Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 385-389 recibido 22-11-2024 revisado 27-01-2025 aceptado 31-01-2025

Artificial Intelligence methods for Social Science Research: Demographic Analysis, Stance Detection and Political Leaning Inference

Métodos de Inteligencia Artificial para la investigación en ciencias sociales: Análisis demográfico, detección de posturas e inferencia de tendencias políticas

Joseba Fernandez de Landa
HiTZ Center - Ixa NLP Group, University of the Basque Country UPV/EHU
joseba.fernandezdelanda@ehu.eus

Abstract: This is a summary of the Ph.D. thesis conducted by Joseba Fernandez de Landa at the University of the Basque Country under the supervision of Ph.D. Rodrigo Agerri. The thesis defense took place in Donostia on October 16, 2024, with the doctoral committee comprising Ph.D. Estela Saquete from the University of Alicante, Ph.D. Xabier Arregi from the University of the Basque Country, and Ph.D. Viviana Patti from the University of Turin. The thesis was awarded the distinction of Cum Laude and received international recognition.
Keywords: Artificial Intelligence, Social Sciences, Social Media, NLP.

Resumen: Este es un resumen de la tesis doctoral realizada por Joseba Fernandez de Landa en la Universidad del País Vasco bajo la supervisión del Dr. Rodrigo Agerri. La defensa de la tesis tuvo lugar en Donostia el 16 de octubre de 2024, con el tribunal de tesis compuesto por la Dra. Estela Saquete de la Universidad de Alicante, el Dr. Xabier Arregi de la Universidad del País Vasco y la Dra. Viviana Patti de la Universidad de Turín. La tesis obtuvo la distinción de Cum Laude y mención internacional.
Palabras clave: Inteligencia Artificial, Ciencias Sociales, Redes Sociales, PLN.

1 Introduction

In recent decades, significant social changes have taken place, primarily due to technological advances in the fields of information and communication. New ways of human interactions are emerging, breaking spatial and temporal barriers and enabling constant connection with the community, allowing communication everywhere and at any time. These online activities leave a digital trace that can be used to represent the behavior of individuals and communities. This massive digitization of social life presents new opportunities for social research.

However, to fully harness this potential, social scientists must complement and expand their methodologies with approaches developed in data science. Artificial Intelligence (AI) began to offer powerful methodologies for analyzing how social dynamics are created and developed in different contexts. Among the most commonly used methodologies are the classification and generation of texts based on Machine Learning (ML) and Natural Language Processing (NLP), while leveraging pre-annotated data to make projections onto new data.

Simultaneously, social research methods or techniques that were effective in the past are beginning to lose credibility and efficiency in today's dynamic society. For example, political polls have come under scrutiny because they failed to predict several significant events, including Brexit or Donald Trump's presidency.

Under these circumstances, it is clear that new tools are needed to interpret and understand society. It is particularly important to develop research methodologies suited to rapidly changing situations, such as policy making, health crises, or social conflicts.

ISSN 1135-5948 DOI 10.26342/2025-74-26

These methods enable research to be conducted quickly and accurately, especially as the computational power to process and analyze large datasets continues to grow. Moreover, traditional methods often struggle to analyze emerging issues, particularly the increasingly common dynamics of the online world. Additionally, research methodologies tend to be rooted in specific fields, yielding strong results within a particular task or environment. However, adapting these methodologies to other fields or domains can be challenging due to a high dependence on annotated data.

Therefore, this thesis explores research methodologies that combine social research and AI to develop approaches that are more generalizable and accurate than current methods. Specifically, by analyzing virtual interactions between humans and textual expressions, we propose and demonstrate methods to predict demographic or ideological characteristics. To achieve this, various types of data from social networks will be utilized to predict and analyze user characteristics. The research will investigate methodologies for data extraction and annotation, as well as approaches for identifying and representing communities through data analysis. To enable accurate predictions, it is crucial to analyze generalizable methods, dynamic techniques that can be adapted to different contexts, tasks, locations, and even languages.

Based on data from social networks, various social characteristics can be inferred, such as age, gender, or ideology from data as a glance. However, the main goal is to automatically infer these social characteristics by exploring and developing approaches that enable social research. To achieve this, we attempted to use and improve tasks based on text classification, making advancements in language processing as well. This work examines how new computational social research techniques can be developed that can be applied in various situations and tasks, ultimately contributing to a more precise and comprehensible picture of an ever-changing digital society.

1.1 Research Lines

The main goal of this thesis will focus on developing techniques for automatically characterizing social media users. This characterization will be based on the interactions between users and the texts they publish. With this data, predictions about users' demographic and ideological characteristics will be made, with the ability to adapt to different realities and contexts.

Thus, in order to make predictions that can be generalized and are more precise, we will use ML and NLP methods. To achieve this goal, we have followed these steps: (i) In the **exploration** phase, demographic characteristics of users were identified, and community detection was addressed using precise data collection, as well as text classification and user representation methods; (ii) In the **development** phase, a methodology capable of obtaining a general user-level representation across various topics and languages was developed, which creates vector representations of socio-political information from user data; (iii) In the **application** phase, the socio-political information representation was applied to accurately predict users' political tendencies in different locations and situations.

Specifically, the research lines of the thesis are organized as follows:

[L1]: Identification of Demographic Characteristics in the Basque Community. The aim of this research line is to analyze what young Basque speakers discuss and with whom they interact on social media, connecting it to the task of identifying demographic characteristics. Specifically, it focuses on identifying speakers of a particular language, suggesting users' ages, and detecting sub-communities. To this end, the study investigates how to identify users belonging to a group with specific characteristics on the social network Twitter, and how to gather their data. To identify users' age, we used advanced NLP methods for experimentation, analyzing how individuals write and annotating users based on their writing style. For community identification, experiments were executed using unsupervised ML, based on interactions related to shared content. In summary, the study explores how to achieve high-accuracy characterization of users through text and interactions.

[L2]: Stance Detection Across Multiple Languages and Targets. The goal of this research line is to generalize the stance detection task leveraging users' texts and interactions. Given the success of utilizing texts

and interactions for characterizing users, the stance detection task has been extended from simple text sequences to the user level inference. This research aims to develop a robust methodology that can be generalized and applied independently trough different targets and languages. To achieve this, we developed data collection and classification methods based on user representations.

[L3]: Political Leaning Inference at User Level. The representations technique developed in the previous research line is examined for its application in the task of political leaning inference at the user level. The study aims to explore whether the data collection and user representation methodologies previously developed can be applied to other areas. Furthermore, the concept of political leaning is expanded from the typical binary view (left/right or liberal/conservative) to a party-level approach, making it applicable to a wider range of leanings and different political realities. In this way, we developed methods for conducting broader and more accurate socio-political analysis.

2 Thesis Overview

The main content of this thesis is structured into 3 chapters, outlined as follows:

Chapter 3 addresses the **identification of demographic characteristics**. Specifically, it explores how to identify the demographic characteristics of social media users and their communities (L1).

Chapter 4 focuses on **stance detection**. To this end, a language-independent stance detection method is developed, applying the task at the user level (L2). This involves addressing user-level data collection and feature extraction based on user interactions.

Chapter 5 explores **political leaning inference**. Here, dynamic identification of users' political leaning is analyzed (L3). Based on interactions, binary and multiparty political approaches are compared, and later, different levels of political involvement are studied. Finally, under these conditions, interaction- and text-based approaches are tested to examine the impact of data hybridization.

3 Contributions

In this thesis we have developed data collection and user representation methods to perform more accurate and generalizable social research. We demonstrate that hybrid methods, based on text and user interactions, are beneficial for a number of text classification tasks, including stance detection and political learning inference. More specifically, and in relation to the different research lines outlined previously, the main **contributions** of this thesis are the following:

- We proposed a new methodology to investigate how AI can be applied to social research. Thus, we explored data collection and labeling techniques, as well as classification models to infer social media user's demographic traits. In order to do so, we first collected *Heldugazte-osoa*[1] corpus, consisting of 6M publications in Basque, enabling further analysis of this under-resourced language. Second, we annotated *Heldugazte*[2] and *Heldugazte-age*[3] datasets to infer the life stage of Basque users (young or adult). Leveraging these datasets, we developed and evaluated different methods for life stage classification, including experiments with monolingual and multilingual Transformer-based language models. Third, we applied our methods to the raw corpus to qualitatively analyze and evaluate their performance in real-world scenarios. Finally, we empirically demonstrated the potential of interactions to convey socio-political information (Fernandez de Landa, Agerri, and Alegria, 2019; Fernandez de Landa and Agerri, 2021b). This contribution corresponds to the **L1** research line.
- We collected and annotated the **VaxxStance dataset**[4], a comprehensive public dataset designed for stance detection on the vaccines topic. This dataset includes both text and interaction data in two different languages (Basque and Spanish), centered on the same topic. Therefore, we encourage experimentation using both social and textual features in multilingual and crosslingual settings (Agerri et al., 2021). This contribution corresponds to the **L2** research

[1] ixa2.si.ehu.es/heldugazte-corpus/heldugazte.osoa.tar.gz

[2] github.com/ixa-ehu/heldugazte-corpus

[3] github.com/joseba-fdl/heldugazte-age-corpus

[4] https://vaxxstance.github.io/

line.

- We presented a novel method, **Relational Embedding**, to represent and exploit interaction data based on one-to-one relations, such as *friends* and/or *retweets*. We experimented on seven publicly available stance detection datasets, showing that our method behaved robustly across various targets and languages without any specific manual engineering. Furthermore, combining our method with textual data systematically improved the results, outperforming even ensembles of large pretrained language models (Fernandez de Landa and Agerri, 2022). This contribution is related to the **L2** research line.
- We proposed a **multi-party framework** to better capture political leaning based on institutional political parties, which proves adaptable to different regions since it is grounded on localized political actors. We annotated a dataset containing labeled users by political party and left-right orientation alongside their retweets from the regions of Basque Country, Catalonia and Galicia. Subsequently, comprehensive experimentation with multi-party (7 political parties) and binary (left-right) frameworks showed that Relational Embeddings outperform other user representation methods even with scarce training data (Fernandez de Landa and Agerri, 2024). This contribution refers to the **L3** research line.
- We delved into political learning inference based on the previously described multi-party framework by focusing on **different levels of user's political engagement**. We annotated another dataset containing labeled users by political party alongside their retweets from the regions of Wales, Scotland and Northern Ireland. Therefore, we include three datasets per region regarding different levels of implication with a political party: members, supporters or sympathizers. We evaluate a range of methodologies to make the most of retweet interactions among social media users to infer their political leaning, showing again that Relational Embedding based approach is effective even along the different levels of engagement (Fernandez de Landa, Zubiaga, and Agerri, 2023). This contribution corresponds to the **L3** research line.
- Finally, we addressed political learning inference leveraging users' texts and interactions, demonstrating that interaction-based Relational Embeddings outperformed state-of-the-art textual approaches. In addition, we proposed **hybrid modeling** of social media users merging text and interaction features, showing that the combination of both data types is required for optimal performance (Fernandez de Landa and Agerri, 2023; Fernandez de Landa, Zubiaga, and Agerri, 2024; Fernandez de Landa et al., 2024). This contribution is related to the **L3** research line.

During the development of the thesis, we also contributed to the generation of valuable resources and findings for the Basque language (Fernandez de Landa, Alegria, and Agerri, 2019; Fernandez de Landa, 2019; Fernandez de Landa et al., 2021; Salaberria et al., 2021; Fernandez de Landa and Agerri, 2021a). The aforementioned resources will be hosted in the Clariah-eus strategic network with the aim of contributing to social sciences and humanities for the Basque language and culture (Alkorta et al., 2024a; Goenaga et al., 2024; Alkorta et al., 2024b).

Acknowledgments

This research has been supported by a predoctoral scholarship granted by the University of the Basque Country UPV/EHU (UPV/EHU-PIF19/208). We are also thankful to the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR.

References

Agerri, R., R. Centeno, M. Espinosa, J. Fernandez de Landa, and A. Rodrigo. 2021. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Alkorta, J., A. Farwell, J. Fernandez de Landa, B. Altuna, A. Estarrona, M. Iruskieta, X. Arregi, X. Goenaga, and J. M. Arriola. 2024a. Clariah-eus: a cross-border clariah node for the basque language and culture. In *Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024)*. CEUR-WS.org.

Alkorta, J., A. Farwell, J. Fernandez de Landa, B. Altuna, A. Estarrona, M. Iruskieta, X. Arregi, X. Goenaga, and J. M. Arriola. 2024b. Clariah-eus: A strategic network helping basque country researchers to participate in european research infrastructures. In *CLARIN Annual Conference Proceedings, 2024*.

Fernandez de Landa, J. 2019. Gazteak eta euskara sare sozialetan. zer, nori, nork: euskarazko txio formal eta informalak sailkatuz eta konparatuz. *Eusko Ikaskuntzaren XVIII. Kongresua Geroa Elkar-Ekin: Mendeurreneko Kongresua*, (18):348–355.

Fernandez de Landa, J. and R. Agerri. 2021a. Euskarazko on-line artikuluetan aipatutako izendun entitate nabarmenen identifikazioa denbora errealean. *EKAIA EHUko Zientzia eta Teknologia aldizkaria*, (40).

Fernandez de Landa, J. and R. Agerri. 2021b. Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

Fernandez de Landa, J. and R. Agerri. 2022. Relational embeddings for language independent stance detection. *arXiv preprint arXiv:2210.05715*.

Fernandez de Landa, J. and R. Agerri. 2023. Hitz-ixa at politices-iberlef2023: Document and sentence level text representations for demographic characteristics and political ideology detection. 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.

Fernandez de Landa, J. and R. Agerri. 2024. Political leaning inference through plurinational scenarios. *arXiv preprint arXiv:2406.07964*.

Fernandez de Landa, J., R. Agerri, and I. Alegria. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6):212.

Fernandez de Landa, J., I. Alegria, and R. Agerri. 2019. Euskaldun gazte eta helduen harremanak twitterren. *III. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Gizarte Zientziak eta Zuzenbidea*.

Fernandez de Landa, J., I. García-Ferrero, A. Salaberria, and J. A. Campos. 2024. Uncovering social changes of the basque speaking twitter community during covid-19 pandemic. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 363–371, Torino, Italia. ELRA and ICCL.

Fernandez de Landa, J., I. García Ferrero, A. Salaberria Saizar, and J. A. Campos Tejedor. 2021. Twitterreko euskal komunitatearen eduki azterketa pandemia garaian. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura*.

Fernandez de Landa, J., A. Zubiaga, and R. Agerri. 2023. Generalizing political leaning inference to multi-party systems: Insights from the uk political landscape. *arXiv preprint arXiv:2312.01738*.

Fernandez de Landa, J., A. Zubiaga, and R. Agerri. 2024. Htim: Hybrid text-interaction modeling for broadening political leaning inference in social media. *arXiv preprint arXiv:2406.08201*.

Goenaga, X., A. Farwell, J. Fernandez de Landa, and X. Arregi. 2024. Constructing the clariah-eus clarin b-centre: First steps. In *CLARIN Annual Conference Proceedings, 2024*.

Salaberria, A., J. A. Campos, I. Garcıa, and J. Fernandez de Landa. 2021. Itzulpen automatikoko sistemen analisia: Genero alborapenaren kasua. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura*.