HUMAN LANGUAGE TECHNOLOGIES

University of Pisa, Department of Computer Science

# Project Report

Academic Year 2022-2023

*Filippo Lari*, 545736 - f.lari@studenti.unipi.it
Msc in Computer Science

*Leonardo Venuta*, 546002 - l.venuta@studenti.unipi.it
Msc in Computer Science

# 1 Introduction

Political ideology is a psychographic trait that can be used to understand individual and social behavior, including moral and ethical values as well as inherent attitudes, appraisals, biases, and prejudices. For instance, a study found that conscientiousness was strongly correlated with the right wing, whereas openness to experience and agreeability were notably more correlated with the left wing. The goal of EVALITA 2023 PoliticIT challenge [1] is to develop an automated method to extract self-assigned gender as a demographic trait, and political ideology as a psychographic trait from a set of tweets written in Italian from several authors that share those traits. This report is structured as follows: In Section 2 we explore the dataset of the challenge highlighting its main properties. In Section 3 we fine-tune two well-known versions of the BERT model for the Italian language. In Section 4 we build two models that only use the embeddings of the fine-tuned BERTs and present the final results of the challenge.

# 2 Data Exploration

The EVALITA PoliticIT dataset [1] consists of tweets from Italian politicians collected during 2020 and 2022 using the *UMUCorpusClassifier* [2]. Starting from these tweets, a set of clusters has been created by composing together tweets by different users that share the following traits: self-assigned gender (male or female), and the political spectrum, which can either be binary (left or right) or multiclass (left, moderate left, moderate right and right). The resulting set of tweets was further preprocessed by removing the Twitter mentions of the politicians or other accounts replacing them with the token *@user*, therefore preventing any model from simply learning to associate a given trait only based on the politician's name. As far as the challenge is concerned, the dataset was divided into a training and test set whose statistics are reported in Table 1.

| | Training | Test |
|---|---|---|
| Clusters | 1298 | 453 |
| Tweets per cluster | 80 | 80 |
| Total tweets | 103840 | 36240 |

Table 1: Distribution of tweets and clusters of the Evalita PoliticIT 2023 training and test sets.

As part of the data exploration phase, we examined the distribution of the labels among the training set obtaining the results of Figure 1 from which it is clear that the dataset is not balanced. In particular, this seems to be less pronounced for the gender and binary ideology, while it is significant for the multiclass case, especially for the moderate right and left labels. For classification problems, it is well-known that an unbalanced dataset is problematic since the model could learn to classify only the most frequent class correctly and still get an overall good accuracy.

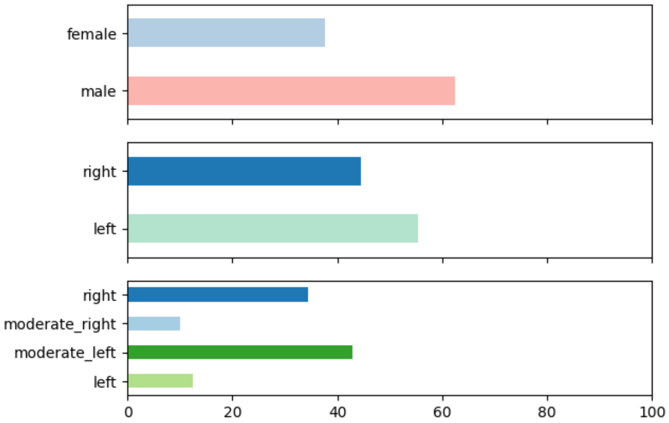

Figure 1: Distribution of tweets for each of the traits under examination: gender (top), binary ideology (middle), and multiclass ideology (bottom)

Finally, Figure 2 presents the most relevant words divided according to political ideology. We can notice that left-wing users refer to the political right, while right-wing users refer to the political left. Considering the right-wing, words related to nationalism such as *italiani* (italians), *Italia* (Italy), *Roma* (Rome), and *difesa* (defense) are the most frequent. Similarly, we can observe that for the left-wing the most frequent words are *Europa* (Europe), *giovani* (youth), *democrazia* (democracy), and *libertà* (freedom). It is also interesting to notice that for the right and moderate right, the pre-processing done by the authors of the challenge failed to remove some references to political parties such as *Fratelli d'Italia* and its acronym *FI*.

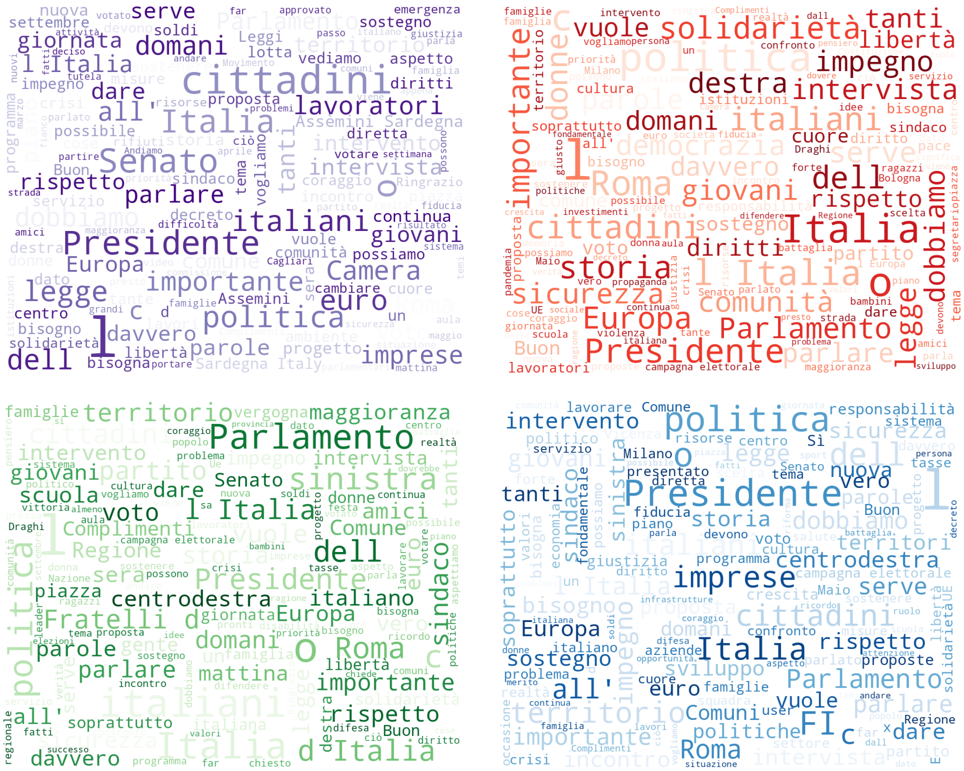

Figure 2: Most relevant word clouds by political spectrum: left (top left), moderate left (top right), moderate right (bottom left), and right (bottom right).

# 3 Proposed Models and Preliminary Results

Our first proposal is a simple model based on a *Bag-of-Words* (BoW) representation of Tweets with a vocabulary size of 5000 words. The resulting model is composed of a dense layer for the classification using the *sigmoid* and *softmax* activation functions respectively for the binary and multiclass tasks. Because of its simplicity, this model represents the baseline of our proposals and is used to measure and understand the advantage of using more complex approaches.
As already pointed out in Section 2, the dataset is heavily unbalanced for multiclass ideology, therefore in measuring the performance of the baseline and any of the following models, we will take into account the macro average $F1$-*score* for each class which is defined as follows:

$$F1 = \frac{1}{3} \cdot \left(F1_{gender} + F1_{binary\ ideology} + F1_{multiclass\ ideology}\right)$$

This measure is preferable in the case of unbalanced datasets because it penalizes models that perform well only on the most frequent class. Considering this measure as the performance evaluation metric, the results of the baseline model are presented in Table 2.

| Model | Average | Gender | Binary ideology | Multiclass ideology |
|---|---|---|---|---|
| Baseline | 0.43 | 0.48 | 0.54 | 0.28 |

Table 2: The average F1-score for the baseline model on the three classification tasks of the EVALITA 2023 PoliticIT challenge.

To improve on the baseline our proposal considers three models based on BERT [3], a popular language representation model by the Google AI research team. The motivations behind this choice are multiple:

- Being a deep neural network that produces word embeddings from unstructured natural language texts, BERT follows the rationale of modern feature-based approaches for building NLP models: first, a general purpose model is obtained by applying different training techniques on a large number of unlabeled texts; then the pre-trained model is made available to be fine-tuned on a variety of downstream NLP tasks without resorting to major changes to the initial architecture.
- Since its introduction BERT-based models have reached the state-of-the-art in numerous NLP tasks including the QA Stanford Question Answering Dataset (SQuAD v1.1) and Natural Language Inference (MNLI).
- Unlike other language models like word2vec or GloVe, BERT generates contextualized embeddings that can capture semantics based on the words in each surrounding.
- Although it is one of the first large language models, it still has a reasonable size that allows its fine-tuning to be completed within a few hours on Google Colab's free GPUs.

Since the EVALITA 2023 PoliticIT challenge concerns the classification of tweets written in Italian, and because of the aforementioned points, the three proposed models are based on two different pre-trained BERTs: AlBERTo [4] and Italian BERT XXL [5], which share the same architecture of the original BERT but present the following differences:

- **AlBERTo**, is a BERT language understanding model for the Italian language, in particular, is focused on the language used in social networks, specifically on Twitter. The training has been performed with 200 million Tweets from a huge corpus of Tweets in the Italian language. To demonstrate its robustness, AlBERTo was evaluated on the EVALITA 2016 task SENTIPOLC (SENTIment POLarity Classification) [6] obtaining state-of-the-art results in subjectivity, polarity, and irony detection on Italian tweets.

- **Italian BERT XXL (It-BERT)**, is a BERT model for the Italian language trained on a recent Wikipedia dump, various texts from the *OPUS corpora collection* [7] and the Italian part of the *OSCAR corpus* [8], with an overall size of 81GB and approximately 13 billion of tokens. To test its effectiveness, it has been evaluated on the EVALITA 2016 task of Part-of-Speech tagging of Social Media [9] obtaining state-of-the-art results.

Given the size of these models and the limitations imposed by the Google Colab environment, we decided to first fine-tune AlBERTo and It-BERT by slightly adjusting their weights to better perform on our task. Both models share the same architecture, the *pooled output* is used as an embedding of the given Tweet and is first processed with a dropout layer using a rate of 0.1, then three dense layers are used to classify gender, binary ideology, and multiclass ideology using respectively a sigmoid activation function for the two binary outputs and a softmax for the multiclass case.

Both models are trained for 4 epochs at the end of which the weight configuration that obtained the lowest loss on the validation set is retained. The hyperparameters used for training are the standard ones suggested by BERT, a batch size of 32 paired with the well-known *Adam optimizer* [10] configured with a learning rate $\eta = 2e^{-5}$, and initial decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

Since in the dataset we have multiple tweets for the same cluster, to correctly evaluate the aforementioned models, we implemented a voting system that for each Tweet of a cluster assigns the gender and political ideology based on a majority vote among the predicted results of each Tweet.

Considering instead the class-imbalance problem pointed out by Figure 1, a common solution is to use so-called *class weights*, which act on the loss computation by giving a larger weight to those classes with fewer examples and a smaller one to the ones with many examples. Since the fine-tuning of each model is an expensive process, we decided to test the usage of class weight only on it-BERT and adopt the best solution on AlBERTo. Table 3 presents the results of this comparison in terms of the F1-score on the validation. As expected, we notice that the class weights help to obtain a better score on the heavily imbalanced multiclass political ideology class, unfortunately, this gain is not sufficient to obtain an overall advantage over the model without the use of class weights.

| Model | Average | Gender | Binary Ideology | Multiclass Ideology |
|---|---|---|---|---|
| it-BERT | **0.86** | **0.86** | **0.91** | 0.82 |
| it-BERT + weights | 0.84 | 0.83 | 0.91 | **0.84** |

Table 3: The F1-score obtained by it-BERT on the validation set without class weights (first row) and with class weights (second row).

After addressing the class-imbalance problem, we fine-tuned AlBERTo without using the class weights and obtaining the results of Table 4.

| Model | Average | Gender | Binary Ideology | Multiclass Ideology |
|---|---|---|---|---|
| it-BERT | **0.86** | **0.86** | **0.91** | **0.82** |
| AlBERTo | 0.80 | 0.77 | 0.88 | **0.82** |

Table 4: The F1-score obtained by it-BERT and AlBERTo on the validation set.

# 4 Extending the Model

After the fine-tuning of Section 3, from Table 4 it is clear that it-BERT performs better than AlBERTo in terms of the average F1-score on the validation set. Therefore we froze all the it-BERT weights and built an architecture on top of it that uses ts embeddings after being through a dropout layer and consists of a dense layer with a *ReLu* activation function, a subsequent dropout layer, and finally, for each of the three classification tasks another dense layer using the sigmoid and softmax activation functions respectively for the gender, binary ideology and multiclass ideology. For the sake of clarity, Figure 3 presents this architecture.

After choosing the architecture of this model, we performed a hyperparameter search guided by the Optuna optimization framework [11]. Starting from a set of possible hyperparameter values, this framework allowed us to explore the hyperparameter space smartly, by using a Gaussian Mixture Model to set up the hyperparameter values to be tested and by pruning unpromising trials.

The hyperparameter space explored by the optimizer is described in Table 5 and is based on preliminary manual trials. Notice that the number of epochs is not in the hyperparameter space because we used the mechanism of early stopping, which allows us to stop training after the score on the validation set did not improve for 2 consecutive

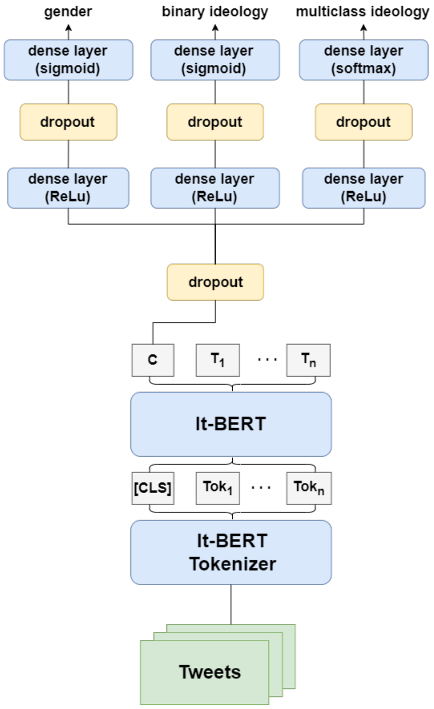

Figure 3: The extension of it-BERT model that only uses its embeddings while keeping its weights frozen.

epochs. The metric we used for the hyperparameters optimization is the F1-score on the validation set.

| Hyperparamters | |
|---|---|
| Learning rate | $[2e^{-5}, 5e^{-5}]$ |
| Batch size | 16, 32 |
| Dropout | [0.1, 0.3] |
| Number of units | 150, 300, 700 |
| Weight decay | [0.02, 0.04] |

Table 5: Hyperparameter space explored with the Optuna framework.

After the hyperparameter search, we re-trained the model with the best configuration of hyperparameters given by Optuna obtaining the results of Table 6. Unfortunately, the model did not improve significantly over fine-tuning, achieving only marginal improvement on the multiclass ideology.

| Model | Average | Gender | Binary ideology | Multiclass ideology |
|---|---|---|---|---|
| it-BERT final | 0.86 | 0.86 | 0.91 | 0.83 |

Table 6: The F1-score on the validation set of the it-BERT model using the best hyperparameters configuration discovered by Optuna.

As a last attempt, since AlBERTo results are not significantly lower than the ones of it-BERT we built a second model that freezes the weights of both AlBERTo and it-BERT, concatenates their embeddings, passes them through a dropout layer, and then uses the same architecture as described above obtaining the model of Figure 4. This idea is inspired by a recent work on the PoliticES 2022 challenge [12], notice that the main difference with our implementation is that, because of the limitations imposed by Google Colab, we were not able to train the full model, instead we used the fine-tuned AlBERTo and it-BERT from Section 3 and just trained the remaining parts of the architecture. As in the case of it-BERT we explored the same hyperparameters space using Optuna. The final results on the validation set are presented in Table 7, from which we can see that the combination of the two models outperforms the single it-BERT.

| Model | Average | Gender | Binary ideology | Multiclass ideology |
|---|---|---|---|---|
| it-BERT + AlBERTo | **0.92** | **0.90** | **0.96** | **0.91** |
| it-BERT final | 0.86 | 0.86 | 0.91 | 0.83 |

Table 7: The F1-score on the validation set of it-BERT and the combination of it-BERT and AlBERTo using the best combination hyperparameters.

# 5 Final Results

After selecting the combination of it-BERT and AlBERTo, we evaluated our final model's performance on the test set, which led us to conclude the challenge. The final results, along with our virtual participation in the EVALITA 2023 PoliticIT challenge, are presented in Table 8 where our team *LabronicTeam* would secure a good third place.

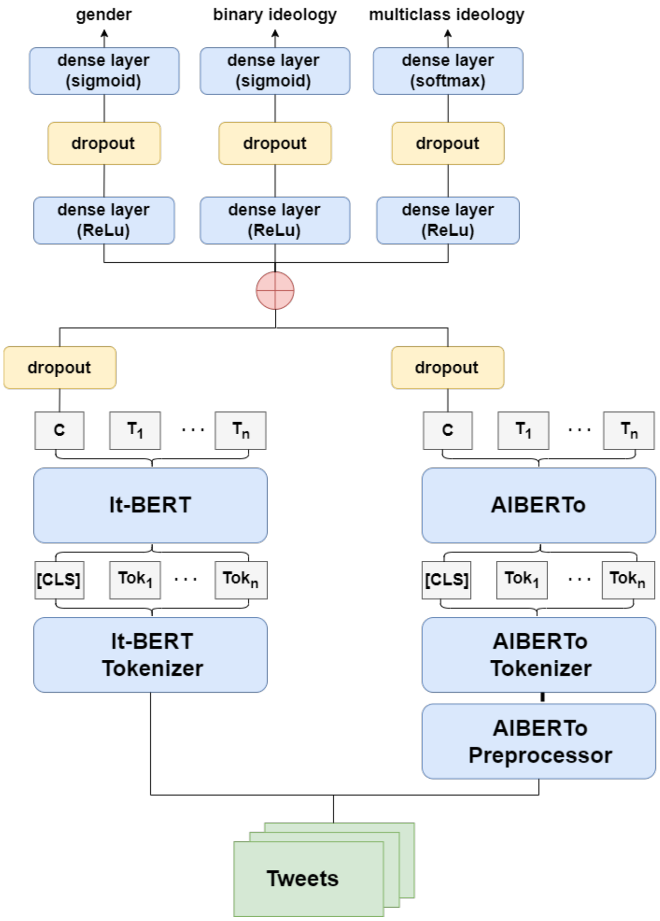

Figure 4: The combination of it-BERT and AlBERTo inspired by [12].

| Team Name | Average | Gender | Binary Ideology | Multiclass Ideology |
|---|---|---|---|---|
| 🥇 TuebingenPoliticIT | **0.824** | 0.792 | **0.928** | **0.751** |
| 🥈 INFOTEC-LaBD | 0.800 | **0.824** | 0.860 | 0.717 |
| 🥉 **LabronicTeam** | 0.785 | 0.780 | 0.881 | 0.694 |
| extremITA | 0.771 | 0.769 | 0.925 | 0.621 |
| INGEOTEC | 0.762 | 0.732 | 0.848 | 0.705 |
| MattiaSangermano | 0.751 | 0.752 | 0.887 | 0.613 |
| ronghao | 0.704 | 0.712 | 0.866 | 0.533 |
| NLP_URJC | 0.684 | 0.661 | 0.770 | 0.621 |

Table 8: The virtual leader board at the end of the EVALITA 2023 PoliticIT challenge, comparing the results of our final model with the ones of the other partecipants.

In conclusion, we would like to point out two aspects on which we could have improved our proposal:

- We could have considered each task separately, thus developing a single model for the different classification tasks, without forcing the same embedding to be a good representation for all the three classes.

- Without the resource limitations of Google Colab we would have been able to fine-tune the entire models presented in Section 4, possibly inside the same Optuna grid-search, in order to use the best configuration of hyperparameters.

- From the final results of Table 8 it is clear that our model performs poorly on the multiclass ideology, which happens to be the most imbalanced class. Because of this, we could have experimented with more sophisticated methods of *data augmentation* for the aforementioned class, like oversampling mechanisms that create new tweets for the imbalanced classes by substituting words with their synonyms. Alternatively, a common practice in this type of challenge is to augment the dataset by collecting other examples following the same method as the authors.

# References

[1] Political ideology detection in italian texts. https://www.evalita.it/campaigns/evalita-2023/tasks/. Last accessed: 2023-08-28.

[2] José Antonio García-Díaz, Ángela Almela Sánchez-Lafuente, Gema Alcaraz Mármol, and Rafael Valencia García. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, 2020-09.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Association for Computational Linguistics, 2019.

[4] Marco Polignano, Pierpaolo Basile, Marco Degemmis, Giovanni Semeraro, and Valerio Basile. AlBERTo: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Italian Conference on Computational Linguistics*, 2019.

[5] Italian BERT. https://huggingface.co/dbmdz/bert-base-italian-cased. Last accessed: 2023-08-28.

[6] Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. Overview of the evalita 2016 sentiment polarity classification task. In *CLiC-it/EVALITA*, 2014.

[7] OPUS corpora collection. https://opus.nlpl.eu/. Last accessed: 2023-08-30.

[8] OSCAR corpus. https://oscar-project.org/. Last accessed: 2023-08-30.

[9] PoSTWITA task. https://corpora.ficlit.unibo.it/PoSTWITA/. Last accessed: 2023-08-30.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[11] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19, page 26232631, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Sergio Santamaria Carrasco and Roberto Cuervo Rosillo. LosCalis at PoliticEs 2022: Political author profiling using BETO and MarIA. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain*, 2022.