Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 227-230 recibido 14-11-2022 revisado 01-02-2023 aceptado 11-02-2023

Linguistic features integration for text classification tasks in Spanish

Integración de características lingüísticas para tareas de clasificación de texto en español

José Antonio García-Díaz
Facultad de Informática, Universidad de Murcia
joseantonio.garcia8@um.es

Abstract: This manuscript summarises the doctoral thesis of José Antonio García-Díaz at the University of Murcia, under the supervision of doctors Rafael Valencia-García and Pedro José Vivancos-Vicente. This doctoral thesis is published by compendium of publications under the industrial doctorate modality. The act of defence took place on Tuesday, July 5, 2022, before the court composed of doctors Salud María Jiménez-Zafra, from the University of Jaén; Miguel Ángel Rodríguez-García, from the Rey Juan Carlos University; and Mª del Pilar Salas-Zarate, from the National Technological Institute of Mexico. The qualification obtained was Outstanding Cum Laude unanimously. In addition, the mention of international doctorate was obtained.
Keywords: Linguistic features, feature integration, automatic document classification, natural language processing.

Resumen: Este documento resume la tesis doctoral por compendio de publicaciones de José Antonio García-Díaz en la Universidad de Murcia, bajo la supervisión de los doctores Rafael Valencia-García y Pedro José Vivancos-Vicente bajo la modalidad de doctorado industrial. El acto de defensa tuvo lugar el martes 5 de Julio de 2022 ante el tribunal compuesto por los doctores Salud María Jiménez-Zafra, de la Universidad de Jaén; Miguel Ángel Rodríguez-García de la Universidad Rey Juan Carlos; y Mª del Pilar Salas-Zarate, del Tecnológico Nacional de México. La calificación obtenida fue de Sobresaliente Cum Laude por unanimidad. Además, se obtuvo la mención de doctorado internacional.
Palabras clave: Características lingüísticas, integración de características, clasificación automática del texto, procesamiento del lenguaje natural.

1 Introduction

Natural Language Processing (NLP) is the branch of Artificial Intelligence (AI) and Linguistics that aims at easing the communication between computers and humans by means of human language.

The scope of this thesis is Automatic Document Classification (ADC), an NLP task which consists in assigning a set of predefined labels to a set of documents. ADC can be applied to Author Profiling (AP), Emotion Detection (ED), Sentiment Analysis (SA), or hate-speech detection among others. To do ADC, computers need practical ways to represent natural language. One of these ways is by means of Linguistic Features (LFs), which represent documents as a vector formed by the percentage of linguistically relevant traits, that indicate *what* a text says, and *how* it says it.

Two research hypotheses are raised in this thesis:

- **RH1**. The inclusion of a set of LFs that capture linguistic traits of the authors can improve the performance of ADC. We address this study in Spanish, including a wide variety of domains concerning infodemiology, hate-speech, humour, or irony among others.
- **RH2**. The inclusion of LFs improves the interpretability to the models with a fewer number of features that generalise better than systems built upon novel Language Models and Transformers.

To accomplish the hypotheses, we have

ISSN 1135-5948 DOI 10.26342/2023-70-18

obtained a taxonomy of LFs in Spanish, and we have developed two software tools: UMUTextStats and UMUCorpusClassifier. The validation of the research hypotheses has been conducted in several scenarios with the validation of the features and the compilation of several linguistic corpora in Spanish.

2 Structure and organisation

Chapter 1 details all the contributions derived from this work. Apart from the abstract, the introduction, its motivation and a state-of-the-art subsection with the methodologies and evaluation used, this chapter describes the system architecture of the two tools developed: UMUTextStats and UMUCorpusClassifier. Besides, it summarises the experimental results obtained during the validation of the tool, which have given rise to the publications that are presented by compendium and the participation in several international workshops.

Chapter 2 presents the research articles that are attached as the compendium of the doctoral thesis. These research articles are about: (1) an ontology-driven aspect-based sentiment analysis system, focused on infodemiolgy (García-Díaz, Cánovas-García, and Valencia-García, 2020); (2) the compilation and evaluation of the Spanish MisoCorpus 2020, focused on misogyny detection (García-Díaz et al., 2021); (3) the compilation process of the Spanish PoliCorpus 2020 and its evaluation with two author analysis tasks: an author profiling task to extract demographic and psychographic traits, and an authorship attribution task in order to obtain which the author of a set of anonymous documents (García-Díaz, Palacios, and Valencia-García, 2022); (4) and the compilation process of the Spanish SatiCorpus 2021, which includes satirical headlines and tweets from a wide variety of countries from Spain and Latin America newspapers (García-Díaz and Valencia-García, 2022).

Chapter 3 contains the conclusions, a summary of all the publications derived from this work, and a list of promising future research lines related to the LFs and ADC in Spanish.

3 Main contributions

The main contributions of this doctoral thesis are the UMUTextStats and UMUCorpusClassifier tools, and their validation in multiple scenarios. Accordingly, this section describes both tools and their validation.

3.1 Tools

3.1.1 UMUTextStats

UMUTextStats[1] (García-Díaz et al., 2022) is a tool for extracting LFs. This tool focuses for Spanish since it is one of the most used languages on the Internet. UMUTextStats is inspired in LIWC (Tausczik and Pennebaker, 2010). However, UMUTextStats solves some deficiencies identified in LIWC for Spanish (Garcí et al., 2007). For instance, LIWC does not capture inflection mechanisms that indicates the tense, mood, and the person to whom the verb refers in Spanish. In addition, LIWC is a commercial tool, and we aim to provide an open-source tool for the Spanish NLP community.

UMUTextStats captures a total of 365 LFs organised within the following taxonomy: (1) phonetics, (2) morphosyntax, (3) correction and style, (4) semantics, (5) pragmatics, (6) stylometry, (7) lexical, (8) psycho-linguistic processes, (9) register, and (10) social media.

3.1.2 UMUCorpusClassifier

The UMUCorpusClassifier tool[2] eases the compilation and annotation of linguistic corpora (García-Díaz et al., 2020), which is a very time-consuming task. Besides, the quality of manually annotated datasets is heavily influenced by disagreements between annotators. Therefore, the lack of supervision of the annotation process can lead to poor quality corpora.

The documents compiled from UMUCorpusClassifier can be classified using distant supervision or manual labelling. Besides, UMUCorpusClassifier allows to coordinate groups of annotators and measure their performance with several metrics concerning inter-annotator agreement.

3.2 Validation

3.2.1 Aspect-based Sentiment Analysis

We evaluate the LFs in an aspect-based SA study focused on infodemiology (García-Díaz, Cánovas-García, and Valencia-García, 2020). For this, a dataset from Twitter with short texts related to different infectious diseases was compiled. Once the dataset was

[1] https://umuteam.inf.um.es/umutextstats/

[2] https://umuteam.inf.um.es/corpusclassifier/

compiled, we extracted the LFs and used them to perform a multi-class SA, achieving an accuracy of 55.3% with the LFs. These results outperformed the rest of the features, which included non-contextual word embeddings trained with a convolutional or recurrent neural networks.

The aspects related to infodemiology were represented within a domain ontology, representing risks, symptoms, transmission methods or drugs related to infectious diseases. In this work, we assumed that one document contains only one sentiment. Accordingly, we ranked the relationship between the sentiment of the tweet with the ontology classes.

The interpretability of the resulting models was measured with the Information Gain of the LFs. We observed that numerals are correlated to negative documents and that the usage of colloquialism is more related to positive and neutral tweets than negative tweets.

Other validations in SA were our participation in TASS 2020 and EmoEvalEs shared tasks.

3.2.2 Hate-speech and misogyny detection

Our contributions regarding hate-speech started with the compilation and evaluation of the Spanish MisoCorpus 2020 (García-Díaz et al., 2021), which includes documents concerning violence against relevant women, messages written from Spain and Latin America, and general traits related to misogyny, such as discredit or dominance among others. The dataset is balanced, and it contains 3 841 misogynous documents. The best accuracy achieved was 85.175% with Support Vector Machine (SVM). Moreover, we observed that the combination of the LFs and the sentence embeddings outperformed the rest of the feature sets. This finding supports our first research hypothesis regarding the improvement of the results for ADC. As expected, we observed that LFs related to offensive language have a strong correlation for misogyny detection. A strong correlation between the grammatical gender and misogyny identification was also found. This is relevant because some words can be interpreted differently according to their gender. We also observed a strong correlation with correction and style features, such as the percentage of misspelled words.

In addition, we participated in the last two editions of EXIST (2021, 2022), and MeOffendES 2021.

3.2.3 Figurative language: Satire, Sarcasm, and Humor detection

We compile the Spanish SatiCorpus 2021 (García-Díaz and Valencia-García, 2022) to distinguish between satirical news and real news. The dataset is balanced and contains news headlines from Twitter. The accounts were selected from different Spanish spoken countries. Moreover, we decided to enlarge this dataset including tweets from Twitter accounts used for impersonating and satirise real relevant people. This dataset was automatically annotated, based on the idea that all tweets from satirical news media are satiric. We evaluated the LFs separately and combined with other types of features using different strategies. Our best result was achieved with a combination of the LFs and BERT with an accuracy of 97.405%. We observed that the number of orthographic errors is more common in non-satirical documents than in satirical documents. In contrast, the number of hashtags more commonly appears in non-satirical documents. Regarding morphological features, the use of pronouns and nouns is good for discerning between satirical and non-satirical documents, being the pronouns more frequently found in satirical documents whereas nouns are more common in non-satirical documents.

Besides, we participate in Hahackathon 2021 and HaHa 2021.

3.2.4 Author Analysis

We explored the reliability of the linguistic features in two experiments regarding author analysis: authorship attribution and profiling, and we annotated each user with their gender, year of birth, and political spectrum on two axes (binary and multiclass) (García-Díaz, Palacios, and Valencia-García, 2022).

Concerning the AP study, we evaluated the LFs with sentence and word embeddings from Word2Vec, FastText, or BERT. We observed that the LFs achieved promising results, outperforming BERT in some traits, such as gender. Moreover, the combination of the LFs with the rest of features usually results in better results than achieved with both features separately. Concerning the interpretability of the results, we observed that morphosyntax is the most relevant category for determining demographic traits. Besides,

we found correlations between the percentage of personal pronouns and verbs, the usage of colloquialisms, topics related to countries and languages.

4 Conclusions

In this doctoral thesis we have proven the effectiveness of the integration of LFs for conducting ADC (RH1) and their interpretability (RH2). Finally, it is worth mentioning that two software tools have been released to the Spanish research NLP community two software tools: UMUTextStats and UMUCorpusClassifier.

Acknowledgements

This PhD. Thesis has been partially supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme, by the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/ AEI/10.13039/501100011033 and by the research project AIInFunds(PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

Garcí, F. A., J. W. Pennebaker, N. Ramírez-Esparza, and R. Suriá. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista mexicana de psicología*, 24(1):85–99.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

García-Díaz, J. A., M. Cánovas-García, R. C. Palacios, and R. Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.*, 114:506–518.

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Gener. Comput. Syst.*, 112:641–657.

García-Díaz, J. A., R. C. Palacios, and R. Valencia-García. 2022. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Gener. Comput. Syst.*, 130:59–74.

García-Díaz, J. A. and R. Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

García-Díaz, J. A., P. J. Vivancos-Vicente, A. Almela, and R. Valencia-García. 2022. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044.

Tausczik, Y. R. and J. W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.