UNED at PoliticEs 2022: Testing Approximate Nearest Neighbors and Spanish Language Models for Author Profiling in Political Ideology*

Alvaro Rodrigo, Hermenegildo Fabregat and Roberto Centeno

NLP & IR group at UNED, Madrid, Spain

Abstract

PoliticES at IberLef 2022 proposes to profile users in a political domain according to their tweets. In our participation, we have tested two different approaches and some combinations of them. Our first approach is based on approximate nearest neighbours, which obtains low scores for individual results but a great score when combining several outputs. On the other hand, we have also tested some BERT-based systems to study the performance of such technologies in this task. While these systems obtain the best score of our submissions (0.74), we still want to study how to combine them with the first approach to take advantage of the best features of each approach.

Keywords

Sentence Encoder, Spanish Language Models, Approximate Nearest Neighbors

1. Introduction

PoliticES at IberLef 2022 is a shared task that proposes to profile users given a set of tweets [1]. More in detail, given a set of tweets for each user, systems must detect their gender, profession and political ideology. While gender and profession are proposed as a binary classification task, political ideology is proposed both as a binary and multi-class task.

Our team had already participated in tasks oriented to profile users based on their tweets, as for example for detecting fake-news spreaders [2] or detecting anorexia trends in forums [3]. Moreover, we also proposed a system for detecting stance in tweets based not only on textual content, but also on social-media information [4]. Given that PoliticES only provides textual data, we have focused on proposing different methods based on text, combining them and evaluating their potential for this task.

On one hand, we have tested an approach based on representing users combining their messages and tag them according to the label or their nearest neighbours. In this approach, we have tested different representations and ways for tagging users.

On the other hand, we have tested the performance of different transformer-based models. For this approach, we have fine-tuned some models pre-trained over different collections. Our aim with this approach was to (1) evaluate the performance of different BERT-based Spanish

IberLEF 2022, September 2022, A Coruña, Spain.

alvarory@lsi.uned.es (A. Rodrigo); gildo.fabregat@lsi.uned.es (H. Fabregat); rcenteno@lsi.uned.es (R. Centeno)
0000-0002-6331-4117 (A. Rodrigo); 0000-0001-9820-2150 (H. Fabregat); 0000-0001-9095-4665 (R. Centeno)

models, (2) evaluate multilingual models for a Spanish task and (3) compare these models among them.

For all our experiments, we have used the dataset provided by the organizers [5], without the addition of any external resource.

Our results show that while individual runs using the nearest neighbours approach obtain low scores, their combination achieves a promising result (our second best score). On the other hand, the BERT-based models obtained similar results, with the best score achieved by a model pre-trained on twitter.

The rest of this paper is structure as follows: In Section 2 we describe the approach based on nearest neighbours, while in Section 3 we describe the approach based on BERT models. Then, we described the submitted runs in Section 4 and give results in Section 4. Finally, we give some conclusions and future work in Section 6.

2. Approximate Nearest Neighbors Approach

Our first approach is based on Approximate Nearest Neighbors (ANN), given its efficiency when processing large data collections, such as those in social networks. Besides, we have already tested this approach in other tasks [6].

The first step is to represent texts using the Universal Sentence Encoder [7]. We use a multilingual model trained using a Deep Average Network (DAN) [8]. The encoder outputs a 512-dimensional vector for each input text. We have tested two variants of this encoder: one based on transformers and a second one based on CNNs.

Since we have several messages for each user, we have tested two approaches for representing a user: (1) we concatenate all the messages of a user in a single text and encode this text and; (2) we represent each user as the average of the embeddings of all their messages.

In this approach, we work under the assumption that among the target classes there could be different topics associated to each label. Then, we can tag users by using the label of users talking about similar topics. In the development period, we check our assumption by representing texts in a low-dimensional space applying TSNE over the encoded texts. We show in Figures 1 and 2 two examples of these representations, where we can see some kind of distinction between labels in each class.

Instead of using a method based on brute force for ANN, we have applied a non-exact technique based on the use of trees. Thus, the results can be easily scaled to larger amount of data. We have employed the Annoy library, which uses tree-like structures for the representation of nodes and random projections for the division of the subspace between adjacent nodes. We have used a space generated by the inner-dot product of the L_2 normalized vectors generated by the Universal Sentence Encoder.

Once the training set is encoded and the nearest neighbor index is generated using Annoy, we label users in the test set based on the labels of their neighbors. For each class, the approach is as follows:

- Profession: given a new user, we tag them with the profession of their nearest neighbour.
- Gender: given a new user, we tag them with the gender of their nearest neighbour.

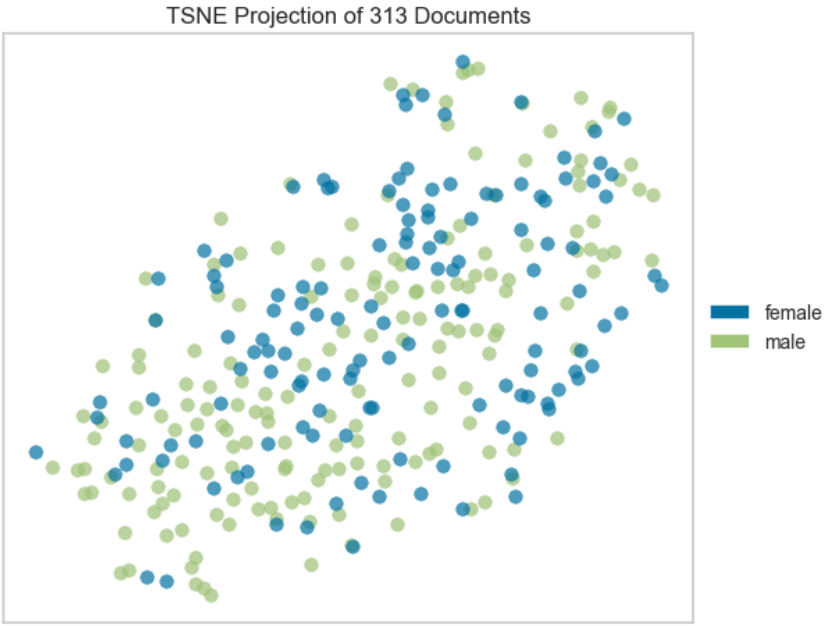

Figure 1: t-SNE representation of documents for the gender label

- Binary ideology: given a new user, we tag them with the binary ideology more repeated among the three nearest neighbours.
- Binary multi-class: given a new user, we tag them with the multi-class ideology more repeated among the fifteen nearest neighbours.

Furthermore, we have tested another approach for tagging the ideology of a user, both binary and multi-class, where we consider that class as dependant of gender and profession. That is, we retrieve the nearest neighbours with the same gender and profession given to the input user.

3. BERT-based Approach

Our second approach is based on fine-tuning different transformer-based models. Our objectives for applying this approach were to test current state-of-the-art BERT-based models for this task and to study if a combination of these models can outperform results of individual systems.

We have performed experiments with the following systems:

- bertin-project-bertin-roberta-base-spanish [9]: RoBERTa-base model trained on the Spanish portion of mC4 [10].
- PlanTL-GOB-ES-roberta-base-bne [11]: RoBERTa-base model trained with data from the National Library of Spain (BNE).

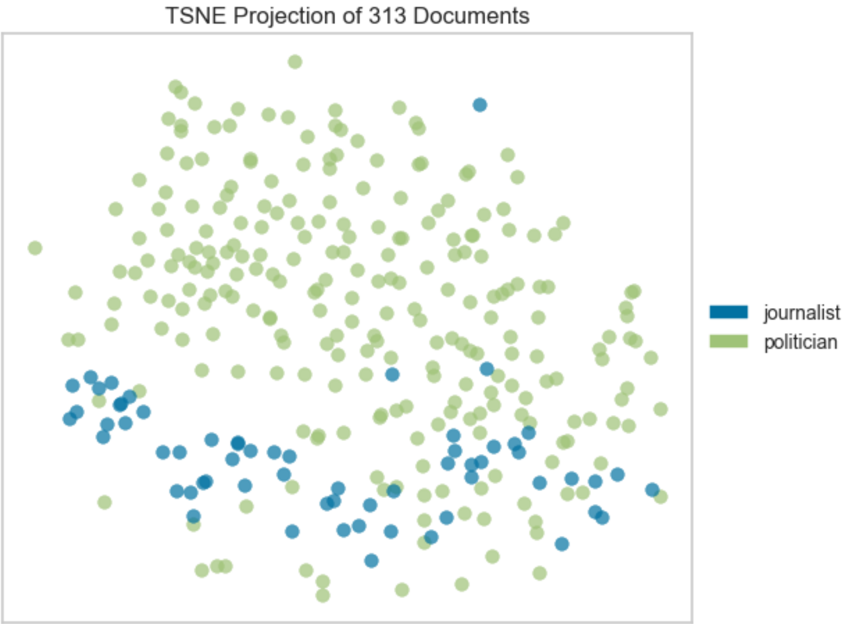

Figure 2: t-SNE representation of documents for the profession label

- bert-base-multilingual-uncased [12]: BERT system trained using different languages including Spanish. By using this system we wanted to test the performance of a multilingual model in this task.
- pysentimiento-robertuito-base-uncased [13]: RoBERTa-base model trained on 500 million Spanish tweets. Thus, we can test the performance of a system trained on tweets instead of more formal texts.
- cardiffnlp-twitter-xlm-roberta-base: RoBERTa-base model trained in 198M multilingual tweets [14]. Here, we can test a multilingual model trained on tweets.

We have fine-tuned each one of these systems for 2 epochs on the training set, given that we detected a lost in performance when fine-tuning for 3, or more, epochs.

4. Submitted Runs

We have submitted 10 runs. We have selected our runs for testing different approaches, giving more importance to the ANN approach and trying to combine different systems. The description of each run is as follows:

- Run 1: CNN version of the Multilingual Sentence Encoder (MSE), average messages of each user and tag ideology independently from other classes.

- Run 2: similar to run 1 except that we tag ideology depending on profession and gender.
- Run 3: CNN version of the MSE, concatenate messages of each user and tag ideology independently from other classes.
- Run 4: similar to run 3 except that we tag ideology depending on profession and gender.
- Run 5: transformers version of the MSE, average messages of each user and tag ideology independently from other classes.
- Run 6: similar to run 5 except that we tag ideology depending on profession and gender.
- Run 7: in this run we apply a voting scheme of different outputs based on ANN depending on: the version of the MSE, if we average or concatenate messages, if we obtain gender independent or dependant from other classes. The most common predicted value among all the systems is chosen as the final prediction. Thus, this run considers the output of previous runs and additional outputs.
- Run 8: we use the output from a fine-tuned RoBERTuito system. This was the BERT-based model that gave us the best results in the development period.
- Run 9: this run combines the 5 BERT-based systems, described in Section 3, based on a voting scheme where the most predicted value is given.
- Run 10: this runs combines the two main approaches followed in this work. More in detail, we combine the first 6 runs with the 5 BERT-based systems. The value more predicted for each class is given as the final output of the run.

In our submitted runs we can see a first group of runs based on ANN, where we have selected 6 systems for the first 6 runs and a combined system based on a voting scheme (run 7). The second group contains runs based on BERT models, with one run for the best model in the development period (RoBERTuito as run 8) and a combination of several models (run 9). Finally, we have also submitted a combination of both approaches (run 10).

By submitting these runs we wanted to test the differences when using different approaches for ANN, and compare them with the results of BERT-based models.

5. Analysis of Results

We show the results of our 10 runs in Table 1. The Table contains F1 results for each class, as well as results of the main evaluation measure (the macro-average f1 of all classes). We have highlighted the best score for each measure. The Table also contains the scores of the best participant, which ranks the best for all the measures except ideology multiclass, where they ranked fourth.

In Table 1, we can see that our best results are given by runs 7 to 10, with the best score (0.74) obtained by RoBERTuito (run 8). With this score, we ranked 13th in the official leaderboard. As we can see in the Table, we are far from the best participant, who obtain an f1 score of at least 0.90 for all the measures, except multiclass ideology.

Regarding runs from the ANN approach, the scores of individual runs (from run 1 to run 6) are lower than the scores of other systems. It seems that the information extracted from the encoder is insufficient to model features for defining each class. We think this is due to the lack of a explicit training and the simplicity of the inference method applied, which is only based on

run	average macro f1	f1 gender	f1 profession	f1 ideology binary	f1 ideology multiclass
best	0.902262	0.902868	0.944327	0.961623	0.800229
run1	0.419777	0.540938	0.428312	0.462267	0.247593
run2	0.367752	0.353140	0.499773	0.436090	0.182003
run3	0.426207	0.579093	0.432870	0.452310	0.240556
run4	0.450610	0.514339	0.486883	0.510991	0.290227
run5	0.451989	0.428904	0.597701	0.538828	0.242522
run6	0.443832	0.490093	0.463602	0.575059	0.246575
run7	0.738497	0.711312	**0.833309**	0.810739	**0.598626**
run8	**0.740889**	0.747162	0.833309	0.818269	0.564815
run9	0.726057	0.732143	0.765326	**0.839939**	0.566822
run10	0.732241	**0.752783**	**0.833309**	0.799924	0.542949

Table 1
Results of the 10 runs. The best score for each measure is highlighted. We also include scores of the best participant.

retrieving the nearest neighbours. Given that results using BERT-based models are higher, we think the ANN approach could benefit from pre-processing the embeddings obtained from the encoder. We want to study this kind of pre-processing.

On the other hand, the combination of several outputs from this approach, in run 7, has obtained results similar to our best approaches. Thus, it seems that the different outputs are complementary or there are some outputs with better results than the selected as the final runs. We want to explore these observations to obtained more information about this approach.

With respect to BERT-based models, they have obtained the best performance of our runs, with a F1 score close to 0.75. However, the combination of these systems was unable to outperform the score of RoBERTuito. We leave as future work a deeper study about the results of each model and how to combine them in other ways.

6. Conclusions and Future Work

Profiling users in social media is an interesting task that is attracting research, as it is shown by the proposal of evaluation tasks. In this paper, we describe our participation in PoliticES at IberLEF 2022, where users are profiled in the political domain.

We have tested two approaches. Our first approach is based on tagging users according to the labels of their nearest neighbours, represented using the Multilingual Sentence Encoder. This approach has shown promising results when combining different outputs from different settings, but it requires a deeper study about how to use it in this task.

Our second approach is based on fine-tuning BERT-based systems pre-trained on different datasets. This approach obtains the best results in our experiments. We plan as future work to study how to combine these systems among them to outperform individual results, as well as how to include information from these systems in the first approach.

Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER,UE), the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under grant PID2019-106942RB-C32 and the projects given to the *ISL: Intelligent Systems for Learning* group (GID2016-39) in the PID 20/21 and 21/22 calls.

References

[1] J. A. García-Díaz, S. M. J. Zafra, L. A. U. L. María Teresa Martín Valdivia, Francisco García-Sánchez, R. Valencia-García, Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022).

[2] M. S. Espinosa, R. Centeno, Á. Rodrigo, Analyzing user profiles for detection of fake news spreaders on twitter, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_132.pdf.

[3] R. Masood, M. Hu, H. Fabregat, A. Aker, N. Fuhr, Anorexia topical trends in self-declared reddit users, in: I. Cantador, M. Chevalier, M. Melucci, J. Mothe (Eds.), Proceedings of the First Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020, volume 2621 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[4] M. S. Espinosa, R. Agerri, Á. Rodrigo, R. Centeno, Deepreading @ sardistance 2020: Combining textual, social and emotional features, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[5] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[6] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, UNED-NLP at eRisk 2022: Analyzing gambling disorders in Social Media using Approximate Nearest Neighbors, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2022.

[7] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, 2018. URL: https://arxiv.org/abs/1803.11175. doi:10.48550/ARXIV.1803.11175.

[8] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 87–94.

[9] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[10] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498.

[11] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[13] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of The Language Resources and Evaluation Conference, LREC, 2022.

[14] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: A multilingual language model toolkit for twitter, in: Proceedings of The Language Resources and Evaluation Conference, LREC, 2022.