Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 235-238 recibido 16-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani

Enfoques de aprendizaje automático para el análisis de sentimientos y temas en opiniones multilingües y en idiomas con escasez de recursos: Del inglés al guaraní

Marvin Matías Agüero-Torales
University of Granada, Spain
maguero@correo.ugr.es

Abstract: The following is a summary of a Ph.D. thesis written by Marvin Matías Agüero-Torales at the University of Granada under the supervision of Ph.D. Antonio Gabriel López-Herrera. The author was examined on Friday, February 4th, 2022 by a committee composed of Ph.D. Enrique Herrera-Viedma and Ph.D. Carlos Gustavo Porcel Gallego from the University of Granada, Ph.D. María José del Jesús Díaz and Ph.D. Salud María Jiménez-Zafra from the University of Jaén, and Ph.D. Jesús Serrano-Guerrero from the University of Castilla-La Mancha. The Ph.D. thesis was awarded the Summa Cum Laude mention.
Keywords: natural language processing (NLP), machine learning, code-switching, low-resource languages.

Resumen: Este es un resumen de la tesis doctoral realizada por Marvin Matías Agüero-Torales en la Universidad de Granada bajo la dirección del doctor D. Antonio Gabriel López-Herrera. La defensa de la tesis se llevó a cabo el viernes 4 de febrero de 2022 ante un tribunal formado por los doctores D. Enrique Herrera-Viedma y D. Carlos Gustavo Porcel Gallego de la Universidad de Granada, Dña. María José del Jesús Díaz y Dña. Salud María Jiménez-Zafra de la Universidad de Jaén, y D. Jesús Serrano-Guerrero de la Universidad de Castilla-La Mancha. La tesis recibió la calificación de Sobresaliente Cum Laude por unanimidad.
Palabras clave: procesamiento de lenguaje natural (PLN), aprendizaje automático, code-switching, idiomas con escasez de recursos.

1 Introduction

In recent years, the internet, especially social media, has become the main source of information, with people sharing their opinions, beliefs, emotions, and experiences online. Researchers from various fields, particularly Natural Language Processing (NLP), have been interested in analyzing this web content. NLP involves using computational techniques to analyze and synthesize natural language, especially text found on the web.

This doctoral thesis proposes using machine learning techniques to analyze opinions in low-resource languages, including Spanish, Guarani, and Jopara, in monolingual, multilingual, and code-switching settings. These techniques include sentiment analysis, which aims to identify the sentiment expressed in a text, and topic modeling, which aims to identify the main topics in a collection of texts.

In this thesis, we followed the path of text mining and NLP at the intersection of computation, artificial intelligence, and computational linguistics, focusing on multilingualism in low-resource languages. Our objectives are to (i) investigate different machine learning approaches that can handle multilingual opinions written in social media (even code-switching), particularly those based on neural networks, (ii) create new linguistic resources for analyzing text in low-resource languages and dialects, particularly those found on social media, and (iii) develop machine learning models for NLP in low-resource languages and dialects in monolingual, multilingual, and code-switching settings. The research aims to gain a deeper understanding

ISSN 1135-5948 DOI 10.26342/2023-70-20

of this problem and provide a comprehensive analysis.

There have been relatively few studies on tasks involving multilingualism with truly low-resource languages (such as Guarani/Jopara) or specific dialects (such as Spanish from Spain) in the literature. This is likely because most research in this area has focused on languages with more resources available, such as a sufficient number of Wikipedia or Common Crawl pages. It is important to carefully study the behavior of state-of-the-art machine learning models, as well as traditional models, to determine which is best suited for addressing the problem of multilingualism, particularly in low-resource languages, and under what conditions.

This thesis may be beneficial to both Spanish-speaking communities (especially in Spain) and Guarani-speaking communities (in Paraguay and surrounding countries such as Argentina, Bolivia, and Brazil) because most NLP systems are designed for use with rich-resource languages. The approaches presented in this work could be applied in various fields and disciplines, including marketing, psychology, sociology, politics, tourism, health informatics, and more, in order to extract insights from written opinions in these languages in various dimensions (such as sentiments, affections, and language type). It is important to have adequate resources for accurately and fairly analyzing written opinions in these languages.

2 Structure

This thesis consists of eight chapters and three appendices, which are described below.

Chapter 1 Introduces the research being conducted, its background and context, as well as the motivation, objectives, and methods of the research.

Chapter 2 Aims to present a thorough review of the use of deep learning to address the problem of multilingual sentiment analysis in social media to the research community. It provides a comprehensive overview of the field and highlights common ideas and issues that have been addressed in the implementation of multilingual sentiment analysis. It also offers a clear summary and discussion to identify potential areas for further research. This chapter is an expansion of a paper published in the journal *Applied Soft Computing* in the special issue 'Soft Computing for Recommender Systems and Sentiment Analysis' (Agüero-Torales, Abreu Salas, and López-Herrera, 2021).

Chapter 3 Focuses on the creation of corpora for low-resource languages and code-switched languages and is divided into the following sections: (i) collection of Spanish COVID-19-related tweets using keywords, language identification tools, and geolocated data for Spanish cities and regions; (ii) collection of Guarani-Spanish (also known as Jopara) Twitter text data for sentiment analysis, which includes challenges such as unbalanced classes due to the limited number of tweets written in Guarani-dominant; (iii) collection of three new, multi-annotated corpora of Jopara Guarani-dominant tweets for affect detection: (a) emotion recognition, (b) humor detection, and (c) identification of offensive and toxic language. The content of this chapter has been adapted from papers published in *Procesamiento Del Lenguaje Natural* (Agüero-Torales, Vilares, and López-Herrera, 2021) and CALCS 2021 (co-located with NAACL 2021) (Agüero-Torales, Vilares, and López-Herrera, 2021), as well as a work submitted to a journal.

Chapter 4 We used NLP techniques to study the discussions taking place on Twitter in Spain at the start of the COVID-19 pandemic. We analyzed the tweets and tracked the evolution of the topics by comparing them to newspaper articles. We also developed a small evaluation framework that involved human judgment. We used both a generative approach and a discriminative approach, which involves identifying the most important keywords and phrases, to represent the topics. The results of this research have been published in the journal *Procesamiento Del Lenguaje Natural* (Agüero-Torales, Vilares, and López-Herrera, 2021).

Chapter 5 Here, various machine learning methods, ranging from traditional approaches to more advanced transformer-based techniques, were applied to the low-resource language Guarani and to a combination of Guarani and Spanish (called Jopara). The performance of the different models was compared and error analysis was conducted to gain further insight into the classifiers' performance in this particular low-resource set-

ting. This chapter is an extension of a previously published paper in CALCS 2021 (co-located with NAACL 2021) (Agüero-Torales, Vilares, and López-Herrera, 2021).

Chapter 6 We describe our efforts to build and pre-train transformer-based language models (Vaswani et al., 2017) using Wikipedia data in the low-resource language Guarani, which faces challenges due to the presence of code-switching. We present a summary of the approaches we took to train a set of BERT models (Devlin et al., 2019) for Guarani and Jopara and evaluate them on tasks related to sentiment analysis. These models overall outperformed the mBERT (multilingual BERT) and Spanish BERT (Cañete et al., 2020, BETO), which do not include Guarani during pre-training, on tasks such as, (i) emotion recognition, (ii) humor detection, (iii) identification of offensive language, and (iv) polarity classification in F1-score and accuracy metrics.

Chapter 7 Summarizes the publications, contributions, and findings of the thesis.

Chapter 8 Presents the conclusions of the thesis. Additionally, we suggest potential areas for further investigation based on the results we have obtained.

Appendices *Appendix A* contains the publications that are part of the thesis. *Appendix B* explains the quantitative analysis for the topic modeling discussed in Chapter 4 and provides information about the annotation guidelines for the Guarani-dominant Jopara corpora used in Chapters 5 and 6. *Appendix C* provides details about the implementation and hyperparameter optimization of the machine learning models used in Chapters 5 and 6, as well as information about the scraped Twitter accounts mentioned in Chapter 3.

3 Contributions

This section provides an overview of the key findings, results, and contributions of the thesis.

3.1 Software prototype

`Gastro-miner`[1] is a cloud-based tool that allows the analysis of users' reviews and opinions written in English about restaurants on social media platforms such as TripAdvisor.com. It allows the collection, storage, cleaning, preprocessing, sentiment analysis, and visualization of review data. The tool was developed using Python, including Scrapy for web scraping, NLTK for NLP, Matplotlib for data visualization, and Django as a web framework. The tool was implemented using virtualization technology such as Vagrant, VirtualBox, and the Docker stack, and data was stored using MongoDB. The sentiment analysis stage used the VADER tool.[2] `Gastro-miner` can be customized for use on other social media platforms, languages, or settings, and the methodology and architecture of the tool are considered a contribution of the thesis.

The results of this study have been published at *Procedia Computer Science* (Agüero-Torales et al., 2019) and were presented at the *Proceedings of the ITQM 2019* and as a poster (Agüero-Torales, López-Herrera, and Cobo, 2018) at the 'Jornadas Científicas de Ciencia de Datos' organized by the *Universidad Comunera* (Asunción, Paraguay), and was awarded first place in the *i-Data* applied data science contest.

3.2 Contributions and resources

The main contributions and resources that have been made available to the research community are listed below:

1. Several corpora for low-resource languages: (a) an unlabeled Spanish Twitter corpus ($\sim$ 1M) about the COVID-19 pandemic outbreak,[3] (b) the first Guarani-dominant Jopara corpus (3,491 tweets) for sentiment analysis,[4] annotated according to a trinary scale (positive, negative, neutral), and (c) the first Guarani-dominant Jopara text-based dataset (2,364 tweets) for affect detection;[5] which includes three multi-annotated corpora: (i) emotion recognition annotated according to four mood categories (happy, sad, angry, other), (ii) humor detection, and (iii) identification of offensive and toxic language.

2. A small evaluation framework with a small guide, that outlines the process followed by native Spanish-speaking anno-

[1] `https://github.com/mmaguero/cloud-based-tool-SA`

[2] `https://github.com/cjhutto/vaderSentiment`

[3] `https://doi.org/10.7910/DVN/6PPSAZ`

[4] `https://doi.org/10.7910/DVN/GLDX14`

[5] `https://github.com/mmaguero/guarani-multi-affective-analysis`

tators to evaluate a sample of the topics discovered in Chapter 4.

3. An annotation mini-guidelines document that outlines the process followed by the bilingual annotators (Guarani-Spanish) as they manually annotated the Guarani-dominant Jopara corpora.

4. A customized tool for language identification.[6] This tool is made up of multiple other tools.

5. A method[7] for discovering topics in Spanish tweets (spoken in Spain) that combines linguistic knowledge with generative and discriminative approaches using the LDA (Latent Dirichlet Allocation) algorithm.

6. A detailed and well-organized set of procedures for creating a Twitter dataset for code-switching and low-resource languages, which outlines the limitations and difficulties encountered during the data-gathering process.

7. A Guarani tokenizer and a set of pre-trained Guarani language models, based on BERT, a widely used transformer-based model, that can be used for a variety of NLP tasks in Guarani or Jopara, such as sentiment analysis. These models were trained with data from Wikipedia in Guarani, including:[8] (a) three monolingual BERT models for the Guarani language and (b) two language models fine-tuned with Guarani (BETO and mBERT respectively).

References

Agüero-Torales, M., D. Vilares, and A. López-Herrera. 2021. On the logistical difficulties and findings of jopara sentiment analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102, Online, June. Association for Computational Linguistics.

Agüero-Torales, M. M., J. I. Abreu Salas, and A. G. López-Herrera. 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107:107373.

Agüero-Torales, M. M., A. G. López-Herrera, and M. J. Cobo. 2018. Gastro-miner: Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor. Caso de Estudio sobre Restaurantes de la Provincia de Granada. In *I Jornadas Científicas en Ciencia de Datos*, page 34, Asunción, Paraguay, October. Universidad Comunera. Abstract (Poster).

Agüero-Torales, M. M., D. Vilares, and A. G. López-Herrera. 2021. Discovering topics in twitter about the covid-19 outbreak in spain. *Procesamiento del Lenguaje Natural*, 66(0):177–190.

Agüero-Torales, M., M. Cobo, E. Herrera-Viedma, and A. López-Herrera. 2019. A cloud-based tool for sentiment analysis in reviews about restaurants on tripadvisor. *Procedia Computer Science*, 162:392–399. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[6] `https://github.com/mmaguero/lang-detection`

[7] `https://github.com/mmaguero/twitter-analysis`

[8] `https://huggingface.co/mmaguero`