Evaluation of Linguistic Features Separately or Combined with Transformers for Solving Automatic Text Classification Tasks in Spanish

José Antonio García-Díaz[1]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

In this paper we describe the evaluation stage and analysis of the UMUTextStats tool for extracting linguistic features applied to text classification in several domains, including the identification of sexist or offensive comments, emotions, and a fine-grained analysis regarding what texts are funny and what mechanisms are involved to make them funny. These subtasks were organised by IberLEF and IberEval 2021 workshops. During the participation on these subtasks, the linguistic features were evaluated separately and combined with state-of-the-art transformers by means of ensembles and knowledge integration strategies, with the objective of achieve competitive results in all tasks. At the same time, we seek to improve our methods to obtain some interpretability of the results. In summary, our results suggest than the combination of different feature sets improves text classification tasks, especially when they are input in the same neural network.

Keywords

Text classification, Feature engineering, Natural Language Processing

1. Introduction

In the past edition of the doctoral symposium organised by the thematic network PLN.net, we described the main objectives related to this doctoral thesis [1]. These objectives consist in the development of a set of linguistic features for Spanish and their inclusion in Natural Language Processing (NLP) tasks, such as forensics linguistic, author profiling, infodemiolgy [2], or misogyny identification [3] among others. We also described our participation in TASS 2020 [4] and MEX-A3T [5] shared tasks, and we described two NLPs tools developed, one for compiling and annotation corpora [6], and the other, inspired in LIWC [7], for extracting the linguistic features [3, 2].

In summary, our main hypothesis is that linguistic features are somehow high-order features than statistical features based on words and their relationships. Examples of these feature sets are n-grams or contextual and non-contextual embeddings. Moreover, we state that applying the linguistic features results in more reliable and interpretable models.

Our previous participation in the symposium was very positive for us, as we received valuable

Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T), 19-20 October 2021, Baeza (Jaén), Spain.

joseantonio.garcia8@um.es (J. A. García-Díaz)

0000-0002-3651-2660 (J. A. García-Díaz)

feedback from our mentors. Specifically, they recommend us to focus on the interpretability of the machine-learning models. In addition, they note that the results achieved in the shared tasks [8, 9], in which we combined the linguistic features with non-contextual embeddings, were limited compared to the results of other participants. Therefore, this year we have focus on improving our pipeline by (1) evaluating other feature sets to combine and compare with the linguistic features; (2) evaluating techniques for combining the features, such as knowledge transfer, and ensemble learning; and (3) evaluating explainable deep-learning techniques. We have measured our progress by participating in four shared tasks of IberLEF 2021 [10], and one shared task from IberEval 2021. In addition, during this year, we have applied our methods for conducting author profiling and hate-speech detection, with two publications that are under review in scientific journals.

2. Validation

After summarising the main hypotheses of this research, here we describe the validation process carried out, that consisted in the participation in the following shared tasks: EXIST-2021 (see Section 2.1), EmoEvalEs 2021 (see Section 2.2), Hahackathon 2021 and HaHa 2021 (see Section 2.3), and MeOffendEs 2021 (see Section 2.4). For each subtasks we include a summary of the main objectives, the methods evaluated as well as the main insights extracted for each one.

2.1. EXIST-2021. Sexist language identification

The shared task EXIST-2021 [11] focuses on the identification and categorisation of sexist language written in Spanish and English. The organisers of this shared task compiled and annotated documents from several micro-blogging platforms. This shared task was divided into two subtasks: (1) a binary classification of sexism utterances, and (2) a multi-class identification of sexist traits, namely, ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence.

We participated in both subtasks with a combination of the linguistic features and transformers. For this, we tackle each dataset independently and combine the results at the end. During our research, we evaluated different types of embeddings, including (1) pre-trained non contextual word embeddings from fastText, word2vec, and gloVe; (2) sentence non-contextual embeddings from fastText; and (3) contextual word embeddings based on transformers. In addition, we evaluate different neural network architectures, including multi-layer perceptrons, convolutional neural networks, and bidirectional recurrent neural networks. We combined each feature set with the functional API of Keras[1], in a knowledge integration fashion, entering each feature into separate hidden layers and combining them before predicting the result.

Three runs were sent. One with the linguistic features, another combining the linguistic features with transformers, and another based in an ensemble of neural networks of linguistic features and contextual and non-contextual word and sentence embeddings. We achieved our best result in task 1, with an accuracy of 75.14% using the ensemble learning approach, and an accuracy of 61.70% for task 2, with the combination of the linguistic features and transformers.

[1] https://keras.io/guides/functional_api/ (Last accessed: 2021-07-17)

These results were not far from the best results achieved in the official leader board, achieved by the team AI-UPV_1, with an accuracy of 78.04% in task 1, and an accuracy of 65.77% for task 2.

We observed that our baseline, consisted in the linguistic features, achieved limited results. This result was the expected with the English dataset, but unexpected with the Spanish dataset, especially as the linguistic features provided promising results regarding misogyny identification [3]. The main differences between both datasets (the Spanish MisoCorpus 2020 and the EXIST-2021) are the number annotators (3 for the MisoCorpus, 5 for EXIST-2021) and the fact that the annotators from EXIST-2021 followed the guidelines from two experts in gender issues. Besides, the Spanish MisoCorpus 2021 contains a large number of tweets from news sites that were labelled as neutral.

To gain some interpretability, we extracted the information gain from the linguistic features. We observed that the linguistic features related to sexual issues and related to female social groups were discriminatory features for the identification of sexism. However, both features appeared less frequently in documents labelled as *stereotyping and dominance*.

2.2. EmoEvalEs 2021. Emotion Detection

The EmoEvalES shared task [12] is focused on extracting the emotions expressed by users on social media, which it is challenging mainly due to the absence of prosodic features and facial expressions. This shared task consists in a multi-class classification for determining if a text contains one of the following classes: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others.

Like the other shared tasks in which we participated, we based our proposal in the combination of the linguistic features and transformers [13]. We achieved 6th position in the official leader board with an accuracy of 68.5990%, falling only 4.1667% below the best result.

In this shared task, we achieve a significant improvement in our pipeline, as we were able to extract the contextual sentence embeddings from BETO [13]. For this, we extracted a fixed representation of 768-length vector from the `[CLS]` token, after fine-tuning the model with the EmoEvalEs dataset [14]. We observe than the performance of this approach was similar to the one achieved using HuggingFace's Trainer[2]. However, the fixed representation of the BERT embeddings provided to us two important benefits: they are easier to combine with other feature sets within the same neural network and the required time for training and performing inference is reduced.

As we expected, regarding the interpretability of our results, we observed a strong correlation between lexicons related emotions with the labels. Lexicons containing sad expressions were strong related to documents annotated as *sadness* and *disgust*. *anger* with the psycho-linguistic process anger. Negative processes were also related to *anger*, *disgust*, *fear*, *sadness*, and *surprise*.

2.3. Hahackathon 2021 and HaHa 2021

Regarding humour, we have participated in two tasks regarding its identification, categorisation, and evaluation. On the one hand, the HaHackathon 2021 shared task [15], proposed in IberEval'2021, focused on texts written in English, and HaHa 2021 [16], focused on Spanish. Both shared tasks were divided into four subtasks each one. HaHackathon focused on

[2]https://huggingface.co/transformers/main_classes/trainer.html(lastaccessed:2021-07-17)

determining if a text is funny or not (binary classification), how humorous it is (regression), and if its humour is controversial or not and how much (binary classification and regression, respectively). HaHa shared with HaHackathon the first two subtasks, but they included two new subtasks for determining what are the mechanisms to make a text funny and what are the targets of the joke, that were, respectively, a multi-classification and a multi-label tasks.

In HaHackathon 2021 we achieved position 45, with a F1-score of 91.60% in the subtask 1a. A RMSE of 0.8847 for subtask 1b, achieving position 47. Position 14 in subtask 1c, with a F1-score of 57.22%. Finally, we achieved position 46 for subtask 2a, with a RMSE score of 0.8740. It is worth mentioning that, as HaHackathon 2021 was focused in English, we only use the subset of the linguistic features based on corpus statistics, such as the type/token ratio (TTR). In HaHa 2021, we achieved the 1st position in Funniness Score Prediction, the 8th position for humor classification subtask, and the 7th and the 3rd position for the subtasks of humour mechanism and target classification, respectively.

For subtask 2 of HaHa 2021, in which we achieved the best result, we observed that stylometry is a relevant linguistic category. We also observe that interjections, verbs in third person, adverbs, augmentative suffixes, and proper nouns were also relevant features. It also caught our attention to find features related to the number of orthographic errors, as they can be committed on purpose as a humoristic device.

2.4. MeOffendES 2021

Finally, we participated in the MeOffendEs 2021 shared task [17], focused on the identification and categorisation of offensiveness, with datasets in European and Mexican Spanish extracted from different social media platforms. This shared task was divided into two subtasks (two subtasks per language variation). On the one hand, the European Spanish subtasks were based on multi-classification, discerning among (1) offensive texts whose target is a person; (2) offensive texts whose target to groups; (3) texts with inadequate language, but not necessary offensive; and (4) non offensive texts. The Mexican Spanish, on the other hand, were binary classification problems. Each linguistic variant included a subtask in which contextual features from the documents could be considered.

In this case, apart from the linguistic features and transformers, we evaluate fine-grained negation features [18, 19, 20, 21] as a result of a collaboration with the Universidad de Jaén. All these features were combined with ensemble learning. Specifically, we evaluated ensembles based on the mode of the predictions, ensembles based on averaging the predictions of each neural network, ensembles based on the highest probability, and ensembles based on training regression machine learning model from the probabilities of the training split. We observed that the ensembles based on linear regression provided the best results whereas the ones based on the highest probability the best precision over the offensive class.

Our official results were promising, as we ranked in the 2nd place in subtask 1 (F1-score of 87.8289%), 1st in subtask 2 (F1-score 87.8289%), 5th in subtask 3 (F1-score of 67.0588%), and 1st in subtask 4 (F1-score of 66.9449%). However, there were less participants in the subtasks that included the contextual features. Regarding the interpretability of the models, we observed in the Spanish dataset that negative psycho-linguistic processes were strong features to discern from non-offensive documents from the others, but that they were not good indicators to discern

among if the target is a person, a group or simply the use of inadequate language.

3. Conclusions and further work

Since I am in the last year of my doctorate, and having previously participated in the previous version of this symposium, we have focus this study on the validation tasks. Specifically, we described our participation in five shared tasks regarding text classification in which we have achieved promising results. We have tried to follow the indications given by our mentors and we feel that their advises have helped us in a great extent. There is, however, a still a lot of room for improvement. For example, we are still focusing on the interpretability based on the linguistic features in isolation, but not in the context of the neural network. To solve this, we will evaluate the ensembles to analyse which features have the documents that are successfully classified correctly by the transformers and not by the linguistic features and vice versa. Moreover, we are adapting tools such as SHAP and LIME [22]. We are also focusing on improving the detection of figurative language [23] to apply to specific domains such as sarcasm, irony, and satire identification [24].

Acknowledgments

This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/ AEI / 10.13039/501100011033). In addition, José Antonio García-Díaz was supported by Banco Santander and the University of Murcia through the Doctorado industrial programme.

References

[1] J. A. Garcıa-Dıaz, Using linguistic features for improving automatic text classification tasks in spanish 2802 (2020).

[2] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 641–657.

[3] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2020) 506–518.

[4] J. A. García-Díaz, Á. Almela, R. Valencia-García, 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, 2020, pp. 187–196.

[5] J. A. García-Díaz, R. Valencia-García, 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, 2020, pp. 287–292.

[6] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142.
[7] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of language and social psychology 29 (2010) 24–54.
[8] M. García-Vega, M. C. Díaz-Galiano, M. Á. García-Cumbreras, F. M. P. del Arco, A. Montejo-Ráez, S. M. Jiménez-Zafra, E. M. Cámara, C. A. Aguilar, M. Antonio, S. Cabezudo, et al., Overview of tass 2020: introducing emotion detection (2020).
[9] M. E. Aragón, H. J. Jarquín-Vásquez, M. Montes-Y-Gómez, H. J. Escalante, L. V. Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, G. Bel-Enguix, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish., in: IberLEF@ SEPLN, 2020, pp. 222–235.
[10] M. Montes, P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Álvarez Mellado, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez Adorno, Y. Gutiérrez, S. M. Jiménez Zafra, S. Lima, F. M. Plaza-de Arco, M. Taulé, Proceedings of the iberian languages evaluation forum (iberlef 2021), in: CEUR workshop, 2021.
[11] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021).
[12] F. M. Plaza-del-Arco, S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021).
[13] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 (2020).
[14] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[15] J. Meaney, S. R. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, Semeval 2021 task 7, hahackathon, detecting and rating humor and offense, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.
[16] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish, Procesamiento del Lenguaje Natural 67 (2021).
[17] F. M. Plaza-del-Arco, M. Casavantes, H. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes-y-Gómez, H. Jarquín-Vásquez, L. Villaseñor-Pineda, Overview of the MeOffendEs task on offensive text detection at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021).
[18] S. M. Jiménez-Zafra, Negation processing in spanish and its application to sentiment analysis, Procesamiento del Lenguaje Natural 66 (2021) 193–196.
[19] S. M. Jiménez-Zafra, N. P. Cruz-Díaz, M. Taboada, M. T. Martín-Valdivia, Negation detection for sentiment analysis: A case study in spanish, Natural Language Engineering 27 (2021) 225–248.
[20] S. M. Jiménez-Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. Urena-López, M. A. Martí,

Sfu review sp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns, Language Resources and Evaluation 52 (2018) 533–569.

[21] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M. T. M. Valdivia, L. A. U. Lopez, Detecting negation cues and scopes in spanish, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 6902–6911.

[22] Y. Rychener, X. Renard, D. Seddah, P. Frossard, M. Detyniecki, Sentence-based model agnostic NLP interpretability, CoRR abs/2012.13189 (2020). URL: https://arxiv.org/abs/2012.13189. arXiv:2012.13189.

[23] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowl. Inf. Syst. 62 (2020) 2105–2137. URL: https://doi.org/10.1007/s10115-019-01425-3. doi:10.1007/s10115-019-01425-3.

[24] M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in twitter: A psycholinguistic-based approach, Knowl. Based Syst. 128 (2017) 20–33. URL: https://doi.org/10.1016/j.knosys.2017.04.009. doi:10.1016/j.knosys.2017.04.009.