Using Linguistic Features for Improving Automatic Text Classification Tasks in Spanish

El Uso de las Características Lingüísticas para Mejorar las Tareas de Clasificación Automática de Texto en Español

José Antonio García-Díaz[1]
[1]Universidad de Murcia. Facultad de Informática.
Departamento de Informática y Sistemas
joseantonio.garcia8@um.es

Abstract: The objective of this doctoral thesis is the design and evaluation of linguistic features in Spanish to apply them in text classification tasks, such as sentiment analysis, hate messages detection, or plagiarism detectors. Currently, the state of the art concerning text classification makes use of deep-learning architectures fed with word embeddings, that are a text representation model in which words are encoded as dense vectors based on co-occurrence properties of the language. Although these models outperformed the results of previous models based on frequency-based vectors such as word and character n-grams, they also result in models whose behaviour is difficult to interpret since they behave as black-box systems and the large number of features they produce. Our hypothesis is that the inclusion of linguistic features to these systems can provide better results at the same time they provide interpretable features. During our research, we have developed two Natural Language Processing tools focused on Spanish: (1) a linguistic features extraction system, and (2) a platform for compiling and supervising corpus for conducting supervised machine learning experiments.
Keywords: Natural Language Processing, Automatic Text Classification, Supervised Machine Learning, Linguistic Feature Extraction

Resumen: El objetivo de esta tesis doctoral es el diseño y la evaluación de características lingüísticas que puedan aplicarse a tareas de clasificación de texto, tales como el análisis de sentimientos, la detección de mensajes de odio, o en sistemas de detección de plagio. El estado del arte actual hace uso de modelos de aprendizaje profundo que usan word-embeddings como entrada. Los word-embeddings son un sistema de representación del lenguaje que codifica en vectores el significado de las palabras agrupando palabras con contexto similar. Aunque estos modelos han mejorado los resultados de los enfoques anteriores basados en modelos de frecuencia de palabras, los modelos que se generan siguen siendo modelos difíciles de interpretar pues funcionan como un sistema de caja negra y que generan gran cantidad de características. Nuestra hipótesis es que la inclusión de características lingüísticas a estos modelos resulta en modelos que devuelven mejores resultados así como modelos más interpretables. Durante nuestra investigación hemos desarrollado dos herramientas para la clasificación de textos en español: (1) una plataforma de extracción de características lingüísticas y (2) una plataforma que permite compilar y etiquetar corpus para llevar a cabo experimentos de aprendizaje supervisado.
Palabras clave: Procesamiento del Lenguaje Natural, Clasificación automática de textos, Aprendizaje supervisado, Extracción de Características Lingüísticas

1 Introduction and background

Automatic text classification is one of the trending tasks concerning Natural Language Processing (NLP) with applications in sentiment analysis, aggressiveness detection, fake news detection, or plagiarism detection among other applications. In a nutshell, automatic text classification consists in assigning one or more pre-determined classes to texts based on its content for organising and

Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network (PLNnet-DS-2020)

making profit of the large amounts of information that exist on the Internet in unstructured textual format (Altınel and Gani, 2018).

The state of the art concerning automatic text classification consists of the design of deep learning classifiers that are capable of learn patterns hidden in the texts of from which they can deduce the most appropriate set of labels that matches each document. In order to work natural language, a computer needs to transform data encoded as natural language into meaningful vectors. There are several approaches to perform this transformation, being the most popular approaches those based on counting methods. For example, the Bag of Words (BoW) model consists in encoding a document as the frequencies of the words that make up that text. BoW and other count-bases models are becoming obsolete giving way to models based on word embeddings, in which words are encoded as distributed representations based on the distributional hypothesis that captures co-occurrence properties of the language (Almeida and Xexéo, 2019). However, although the word embeddings models have outperformed count-based approaches (Rudkowsky et al., 2018), both approaches produce black-box models in which it is difficult to explain the behaviour of the model (Danilevsky et al., 2020). Linguistic features, on the other hand, represents texts by means of a vector formed by the percentage of psycho-linguistically relevant words, with the aim of classifying those words that indicate what the text says and how it says it. We argue that linguistic features can be combined with count-based features as well as with word embeddings in order to build better models as well as they provide interpretability of their behaviour.

The objective of this doctoral thesis is the design and evaluation of linguistic features in Spanish and apply them for solving automatic text classification tasks. We focused our proposal on Spanish because it is the third most used language on the Internet, only behind English and Chinese. Although there are linguistic extraction tools available in Spanish, as far as our knowledge goes, they do not handle all the specific phenomena of Spanish. For example, Spanish, unlike English, makes use of inflection to indicate the tense and mood of verbs, along with the person to whom they refer. These kinds of features are not available in the systems we have evaluated. For this reason, we are developing a set of linguistic features designed specifically for the Spanish and we have developed a tool for extracting those linguistic features from a wide variety of sources, called UMUTextStats[1]. As a extra contribution, we have developed another tool, named UMUCorpusClassifier[2], which helps compile and manage groups of annotators to create their linguistic corpus for supervised machine learning experiments (García-Díaz et al., 2020).

2 Research Hypotheses

The research hypotheses we investigate in this work are related to the inclusion of linguistic features for solving automatic text classification tasks. Our first research hypothesis states that (1) the inclusion of linguistic features that capture stylometric traits of the authors can improve automatic text classification systems creating more robust models. (2) Our second research hypothesis states that the inclusion of linguistic features can provide interpretability to the models with a fewer number of features that can generalise better.

To validate these research hypotheses, we established the following objectives:

- Design and development of a tool capable of generating a vector composed of linguistic features. These linguistic vectors can be used as input of different machine learning models in order to build automatic text classifiers for solving a wide variety of NLP classification tasks, such as sentiment analysis, aggressiveness detection, or satire identification.
- Use the linguistic features to validate our hypotheses in different automatic text classification tasks. In this sense, we are evaluating the linguistic features separately and in combination with word embeddings to feed traditional machine learning and deep learning architectures such as Recurrent Neuronal Networks and transformers like BERT. Some of the domains in which we evaluating our proposal are: (1) infodemiology, to measure public opinion regarding public

[1] https://pln.inf.um.es/umutextstats/
[2] https://pln.inf.um.es/corpusclassifier

health concerns; (2) misogyny and aggressiveness detection, in order to help to build safer places for everyone in social networks; and (3) and author profiling, in order to improve plagiarism detectors. Moreover, we are also evaluating our proposal by participating in several shared task regarding text classification from different workshops.

- Compilation and annotation of linguistic corpus. We are compiling different linguistic corpus to validate our hypotheses and releasing them for the scientific community.

3 Methodology

In this section we describe the two NLP tools we have created for support us to validate our hypotheses.

On the one hand, UMUTextStats is a linguistic feature extraction tool designed for Spanish. This tool can extract a vector made up of the percentages of words and expressions that fit into a series of psycholinguistic features. This tool is inspired in LIWC (Tausczik and Pennebaker, 2010) (pronounced *Luke*), which is a tool for the extraction of linguistic features capable of analysing a set of documents and generating a vector with the percentages of a series of pre-established categories related linguistics. Although it was originally designed for English, LIWC has a version translated to Spanish. This translation process was analysed in (Ramírez-Esparza et al., 2007), in which some drawbacks were identified: (1) translation issues between English and Spanish, (2) an arbitrary design of the dimensions, (3) grammatical phenomena of Spanish not considered, (4) insufficient verb conjugations and, (5) the lack of studies with Spanish sources. In addition, it is important to note that LIWC is a commercial tool, which ended up motivating the development of a free tool for the NLP community in Spanish.

UMUTextStats is extensible and allows defining dimensions from a set of predefined abstract dimensions, where we highlight:

- Dictionary dimensions. Allows to find regular expressions that appear in a certain catalogue of terms. This dimension also allows to indicate counterexamples. Using counterexamples makes it easier to design a simple regular expression on a term, and then list the exceptions, as is the case with grammatical gender. Some of the resources were compiled from available lexicons such as (Molina-González et al., 2013).
- Dimensions based on regular expressions. Allows to specify regular expressions to, for example, detect expressions within quotation marks, which is indicative of the use of quotes or words that acquire a certain special tone.
- Typography-based dimensions. It allows detecting the percentage of words written in capital letters, which may be an indication of a high tone of the voice, an interesting feature for detecting violence over the Internet.
- Custom dimensions. In addition to these features, it is possible to extend the tool for including custom features. For example, we have included features that captures grammatical and stylistic errors.

Each dimension can be configured to operate with different versions of the same text. Therefore, some dimensions can operate on a filtered version that makes it easier to search for terms in the dictionary, while the original version can be used to measure characteristics such as the percentage of words in capital letters.

These feature types helped us to design the linguistic features that can be categorised as follows: (1) grammatical features, to measure Part-of-Speech (PoS) words that include a list of a thousand Spanish popular verbs and their respective conjugations obtained from online resources as well as discourse markers and a wide variety of adverbs, adjectives and pronouns; (2) spelling mistakes, in order to detect words and expressions that capture misspellings that could indicate that the author did not paid enough attention to review their writing, or features that indicate informal speech language, such as colloquialisms, popular abbreviations in texting, or non-fluent markers; (3) stylometric features, to capture the average length of the documents, their number of sentences within the texts as well as their type (declarative, interrogative, exclamatory) as well as other punctuation and symbols markers to capture sentence dividers to measure the rhythm of the

text; and (4) we employed dictionaries to capture different topics such as health concerns, food, or animals among other topics as well as different lexicons to capture positive and negative emotions.

On the other hand, we have developed another tool in order to facilitate the design of experiments to validate UMUTextStats, we have developed another tool called UMUCorpusClassifier. This tool allows to compile datasets from Twitter from a query-string and a geographic location and define a set of custom labels. Then, the researchers can organise and supervise groups of annotators. This platform allows different annotators to label the same tweet. Consequently, the final corpus enhances those documents that have the most consensus among annotators, allowing researches to discard (or to prioritise) those tweets that generate more controversy.

4 Validation

The following section describes the validation experiments that we have already performed to validate our work that includes domains such as infodemiology (see Section 4.1) and misogyny identification (see Section 4.2). We also describe our participation in some shared tasks regarding automatic text classification including TASS'2020 (see Section 4.3), MEX-A3T'2020 (see Section 4.4), and AI-SOCO'2020 (see Section 4.5).

4.1 Sentiment Analysis applied to Infodemiology

Infodemiology investigates the use of information available on the Internet in order to improve public health services. To validate the linguistic features, we first compiled a corpus consisting of tweets from Ecuador based on virus keywords such as Zika or Chikungunya. The final version of the corpus is composed of 10,843 positive tweets, 10,843 negative tweets and 7,659 neutral tweets, all written in Spanish and supervised by 20 students from the University of Guayaquil who performed a manual classification of the tweets. This corpus was released to the scientific community and it can be found a detailed description of the annotation process in the paper (García-Díaz, Cánovas-García, and Valencia-García, 2020).

Then, we evaluated the linguistic features along with different deep-learning architectures to perform a multi-class evaluation of sentiments (negative-neutral-positive). The results of this evaluation are depicted in Table 1, in which we can observe that the linguistic features in isolation achieved the best accuracy identifying correctly the 55.3% of the tweets. Moreover, when we combined the linguistic features and word embeddings employing other deep-learning architectures, we observed that the models based on Recurrent Neuronal Networks such as Long-Sort Term Memory (LSTM) and its bidirectional variant (BiLSTM) benefits for the combination with linguistic features while the accuracy with a Convolutional Neuronal Network (CNN) only decreased slightly.

Feature set	Accuracy
LF	**55.3**
LSTM	46.8
LSTM+LF	51.0
BiLSTM	42.9
BiLSTM+LF	54.2
CNN	49.3
CNN+LF	49.1

Table 1: Comparison of the accuracy of different feature sets in a multi-class sentiment analysis experiment in the infodemiology case-study

Next, we developed an aspect-level sentiment analysis system towards the infectious diseases. The aspects were extracted by using an ontology that models the infectious disease domain with concepts such as risks, symptoms, transmission methods or drugs. Then, we measured the relationship between these concepts in order to determine the degree to which one concept influences other concepts and we used to average the sentiments extracted for each document by using deep-learning models that combined statistical and linguistic features. Finally, we created a graphical interface in which final users can see at a glance the sentiment for each concept and analysed how each concept influences the sentiment of the others. This interface is depicted in Figure 1.

It is worth noting that we previously used a preliminary version of the corpus to evaluate different statistical approaches based on word and character n-grams. This evaluation is described in (Apolinardo-Arzube et

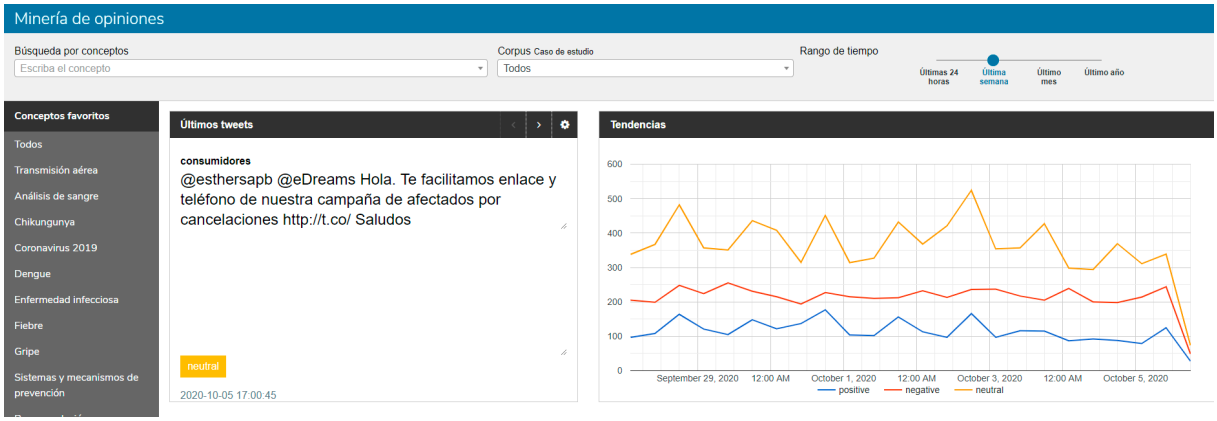

Figure 1: Graphical user interface for the aspect-based sentiment analysis in the infodemiology case study

al., 2019) and this work helped us to better understand these statistical models as well as to provide a baseline for further experimentation.

4.2 Misogyny identification

We evaluated the linguistic features in order to build a system capable of detecting misogynistic behaviour on social networks. Our contribution was two fold. On the one hand, we applied sentiment analysis and social computing technologies for detecting misogynous messages in Twitter and, on the other, we compiled a corpus composed of three subsets concerning of: (1) violence towards relevant women, (2) messages harassing women in Spanish from Spain and Spanish from Latin America, and (3) general traits related to misogyny (García-Díaz et al., 2020).

In this case study, we combined the linguistic features with average word embeddings (also known as sentence embeddings) from fastText (Grave et al., 2018) in order to understand which linguistic phenomena principally contribute to the identification of misogyny. We evaluated our proposal with three machine learning classifiers, achieving the best accuracy of 85.175% with a Support Vector Machine (SVM). In Table 2 we can observe the accuracy for each feature set evaluated with the SVM. This comparison involved a baseline model based on BoW, the average word embeddings from fastText (AWE), the linguistic features in isolation (LF), and the combination of linguistic features and the average of words embeddings (AWE+LF). We can observe that the combination of linguistic features and the average of word embeddings outperformed the rest of the feature sets which supports our first hypothesis regarding the improvement of the results for text classification tasks.

Feature set	Accuracy
BoW	73.798
AWE	81.020
LF	78.938
AWE+LF	**85.175**

Table 2: Comparison of the models with a BoW baseline, average of words embeddings, linguistic features in isolation or combined with sentence embeddings using SVM in the misogyny case study

Regarding the interpretability of the model we calculated the Information Gain (IG), a metric employed in decision trees to decide when new branches need to be created. Figure 2 includes the 20 top features with major information gain for our model in which we detected that the usage of offensive language was the most discerning feature. Other discriminatory features were the number of words that are grammatically feminine and some features concerning grammatical errors, such as mistakes in writing or the percentage of misspelled words.

In addition, we evaluated our proposal with existing two corpora for misogyny and aggressiveness detection. On the one hand, with the AMI'2018 dataset (Fersini, Rosso, and Anzovino, 2018) (see Table 3) in which we outperformed the results of the participants of the shared task by combining the linguistic features and the word embeddings with a linear SVM. Moreover, the

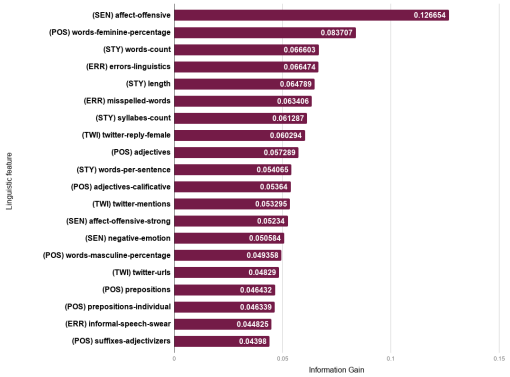

Figure 2: Information gain of the linguistic features in the misogyny case study

model trained with SMO (a SVM variant) also improved the baselines provided. However, there are two more tasks proposed in AMI'2018 regarding the identification of different misogynistic traits and concerning if the message was directed to particular women or in general. In those cases, our proposal achieved an accuracy slightly worse than the proposed baseline. With regard HatEval'2019 (Basile et al., 2019) (see Table 4), our proposal outperformed the baseline proposed as well as the best results of the participants of the shared task with the combination of the average of word embeddings with the linguistic features with an accuracy of 75.4505%.

Run	**Accuracy**
AWE+LF (LSVM)	**81.5217**
14-exlab.c.run3	81.4681
JoseSebastian.c.run1	81.4681
AWE+LF (SMO)	79.2271
AMI-BASELINE	76.7750

Table 3: Comparison of our proposal for misogyny identification with the AMI dataset

Feature set	**Accuracy**
AWE+LF (SMO)	**75.4505**
AWE+LF (LSVM)	73.3041
Max	73.0000
SVG HatEval Baseline	70.1000

Table 4: Comparison of our proposal for misogyny identification with the HatEval dataset

4.3 Sentiment Analysis (TASS-2020) (Track 4)

We participated in the two tasks proposed in the TASS'2020 (García-Vega et al., 2020) from IberEval. The first task consisted in the classification of tweets according to general sentiments of tweets written in several Spanish varieties, and the second task consisted in a multi-class fine-grained distinction among six basic emotions. Our proposal was grounded on the combination of linguistic features in isolation or combined with word-embeddings with a CNN or combined with average word embeddings trained with a SVM. Our proposal achieved the best precision rate regarding emotion detection (Task 2) and competitive results with respect to the general sentiment classification in which tweets written in different varieties of Spanish were mixed. Our participation is described in (García-Díaz, Almela, and Valencia-García, 2020). For the sake of simplicity, we only included in this paper the results of the subtask 1.2 regarding sentiment analysis mixing all the tweets written in different Spanish variants.

Model	**F1**
CNN (LF + WE)	0.336824
SMO (LF)	**0.357876**
SMO (LF + AWE)	0.334466

Table 5: Comparison of our runs for the subtask 1.2 regarding sentiment analysis

4.4 Aggressiveness Identification (MEX-A3T 2020) (Track 6)

We participated in the IberEval 2020 task MEX-A3T (Aragón et al., 2020) focused on aggressiveness identification in tweets written in Mexican Spanish. We based our proposal in the combination of linguistic features and pre-trained word embeddings. In the first run, we applied a SVM with a combination of linguistic features and sentence embeddings; whereas in the second and third run the linguistic features were combined with two deep-learning models: a Convolutional Neural Network and a Bidirectional Long-Short Term Memory. Our results did not outperform the baseline proposed by the organisers, but we could provide an interpretable model.

Our participation is described at (García-

Díaz and Valencia-García, 2020) and our results compared to the baseline and the rests of the participants are depicted in Table 6. We can observe that none of our proposals outperformed the two baselines proposed by the organisers of the shared task with a difference of 0.54% between the second baseline with our second run.

Model	F1 macro
best-result	0.8596
baseline1	0.7983
baseline2	0.7770
CNN (LF + WE)	**0.7716**
BiLSTM (LF + WE)	0.7644
SVM (LF + AWE)	0.7161

Table 6: Comparison of our runs with the two base-lines and the winner of the MEX-A3T'2020 task

4.5 AISOCO'2020

We participated in the AISOCO'2020 shared task from FIRE workshop concerning authorship identification of source-code. In our approach we combined the character n-gram model with author traits in which we applied some of the linguistic features developed with UMUTextStats to capture stylometric features from the authors. Our proposal achieved an accuracy of 91.16% over the test dataset reaching the sixth position in the official ranking and outperforming some baselines based on transformers such as RoBERTa (see Table 7). As the source code was mainly written in English, and not all the source codes from the datasets had comments, we only could apply some of the stylometric linguistic features in our proposal.

Team	Accuracy
AlexCrosby	0.9511
yang1094	0.9428
mutaz	0.9336
AI-SOCO RoBERTa (1)	0.9288
zz	0.9219
FSU_HLJIT	0.9157
UMUTeam	**0.9116**
AI-SOCO RoBERTa (2)	0.9102

Table 7: Results of the official participants and the baselines at the AISOCO'2020 task

5 Further work

Currently, we are designing more experiments that involve author profiling and forensic linguistics in order to improve our tools and validate our hypotheses. We are making efforts to include new deep-learning approaches that includes transformers such as ELMo and BERT to combine with the linguistic features. In addition, we will focus on explainable techniques to gain better understanding of local and global predictions achieved with the linguistic features (Danilevsky et al., 2020). We will also focus on the observation of how linguistic phenomena vary among the different variants of Spanish on general-purpose domains or specific ones such as those related to hate-speech or misogyny.

With regard UMUTextStats we are working with linguistics from the University of Murcia to advise us on establishing a better taxonomy for the linguistic features. Regarding UMUCorpusClassifier, we are improving the detection of duplicate tweets because during the experiments, we identified cases of very similar tweets have been detected where only a comma or punctuation symbol has been varied. Beta versions of these tools are currently available as a web services to be used by the research community.

Acknowledgements

This work has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme.

References

Almeida, F. and G. Xexéo. 2019. Word embeddings: A survey. *CoRR*, abs/1901.09069.

Altınel, B. and M. C. Gani. 2018. Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, 54(6):1129–1153.

Apolinardo-Arzube, O., J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, and

R. Valencia-García. 2019. Evaluating information-retrieval models and machine-learning classifiers for measuring the social perception towards infectious diseases. *Applied Sciences*, 9(14):2858.

Aragón, M., H. Jarquín, M. M.-y. Gómez, H. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, and J. Posadas-Durán. 2020. Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*.

Basile, V., C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Danilevsky, M., K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Fersini, E., P. Rosso, and M. Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228.

García-Díaz, J. A., A. 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., Á. Almela, and R. Valencia-García. 2020. Umuteam at tass 2020: Combining linguistic features and machine-learning models for sentiment classification. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*, pages 187–196.

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. 2020. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 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 Generation Computer Systems*, 112:641–657.

García-Díaz, J. A. and R. Valencia-García. 2020. Umuteam at mex-a3t'2020: Detecting aggressiveness with linguistic features and word embeddings. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*, pages 287–292.

García-Vega, M., M. C. Díaz-Galiano, M. Á. García-Cumbreras, F. M. P. del Arco, A. Montejo-Ráeza, S. M. Jiménez-Zafra, E. M. Cámarab, C. A. Aguilar, M. Antonio, S. Cabezudo, et al. 2020. Overview of tass 2020: Introducing emotion detection. *Proceedings of TASS*.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. *CoRR*, abs/1802.06893.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Ramírez-Esparza, N., J. W. Pennebaker, F. A. García, R. Suriá Martínez, et al. 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.

Rudkowsky, E., M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair. 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

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.