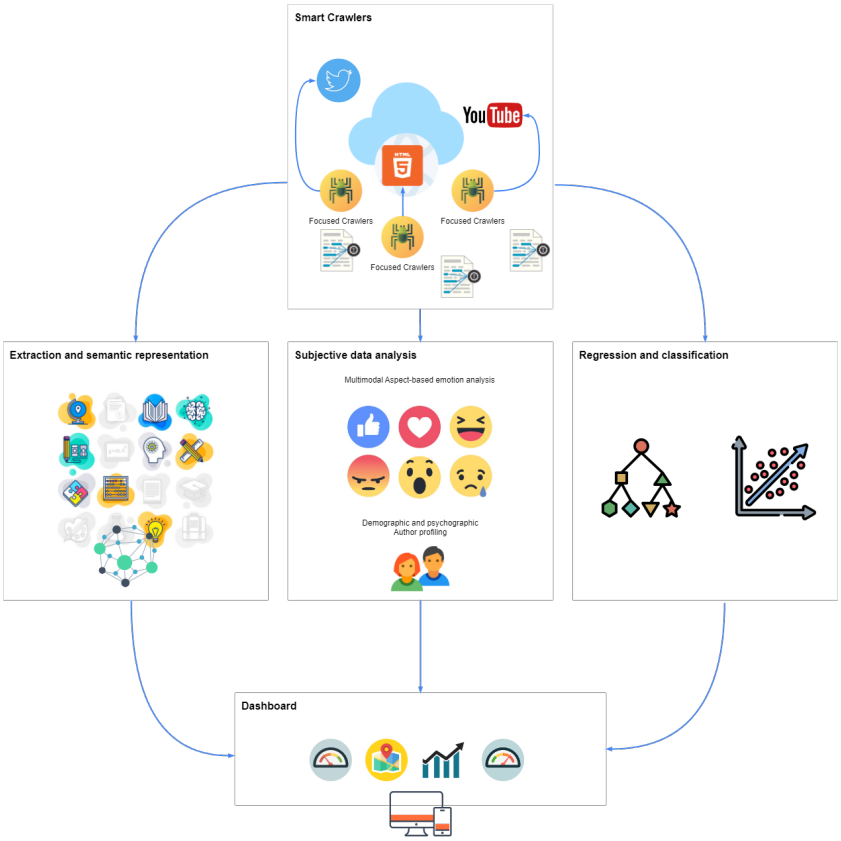

Figure 1: System architecture

- (OB3) Create a subjective text analysis system that uses state-of-the-art language models to build an aspect-based multimodal emotion analysis based on aspects extracted from ontologies.
- (OB4) Develop regression and classification models for demand forecasting based on product types and sales forecasts.
- (OB5) Create an intelligent retail demand management platform based on intelligent analysis of structured data and social networks. This includes integration of the previous modules, validation processes and testing to obtain the final system. The final modular system will consist of a configuration section, a dashboard view, and an alert system to notify users of potential problems.

2. System architecture

The architecture of the proposed system is shown in Figure 1. Below is a brief description of each of these components.

2.1. Smart Crawlers module

This component consists of a set of smart crawlers that collect information from websites and social networks. The crawlers can adapt to different domains, although they share a common interface. In the case of social networks, specific crawlers have been developed separately, such as X (formerly Twitter) or YouTube. Each smart crawler is configurable and different strategies can be defined, such as the frequency with which each crawler starts monitoring and the criteria it must have to decide whether a page is linked or not based on the content and specific keywords.

The smart crawlers can extract metadata from the web pages and multimedia content. On the one hand, metadata in in formats such as JSON-LD[1] (JSON for Linked-Data) to try to give interoperability to the data and have more confidence in the subject of linked data. Metadata about published and updated data is also tracked. On the other hand, multimodal content such as audio and

[1]https://json-ld.org/

images. For audio data extraction, we use Whisper [1], an automatic speech recognition (ASR) model based on the Transformers architecture. Whisper has demonstrated human-level performance and robustness, making it a valuable tool for accurately transcribing audio data. For image data extraction, we use Vision Transformer (ViT) [2], which enables the extraction of visual features and image descriptions. ViT has shown remarkable capabilities in handling image data by transforming images into sequences of tokens, allowing efficient processing and extraction of relevant information.

2.2. Extraction and Semantic representation module

The goal of this module is to extract information from previously compiled and indexed documents using an ontology that includes concepts related to retail management. This ontology was created by merging and adapting existing ontologies and applies to different subdomains related to retail management. The retail domain includes product concepts and types, distribution areas and their relationship with external agents. This ontology has been enriched with the information about products and services that each organization has in its own database. For example, when a supermarket wants to use the system, its entire product catalog is loaded into the ontology to identify these concepts as possible aspects to be discovered.

Relevant concepts and their relationships are identified from the documents using Stanza's modules for Dependency Parsing, Named Entity Recognition (NER), Part-of-Speech Tagger (PoS) [3]. In addition to ontology enrichment, KnowGL [4] is used and the dataset is extended by merging Wikidata with an extended version of the REBEL dataset [5]. More information about this system can be found at [6]. Once the system The concepts of the ontology are linked to the collected data using extended TF–IDF [7].

Another stage if topic extraction. For this, we developed a Topic Modeling model based on KeyBERT[2] and BERTopic [8]. To do this, we first build a dataset from the collected evidence, including news and social networks related to retail management. These documents are stored in markdown format to preserve information about the structure of the document, such as titles and divisions by content sections. We preprocess the dataset by removing the markdown tags as external links. We then apply a stemming process and use Stanza's PosTagger to remove non-relevant grammatical information. Once the dataset is assembled, we train the topic modeling model as follows. First, we encode each document as an embedding using Sentence Transformers [9] models such as `paraphrase-spanish-distilroberta`. This representation allows us to encode documents as dense vectors that store the relationships between semantically similar entities. Second, we apply a dimensional reduction process using UMAP [10]. Third, we use HDBSCAN [11], which is a density-based clustering algorithm. Fourth, we extract the topics from the clusters. using a customized variant of TF–IDF.

[2] https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea

2.3. Subjective data analysis module

The subjetive data analysis module consists of two main components: (1) an aspect-based multimodal emotion analysis system and (2) and an author profiling model.

The aspect-based multimodal emotion analysis component focuses on identifying and analyzing emotions expressed within specific topics. Aspect-based emotion analysis is a valuable asset in product reviews, as it facilitates the identification of emotions expressed toward different features or attributes of the product, such as its performance, design or usability among others. We first build a new multimodal dataset by merging existing textual datasets such as EmoEvalEs [12] with custom multimodal data extracted from social networks such as YouTube, where audio and its transcription have been extracted using Whisper [1]. Next, we train a multimodal emotion analysis model using textual features extracted from MarIA [13] and acoustic features from Wav2Vec [14]. Finally, we use the extraction and semantic representation module to determine the aspects. We use this approach because ontologies have been shown to be effective in aspect-based sentiment analysis in the past [15, 16].

The Author Profiling (AP) component is used to extract demographic traits from the authors of the compiled documents in order to obtain a better segmentation of the compiled data. The demographic traits are the age range, gender and location. For this purpose two novel datasets are developed for conducting AP in Spanish: the (1) Spanish IncluCorpus 2023 and the (2) Spanish CCAACorpus 2023. Both datasets are compiled from Twitter using the UMUCorpusClassifier tool [17]. From the first list of users, we selected those whose accounts were public and had published at least 100 tweets. Next, we manually checked the users to ensure that only those users were included whose location in their profile could be recognized as a place in Spain, or whose gender could be recognized by analyzing their profile name or description. In the case of gender, we first checked to see if the user was gender non-binary. If they could not be classified as such, we tried to classify them as male or female. For the former, a list of words and phrases associated with non-binary gender was created. For example, it is common for people of this gender to use the pronoun

"elle" in their description. If any keyword from the list appeared in the user's description, they were classified as non-binary gender. For the second case, a dictionary of male and female names was created by consulting various websites that suggest names for babies according to gender. Names that were considered unisex were discarded and, analogous to the previous case, if the user's name matched one in the dictionary, it was classified as the appropriate gender. To train the author profiling modules, we first merge several existing datasets based on authorship analysis and then extract user-level features based on sentence embeddings from MarIA [13] and linguistic features [18].

2.4. Regression and classification modules

In this module, several machine learning models focus on demand forecasting. These models include classification and clustering models to classify trends and group similar products based on a set of patterns and characteristics. On the other hand, regression models are used to estimate the optimal demand for a given product in a given situation. Explainable Artificial Intelligence (XAI) techniques [19] will help to interpret the results of the machine learning models.

On the one hand, classification models are used to predict the class or category to which the data belongs. For example, they can be used to predict whether or not a product belongs to a particular category. On the other hand, clustering models are used to group similar data into clusters or groups, i.e., products that share similar characteristics or exhibit similar patterns of behavior on various dimensions, such as sales characteristics, prices, or similar product characteristics. The models evaluated include decision trees, support vector machines and logistic regressions, as they are more interpretable than deep neural networks. For regression models, on the other hand, various time series-based models were evaluated to model a dependent variable over time and other independent variables. In addition, external variables such as weather, current economic conditions, commodity prices, or other exogenous variables were included to improve the prediction of future demand trends. The models evaluated include LSTM, ARIMA, and Prophet.

2.5. Dashboard

The final system will be available on a web dashboard. This dashboard will allow the creation and configuration of campaigns in which managers and stakeholders will be able to define the products, websites and social networks to be monitored, as well as other spatio-temporal criteria. Once configured, all the data will be displayed on a dashboard composed of several semantic KPIs, which are independent and autonomous components capable of reading and processing data periodically and whose output can also be configured, including graphs or tables, among others. In addition, options will be implemented to allow users to export the data obtained. Figure 2 displays a screen capture of the dashboard.

The system also allows personalized alerts to be sent through multiple channels, such as email, when the data exceeds a configurable threshold.

3. Future work

The project will be completed during 2024 and most of the goals have been successfully achieved. However, there are some improvements that can be made in the near future.

On the one hand, future work for this project includes several key areas aimed at enhancing the retail management decision support platform. First, there is a focus on expanding data sources to include emerging social media platforms, additional open data repositories, and real-time market data feeds. This broader data integration will provide retailers with a more comprehensive view of market trends and consumer sentiment. Second, the project aims to refine and enhance sentiment analysis capabilities, including multilingual support. These advances will enable the platform to extract more valuable insights from textual data.

Meanwhile, the platform's user interface and user experience are being improved based on user feedback and usability testing. This includes providing customizable dashboards to meet the diverse needs of retail professionals. Integration with third-party services and APIs is also being explored to extend the platform's usefulness, such as integration with e-commerce platforms, marketing automation tools, and customer relationship management (CRM) systems. Finally, scalability and performance optimizations will be critical as the volume of data processed increases, involving distributed computing techniques, database query optimizations, and the use of cloud infrastructure to ensure efficiency and responsiveness.

Acknowledgments

This work was funded by the Spanish Government, Ministerio para la Transformación Digital y la Función Pública through the "Recovery, Transformation and Resilience Plan" and also funded by the European Union NextGenerationEU/PRTR through the research project 2021/C005/00149877.

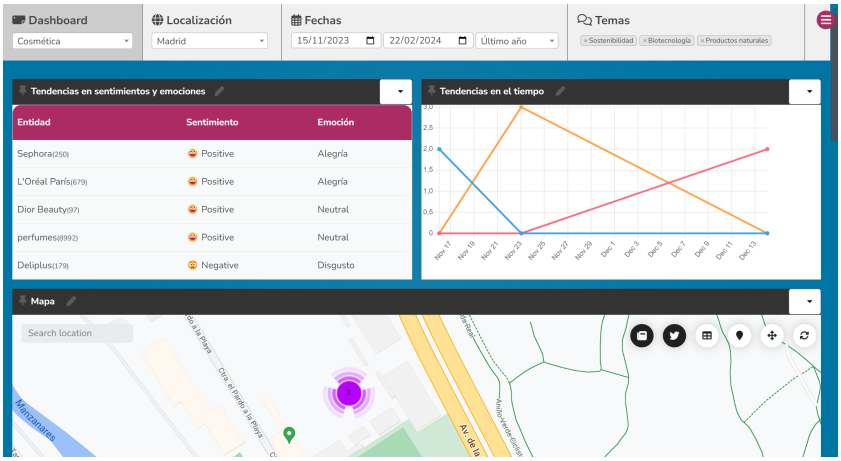

Figure 2: Dashboard

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, arXiv preprint arXiv:2212.04356 (2022).

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[3] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, arXiv preprint arXiv:2003.07082 (2020).

[4] G. Rossiello, M. F. M. Chowdhury, N. Mihindukulasooriya, O. Cornec, A. Gliozzo, Knowgl: Knowledge generation and linking from text, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[5] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.

[6] R. Pan, J. A. García-Díaz, D. Roldán, R. Valencia-García, Knowledge graph for retail commerce, in: International Conference on Technologies and Innovation, Springer, 2023, pp. 173–185.

[7] M. Á. Rodríguez-García, R. Valencia-García, F. García-Sánchez, J. J. S. Zapater, Creating a semantically-enhanced cloud services environment through ontology evolution, Future Generation Comp. Syst. 32 (2014) 295–306. URL: https://doi.org/10.1016/j.future.2013.08.003. doi:10.1016/j.future.2013.08.003.

[8] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[9] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020. URL: https://arxiv.org/abs/2004.09813.

[10] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[11] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering., J. Open Source Softw. 2 (2017) 205.

[12] F. M. Plaza-del Arco, S. M. Jiménez Zafra, A. Montejo Ráez, M. D. Molina González, L. A. Ureña López, M. T. Martín Valdivia, Overview of the emoevales task on emotion detection for spanish at iberlef 2021, Procesamiento del Lenguaje Natural (2021).

[13] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Maria: Spanish language models, arXiv preprint arXiv:2107.07253 (2021).

[14] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[15] M. del Pilar Salas-Zárate, R. Valencia-García, A. Ruiz-Martínez, R. C. Palacios, Feature-based opinion mining in financial news: An ontology-driven approach, J. Inf. Sci. 43 (2017) 458–479. URL: https://doi.org/10.1177/0165551516645528. doi:10.1177/0165551516645528.

[16] J. M. Ruiz-Martínez, R. Valencia-García, F. García-Sánchez, et al., Semantic-based sentiment analysis in financial news, in: Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web, 2012, pp. 38–51.

[17] 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.

[18] 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.

[19] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, B. Sikdar, A review of trustworthy and explainable artificial intelligence (xai), IEEE Access (2023).