Intelligent Radar for Aragonese Tourism

Rosa M. Montañés-Salas[1], Paula Peña-Larena[1], María del Carmen Rodríguez-Hernández[1], Luis García-Garcés[1], Pablo Pérez-Benedí[1], Sergio Mayo-Macías[1], Enrique Meléndez-Estrada[1], Rafael del-Hoyo-Alonso[1] and José Luis Galar-Gimeno[2]

[1]*Aragon Institute of Technology (ITAINNOVA), María de Luna, 7–8, Zaragoza, Spain*
[2]*Tourism of Aragon, Avda. Ranillas, 3A, 3rd floor, Office 3D, Zaragoza, Spain*

Abstract
This paper describes the background and architecture of the Intelligent Radar for Aragonese Tourism (*RITA*), a data-driven social media surveillance system, which aims to enhance the tourist experience by helping the Aragonese government to make informed data-driven decisions and therefore empowering its tourism sector. The system is built over a customizable platform that integrates multiple data mining techniques to collect, clean, process and extract explicit and implicit knowledge from various sources such as social media networks, web pages, RSS feeds and structured data files. RITA employs state-of-the-art Natural Language Processing technologies combined with data analysis and modelling techniques to analyse social perception of the region and link that information with organizational data. The platform integrates pre-trained and fine-tuned language models based on transformers architectures for solving different NLP tasks including opinion and emotion analysis, semantic classification and entities recognition. The knowledge gathered is made available to the tourism professionals via an interactive and customizable web application.

Keywords
Natural Language Processing, Transformers, Social Media, Tourism

1. Introduction

Social media networks are an integral part of human daily life, they have positioned as one of the most popular forms of communication, entertainment and social connection [1]. People generates and shares almost any type of content regarding their opinions, interests, experiences and desires, which constitutes a huge and invaluable source of data and knowledge about the human behaviour. The tourism domain may be highly benefited from the amount of information shared publicly among citizens by being provided with appropriate data intelligence tools to discover at first hands and analyse tastes, inclinations and concerns of tourists and other interested groups of people ([2], [3], [4]).

Multiple Natural Language Processing (NLP) challenges, such as multilingualism, the use of formal and informal expressions, ambiguity and others, must be faced when using social media data for human-related research. These challenges, coupled with the multimodal nature of user-generated data, may be better addressed by combining cutting-edge techniques with conventional methods. In this context, ITAINNOVA has re-designed and evolved the solution showcased in [5] that was conceived as a unified self-monitoring system for a particular user, a place where any individual could stay updated about its virtual social environment by thoroughly analysing the virtual interactions and extracting implicit knowledge from them. The capability to extract and organize valuable information from social networks is primarily enabled by the use of various natural language processing techniques, such as semantic categorization, entity extraction and opinion inference. Moving the focus from a single user to a more professional setting, makes the system applicable as a working tool for any public or private corporation to empower their decision-making processes. The main objective pursued is to develop a multimodal data-driven social intelligent platform.

Both private companies and public administrations are recognizing the imperative to integrate digitalization and artificial intelligence (AI) tools into their organizational processes. Given the significance of the tourism sector in Aragón and the desire to enhance its robustness, as well as gain profound insights into its users and potential visitors, Aragonese Tourism embarked on a collaborative endeavour with ITAINNOVA to develop a decision support tool based on the data-driven social platform. The

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations
rmontanes@itainnova.es (R. M. Montañés-Salas); ppena@itainnova.es (P. Peña-Larena); mcrodriguez@itainnova.es (M. d. C. Rodríguez-Hernández); lggarcia@itainnova.es (L. García-Garcés); pperez@itainnova.es (P. Pérez-Benedí); smayo@itainnova.es (S. Mayo-Macías); emelendez@itainnova.es (E. Meléndez-Estrada); rdelhoyo@itainnova.es (R. del-Hoyo-Alonso); jlgalar@aragon.es (J. L. Galar-Gimeno)
0000-0003-4636-5868 (R. M. Montañés-Salas); 0000-0001-5750-6238 (P. Peña-Larena); 0000-0002-0062-9525 (M. d. C. Rodríguez-Hernández); 0000-0002-7485-7600 (P. Pérez-Benedí); 0000-0002-3461-4700 (S. Mayo-Macías); 0000-0002-8385-4954 (E. Meléndez-Estrada); 0000-0003-2755-5500 (R. del-Hoyo-Alonso); 0000-0001-6592-7067 (J. L. Galar-Gimeno)

primary objective of this system is to aggregate citizens information from diverse sources and modalities. While surveys and traditional statistical analysis offer precise and valuable data, it is essential to acknowledge that individuals convey a wealth of information through spoken and written language, specially through open social media platforms, which requires the use of advanced AI and NLP techniques into the decision-making process [6]. RITA emerged from the need to better understand the needs and expectations of the tourism sector in Aragon.

In this paper, the social media monitoring system for the Aragonese tourism is presented following this structure: after the introduction, a description of the general platform designed is outlined. In section 3, the final system developed on top of the social media platform is presented. And the last section concludes with an overview of the main outcomes achieved, as well as suggestions for further improvements.

2. Social media platform

ITAINNOVA has designed a highly scalable and customizable system aiming to fit the needs of current social and marketing research, named *Social Media* platform. The long-term objective is to work as a base architecture to integrate heterogeneous data sources, which are curated, fused and modelled by means of advanced artificial intelligence techniques, deploying a wide range of smart services and allowing to develop data resources including datasets and in-domain knowledge models (large language models, multimodal models and data models). The operational architecture is shown in figure 1.

This platform is designed with the vision of serving as a decision support tool and as a domain knowledge repository. The core of the system corresponds to the block identified as *Social-Media* together with all the data and models repositories displayed in the lower part. Social-Media implements and exposes a series of microservices which are consumed by the client module ("Client API"), responsible for adapting and executing the customized processes for the specific domain.

As a general guideline, the platform is designed over the data mining main pillars: gather useful information; process, clean and obtain relevant data; analyse and model the explicit and implicit knowledge contained in the data; and disseminate results. An overview of the implementation of these steps is presented in this section.

2.1. Information retrieval

The first step in the development is to properly establish the focus of the information to be retrieved and analysed, i.e. to determine which information sources will be included in the system, the specific domain of the information to be retrieved, the data types supported and the feasibility of their inclusion.

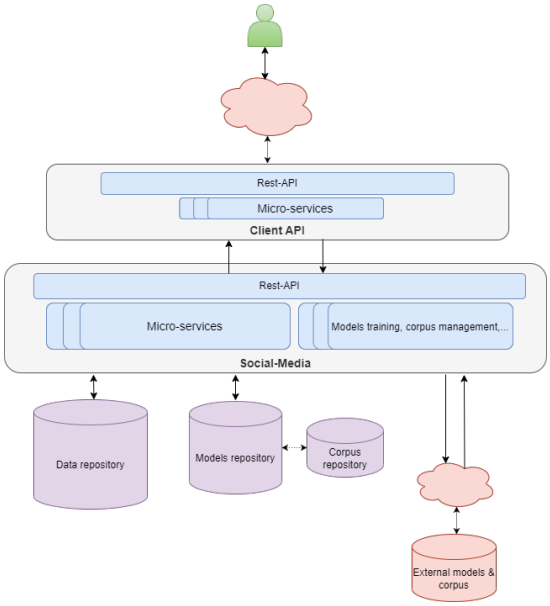

Figure 1: Operational architecture.

The current state of development of the platform supports intake of textual and numerical data from diverse sources: social media networks, web pages, RSS feeds and structured data files. The primary social network integrated is Twitter, which exposes a public API composed of a set of endpoints that permit to explore and manage Twitter entities, depending on the access level authorized. A monitoring system will typically require tracking of the most recent events in the domain under research, thus, by default, standard endpoints will be queried.

In order to store all this information, a homogenized data model has been designed on a document-oriented database engine. This model is composed of two different entities: the first one called "social-networks" in which the information on publications from the sources described above (data and metadata) is structured and stored, in which the differential concepts of these sources have been homogenized; and the second one called "users" in which the public information of the authors of the publications is linked.

2.2. No-NLP: Not only NLP

The following phases of data cleaning, processing and modelling are crucial in order to obtain useful information, implicit knowledge, patterns and trends on large

amounts of data. At this stage, two main approaches are considered: applying different state-of-the-art algorithms on the core data: the text interactions, and then exploit relations with the rest of the analytical data to deduce higher level associations. Therefore, primarily transformer-based NLP techniques have been applied to analyse textual content [7], identify salient elements and infer information at the semantic level, as described following.

After the extraction of social media content, a prefiltering step is performed based on the keywords configured for searching and the metadata retrieved, applying specific domain rules. Over the most relevant candidates obtained from this filtering process, a data processing pipeline is applied for extracting structured and meaningful information, as depicted in figure 2.

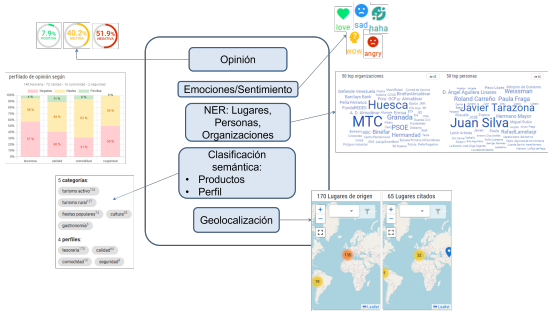

Figure 2: Information processing.

Opinion analysis tries to identify and extract subjective information expressed in human linguistic production (spoken and written), that is, it tries to determine the attitude of the interlocutor with respect to the general topic or particular aspects of a message. This attitude can be classified in different ways, in this case a three-level categorical evaluation has been established: positive, negative and neutral. A multilingual XLM-RoBERTa fine-tuned model for twitter sentiment analysis [8] has been integrated in the system for this submodule.

Emotion analysis: alongside the previous analysis, it is feasible to perform an analysis of the emotions and feelings expressed over a topic in written and verbal communication [9]. In this case, the model has to distinguish more precise sentiments and assign one or several corresponding tags to the documents. The tags selected are: like, love, haha, wow, sad and angry, inspired in Facebook's reactions options. A Spanish multilabel dataset, consisting of more than two thousand documents retrieved from social networks in the past, was used to fine-tune the multilingual BERT [10] developed in this module.

Named-Entity Recognition (NER) consists of automatically extracting from the textual sequence a set of terms of interest referring to a given concept (entity). Depending on the working domain, the entities defined may vary, but for a general perspective, a standard NER on places, organizations and persons is applied. An already fine-tuned multilingual BERT entity recognizer is integrated in the system. In this task, both the sentence's grammar and lexical ambiguity must be taken into account in order to get valid results. To overcome possible mistakes produced on the lexical level, the acknowledged approach based on gazetteers is still considered at the end of the NER processing, by making use of the internal and external domain-knowledge available.

Semantic classification refers to the task of assigning one or more predefined categories to a statement according to its overall meaning and the subject to which it refers. This classification will make it possible to group documents and organize large amounts of information regarding their semantic interpretation. In the first user-oriented version of the monitoring system, ontology-based strategies were explored and implemented. They allow collecting terms and concepts from specific domains in a structured way that can be explored during text analysis by means of different algorithms, enabling the assignation of a conceptual category to that document or profiling users [11]. Results obtained were quite reasonable, although this approach required a periodic update of the knowledge bases used (ontologies, dictionaries, thesauri, lists, etc.) as the data sources change over time. To improve this process, leveraging the power of Transformers that exploit to the maximum the concept of *transfer learning* [12] was studied and tested, including finally a multilingual *zero-shot classifier* fine-tuned on the XLM-RoBERTa-Large pre-trained model [13] on a combination of data from 15 languages, such as English, French, Spanish, German, Hindi, etc.

Geocoding: the functionalities of the platform include a geolocation service that allows finding the information associated with a location, expressed through its coordinates or through its usual name (i.e.: *Sos del Rey Católico*). This component is implemented through queries to the open source service Nominatim[1], which makes use of OpenStreetMap data. The names of locations, from which its geo-positioning is desired, are retrieved from the textual publications mentions by invoking the NER service.

Further data analysis is applied at the end of the data processing pipeline. Statistical analysis on numerical data extracted from the social networks is performed, by means of computing standard metrics including means, deviations, comparison among different variables sample sizes, text frequencies, common pattern and so forth. Additionally, these techniques are applied over different cross-data sets as well, as for example the set obtained

[1] https://nominatim.org/

by joining the number of documents which are speaking about a topic with the opinion the user is expressing about that topic.

2.3. Data exploitation

The results produced during the previous stages should be analysed and shared with the interested parties. The Social-Media platform is designed to be highly flexible by exposing a set of web-services through a simple REST-API that can be invoked from any compatible client that enables to integrate seamlessly the results on a web application, external business intelligence systems or in client reports, depending on the use case defined by the user.

3. *RITA*

The acronym *RITA* comes from its Spanish initials: *Radar Inteligente de Turismo de Aragón* (Intelligent Radar for Aragonese Tourism) which stems from the collaboration between the Aragon Institute of Technology (ITAINNOVA) with the Society for the Promotion and Management of Aragonese Tourism (Aragonese Tourism organization from now on). Both entities are cooperating on researching and executing new innovative tourism strategies based on digitalization for the development and improvement of the tourist sector in their region. Listening to the social needs and tastes directly and in a non-intrusive way, letting them express freely and comfortably from any place, can only be reached by consulting public intercommunication spaces such as the Internet and social media networks. Fusing that information with data retrieved from tourism offices, touristic places visiting reports and open data knowledge, would lead experts to comprehend thoroughly which aspects influence the current touristic panorama and which ones should be attended to then design and offer new tourist experiences. With this aim, Aragonese Tourism corporation and ITAINNOVA begin the designing and development of their intelligent radar.

3.1. Tourism domain

For the use case and at the current phase, two information sources have been considered: social networks (Twitter) and web pages (blog posts, RSS feeds).

Pre-filtering and post-filtering stages are adapted to this particular domain: content related to Aragonese towns, regions and touristic places. In this sense, metadata is examined in order to check whether the query has returned mistaken publications and discard data in that case. Moreover, NER locations identified and geolocated are matched against searches configured and filters those that appear clearly out of the scope.

Opinion and emotion analysis are maintained as explained in section 2, while the NER results are curated in order to extract more precise results. Entities extracted for each class are automatically reviewed applying some in-domain rules devised by tourist specialists: places identified are geolocated and matched with information on villages and regions; likewise, people names retrieved are examined in detail, as it was found that a number of surnames match locations names in Spanish; moreover, all the entities are filtered in terms of character length and social network related stopwords in order to discard meaningless information.

Semantic classification is particularly customized for the use case. Two different criteria are considered to structure the content based on its meaning: products and profiles. The group of *products* identifies the typology of the tourist offer mentioned, they refer to a combination of places, festivities, activities, natural resources, material and immaterial attractions, etc. The following products are defined within *RITA*: active tourism, rural tourism, popular festivities, culture and gastronomy. *Profiles* refer to the characteristics of interest that resembles the attitude and aspects that tourists demand when visiting different places in the autonomous community. The technical experts from Aragonese Tourism corporation have selected this labels: safety, comfort, treasury and quality. The zero-shot classifier fits these classification needs with ease, as up to our knowledge, there was no public, multilingual or Spanish, dataset available to fine-tune a language model classifier and the labels considered may vary regarding the context.

3.2. Web interface

A web application has been designed and published, which includes a dashboard displaying all the processed information and different interactive components that allow filtering the results to visualize the desired contents in an interactive way. It includes a timeline of publications that can be sorted by date, relevance, number of "likes" and number of retweets. On one side there is a set of possible filters to be applied: dates, sources, opinion, categories, profiles and tags (regions and municipalities searched in social networks) as depicted in figure 3. On the other side are displayed different statistics of users who have published along with several graphs that allow to know the average opinion of the publications, the evaluation of the aspects considered from the tourist point of view (profile), most used hashtags, the identified places where the publications have been made and which places have been mentioned, as well as organizations and people detected in the texts in form of word clouds (see figure 4).

The use of the application is intended for technical and management staff of the Aragonese Tourism corpo-

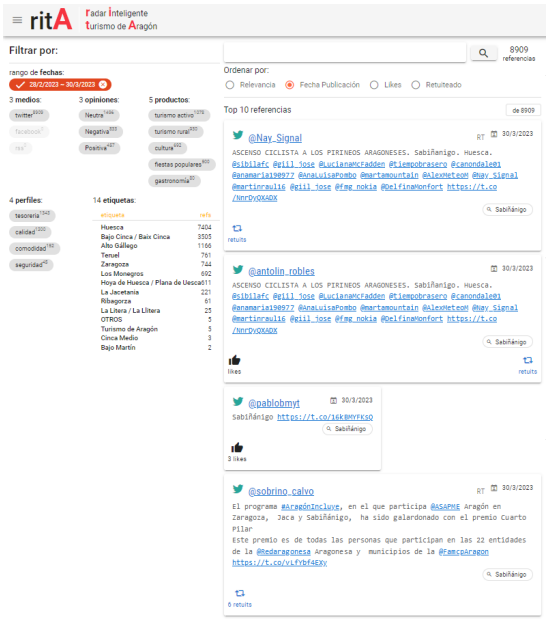

Figure 3: RITA's web interface: timeline and filters.

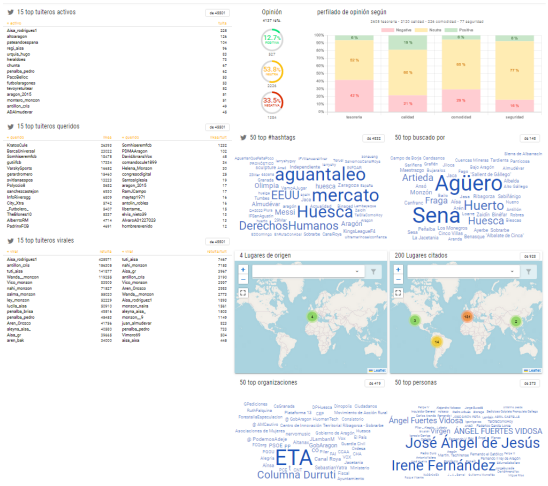

Figure 4: RITA's web interface: data-driven results.

ration, to be incorporated as a dynamic consulting tool where they might explore news from their action area, generate periodical reports, obtain an overall vision of the typology and expectations of potential visitors and, ultimately, get real snapshots of the current situation of the sector in Aragon from more personal perspectives.

4. Conclusions and future work

The Intelligent Radar for Aragonese Tourism, *RITA*, has been introduced in this paper. It is built on top of the Social-Media platform designed by ITAINNOVA, which aim is to evolve to a multimodal data-driven social intelligent platform used to support and enhance any kind of advanced decision-making tools. The radar emerges as a digital solution to empower the tourist sector in Aragon, with the objectives of monitoring social perception of the region and its touristic attractions, and link that information with organizational data by applying natural language processing and other state-of-the-art techniques to detect current situation of the sector and enhance tourist experience. It can be stated that the RITA solution enables the Aragonese government to make informed decisions based on data, which reduces the cost of data acquisition. The platform is capable of gathering valuable information from heterogeneous sources, focused on user-generated content, which provides a more comprehensive understanding of the tourism sector in Aragon. This data-driven approach empowers the government to make smart decisions that can improve the tourism experience and eventually promote the region's growth.

These objectives are fulfilled by the use of a wide range of language technologies combined with data analysis and machine learning techniques. The platform integrates several pre-trained and fine-tuned language models mainly based on transformer architectures for solving different NLP tasks, such as opinion analysis, emotion analysis, named-entity recognition and semantic classification. It also relies on in-domain knowledge in the form of rule-based filters and gazetteers, to improve the insights retrieved from the machine learning processes. Combining different data types and statistical analysis allow capturing implicit relations in the information that feeds the system.

Nonetheless, there is still room for improvement and some future work lines are considered: extending the number and modalities of data sources is the main target defined in the platform roadmap. Adding external and internal data would lay a foundation to incorporate advanced multimodal transformer models for enriching hidden pattern discovery and gather more data which, adequately analysed and modelled, will answer more complex questions to the tourism technical experts. Some types of data being under evaluation are public customer reviews, public accessible reports from the National Statistics Institute and other regional tourist activity reports. With the emergence of generative large language models [14], such as GPT, a new landscape opens up in the data analysis and knowledge inference, which will be leveraged in following updates of the platform. On the end-user side, the web interface will append

new specific dashboards to analyse more easily particular aspects such as the opinions and emotions associated with a certain region or product, temporal statistics and trends or social interactions.

Acknowledgments

This work has been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon, by IODIDE group of the Government of Aragon, grant number T1720R and by the European Regional Development Fund (ERDF).

References

[1] E. Ortiz-Ospina, M. Roser, The rise of social media, Our world in data (2023).

[2] A. Huertas Herrera, M. D. Toro-Manríquez, R. Soler Esteban, C. Lorenzo, M. V. Lencinas, G. Martínez Pastur, Social media reveal visitors' interest in flora and fauna species of a forest region, Ecosystems and People 19 (2023) 2155248.

[3] R. Nunkoo, D. Gursoy, Y. K. Dwivedi, Effects of social media on residents' attitudes to tourism: Conceptual framework and research propositions, Journal of Sustainable Tourism 31 (2023) 350–366.

[4] F. J. Lacarcel, R. Huete, Digital communication strategies used by private companies, entrepreneurs, and public entities to attract long-stay tourists: a review, International Entrepreneurship and Management Journal (2023) 1–18.

[5] R. Montanés, R. Aznar, S. Nogueras, P. Segura, R. Langarita, E. Meléndez, P. Pena, R. Del Hoyo, Monitorización de social media, Procesamiento del Lenguaje Natural 61 (2018) 177–180.

[6] N. A. Alghamdi, H. H. Al-Baity, Augmented analytics driven by ai: A digital transformation beyond business intelligence, Sensors 22 (2022) 8071.

[7] L. Tunstall, L. Von Werra, T. Wolf, Natural language processing with transformers, " O'Reilly Media, Inc.", 2022.

[8] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: A multilingual language model toolkit for twitter, CoRR abs/2104.12250 (2021). URL: https://arxiv.org/abs/2104.12250. arXiv:2104.12250.

[9] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of bert-based approaches, Artificial Intelligence Review (2021) 1–41.

[10] 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.

[11] P. Peña, R. Del Hoyo, J. Vea-Murguía, C. González, S. Mayo, Collective knowledge ontology user profiling for twitter–automatic user profiling, in: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), volume 1, IEEE, 2013, pp. 439–444.

[12] R. Qasim, W. H. Bangyal, M. A. Alqarni, A. Ali Almazroi, et al., A fine-tuned bert-based transfer learning approach for text classification, Journal of healthcare engineering 2022 (2022).

[13] 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.

[14] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).