INFOTEC-LaBD at PoliticIT: Political Ideology Detection in Italian Texts

Hiram Cabrera-Pineda[1,*], Eric Sadit Téllez[1,2,3] and Sabino Miranda[1,3,4]

[1]INFOTEC Aguascalientes, México
[2]CICESE, Ensenada, Baja California, México
[3]CONAHCYT, México
[4]UPIITA-IPN, Ciudad de México, México

Abstract

This working notes presents our approach to clusters of texts profiling in the PoliticIT 2023 challenge, utilizing a non-linear low-dimensional representation of term distribution entropy. Our proposed algorithm is designed to learn a 3-dimensional model of text data, effectively capturing essential features for accurate profiling. Furthermore, it offers valuable insights through cluster analysis and visualizations, enabling a deeper understanding of the underlying patterns. The method employed in our algorithm uses a bag-of-words representation and incorporates weighting schemes based on the term's distribution entropy. By leveraging these techniques, we are able to extract meaningful information and uncover significant characteristics related to clusters of texts profiling. To evaluate the effectiveness of our proposed algorithm, we conducted experiments on the PoliticIT 2023 dataset, encompassing three tasks: gender identification, and binary and multiclass political ideology classification. The obtained results demonstrate the competitiveness of our solution across all three tasks, highlighting its efficacy in accurately predicting the attributes of clusters of texts. One notable advantage of our algorithm is its explainability. It offers insights into the reasoning behind its predictions, allowing users to understand the factors influencing cluster of text behavior. This transparency enhances the practicality and utility of our algorithm as a powerful tool for cluster of text profiling and behavior analysis.

Keywords

explainable user-profiling, low-dimensional representations, political parties identification, term's distribution entropy weighting

1. Introduction

User profiling is a powerful technique used to extract valuable information about individuals, including their interests, demographics, behavioral patterns, and even their political preferences. Cluster profiling aims to identify patterns, trends, and similarities within groups of texts written by different users with similar traits or characteristics based on specific criteria. By carefully analyzing data related to users, we can gain deep insights into their unique characteristics and preferences. These insights have far-reaching applications in various domains, including enhancing user experiences, personalized advertising, detecting and preventing malicious activities, and gaining a deeper understanding of societal dynamics and preferences. In this manuscript, our focus lies specifically on political preferences as participation in the PoliticIT challenge of the EVALITA 2023 forum.

The task consists of predicting user demographics like gender and political ideology from 36,240 Twitter messages in the Italian language.These tweets cover various topics, from news and current events to personal thoughts and experiences. Twitter is a valuable data source for user profiling, offering valuable insights into users' interests, demographics, and behaviors. The practical applications of these tasks span various domains, including personalized marketing, content recommendation, and social science research [1].

Automatic user profiling gives us insight into a population's characteristics, preferences, and ideological orientations. There are several examples of this, including the PAN@CLEF and FIRE series. These tools cover a wide range of objectives such as age, gender, language variety identification, and personality recognition in different languages and genres, like blogs, reviews, social media, and Twitter [2, 3, 4, 5, 6]. Forums like MEX-A3T [7] delve into identifying occupation types and places of residence. At the same time, the IberLEF@SEPLN forum focuses on gender, profession, and political ideology profiling, see [8, 9], all of which contribute significantly to the advancement of user profiling techniques.

This paper provides an overview of the author profiling task, focusing on its relevance in political domains, specifically in the context of the PoliticIT challenge. Our approach is described in Section 2, where we outline the methodology we employed. In Section 3, we delve into

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
hiram.cabrerap@infotec.mx (H. Cabrera-Pineda); eric.tellez@ieee.org (E. S. Téllez); smiranda@ieee.org (S. Miranda)
0000-0003-1419-761X (H. Cabrera-Pineda); 0000-0001-5804-9868 (E. S. Téllez); 0000-0002-9421-8566 (S. Miranda)

the details of our dataset and models, providing a brief analysis. Section 4 presents the experimental results and their discussion. Finally, in Section 5, we reflect upon the achievements of our approach and discuss their implications.

2. Our approach

Our primary objective is to develop robust and interpretable representations that achieve high performance and allow for in-depth analysis of the model's label predictions. Instead of relying on a predefined set of features, we generate 3D maps that capture the underlying similarity structure of the original high-dimensional representation. This spatial representation lets us explore similar clusters of texts by directly examining their messages and analyzing their shared vocabulary.

In a previous publication [10], we introduced our overarching approach. To tackle the PoliticIT 2023 challenge, we have streamlined our approach by reducing the number of hyperparameters. Our cluster of texts profiling methodology encompasses three main modules, providing a more concise and focused framework.

Vector spaces models. We implemented various preprocessing steps to prepare the data; for instance, we converted all messages to lowercase, normalized blank spaces to a single space, and removed diacritic marks. Token numbers 1-9 were preserved to capture important information on small numbers, while other numbers were replaced by 0 (to reduce dimensionality). We employed three types of tokens: unigrams, bigrams, and character q-grams of size four. Each token is modeled as a distribution along classes, and we compute the token's weight based on the distribution's entropy using the formulation of [10], that is, for each token t

$$\mathsf{entweight}(t) = 1 + \frac{\sum_{c \in L} p_c^t \log p_c^t}{\log \#L}$$

where L is the set of tables and p_c^t is the probability of token t in class c. Note that the numerator's $\log$ produces negative numbers. Therefore, the weight is bounded between 0 (tokens with low-discrimination power) and 1 (high-discriminant tokens). This formulation ignores smoothing constants as needed by our original formulation. Instead, we reject tokens from the vocabulary if they do not occur in at least M clusters of texts. This change reduces the memory required by our models and speedups computations.

Non-linear dimensional reduction. The non-linear dimensional reduction module uses the Uniform Manifold Approximation and Projection (UMAP) [11] to produce a low-dimensional vector space (UMAP model) from the resulting vector space in the previous Module. We made three-dimensional projections of the data for visualization and as the input of classifiers. UMAP projection requires the k nearest neighbor graph (using our high-dimension vector space and cosine similarity) to capture the dataset's structure. More detailed, a fuzzy smoothed version of the graph is created, and then, the eigenvectors of the normalized Laplacian matrix are computed to initialize low-dimensional embedding vectors that are finally optimized to preserve the k nearest neighbor graph in the low-dimensional projection. This procedure captures similar properties to the spectral clustering [12] while producing insightful visualization. The explainability aspect utilizes UMAP projections to create visualizations for understanding data distribution, cluster relationships, and separability. These visualizations provide intuitive representations of clusters, highlighting proximity, separability, and discernible patterns or trends.

Classification. The supervised learning stage uses the low-dimensional vectors to train a classifier to predict the cluster's political ideology, or gender. We use SVM classifiers with linear and non-linear kernels in the sklearn library [13]. We perform a model selection procedure for tuning each classifier.

3. Methodology and model analysis

This manuscript addresses the challenge of profiling political preferences within the context of the PoliticIT challenge held in the EVALITA 2023 forum. The task involves predicting user gender and political ideology based on a dataset comprising more than 36 thousand Twitter messages written in Italian. To ensure privacy and ethical considerations, an automated clustering approach was employed, grouping tweet messages from different users who shared all the evaluated traits [14].

The provided training corpus consists of 103,840 messages collected from Twitter, aiming to extract the author's traits from Italian texts. This shared task entails gathering demographic traits, i.e., gender and political ideology as psychographic traits. See [14] for more details.

The training dataset includes tweets from 1,298 clusters of texts, each contributing 80 tweets. However, it is important to note that the dataset exhibits varying degrees of class imbalance across different class categories. Considering the gender task, the proportion is 63.4% vs. 37.6% for males and females, respectively. There are 12.5%, 43%, 10.1%, and 34.4% between left, moderate left, moderate right, and right regarding multiclass political ideology. The proportion is 55.4% vs. 44.5% for

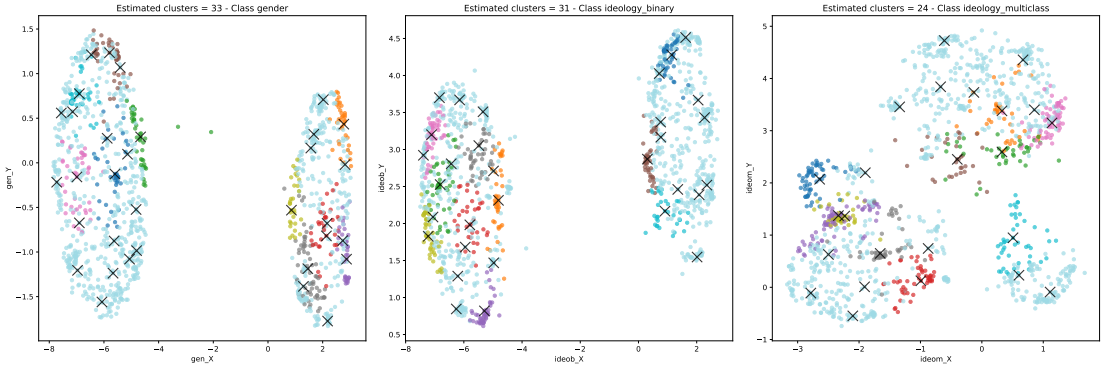

Figure 1: 2D UMAP projections for each task with $M = 5$ and $k = 50$.

binary ideology, i.e., a slight imbalance between the left and right categories, respectively. We are asked to create models to predict these labels for a test dataset, i.e., a list of 453 clusters (80 messages per cluster); the label distribution of the test dataset is unknown.

As explained in Section 2, we create low-dimensional projections that concisely represent the dataset and its associated labels. This visualization serves the purpose of uncovering group properties and revealing any underlying cluster structures. It is important to note that our method utilizes a k nearest neighbor graph constructed from a vector space generated by the entropy weighting model. To ensure meaningful results, we exclude tokens with low frequency, specifically those that appear in fewer than M clusters. Hence, our primary hyperparameters revolve around k and M.

We experimented with different parameter values to create UMAP low-dimensional projections. Specifically, we varied the values of k and M, with k ranging from 10 to 50 and M ranging from 3 to 35. We generate three-dimensional embeddings using a spectral layout for embedding initialization and optimized during 100 epochs. We also used three negative samples per point (user vector).

Based on the combinations of k and M parameters, we obtained 35 unique UMAP projections for each task in the challenge, including gender, binary ideology, and multiclass ideology. These projections provided visual representations that captured the data's underlying structure and relationships.

In Figure 1, we present an overview of different parameter combinations (k and M) and their impact on the data representation. Please note that although our models use three-dimensional projections, the figures shown are two-dimensional for easier visualization. The figures display centers obtained through the Affinity Propagation (AP) clustering algorithm [15] applied to the low-dimensional embeddings. These results offer insights into patterns, separations, and overlaps within the data, aiding the analysis and interpretation. It's important to clarify that the centers represent groups of near clusters of texts, not individual classes.

We used the homogeneity score to evaluate the extent to which the clusters are homogeneous, meaning that the samples within each group are similar to each other [16]. It measures the similarity of the clusters in terms of their composition and whether they predominantly consist of samples from a single ground truth class. A high homogeneity score indicates that the clusters formed by the AP algorithm are internally consistent. However, it's important to note that the homogeneity score does not directly measure how well the predicted clusters match the ground truth labels. It doesn't assess whether the cluster labels perfectly align with the true class labels. Instead, it assesses the consistency and purity of the groups within themselves.

We provided homogeneity heatmaps in Figure 2. They use color intensity to represent the homogeneity scores for different combinations of k and M. Darker colors indicate higher homogeneity.

By examining the homogeneity score heatmaps, we can gain insights into the clustering quality and make better decisions regarding the choice of parameters k and M. This helps us understand how different parameters impact the quality of the clusters.

3.1. Model Selection

Our participation in the PoliticIT challenge involved the development of multiple models using the provided training data. To select the most suitable models, we employed a Grid Search approach and conducted model selection based on maximizing the macro-F1 score. The evaluation process utilized five-fold stratified cross-validation.

For the creation of our classification models, we utilized SVM classifiers with both linear and non-linear ker-

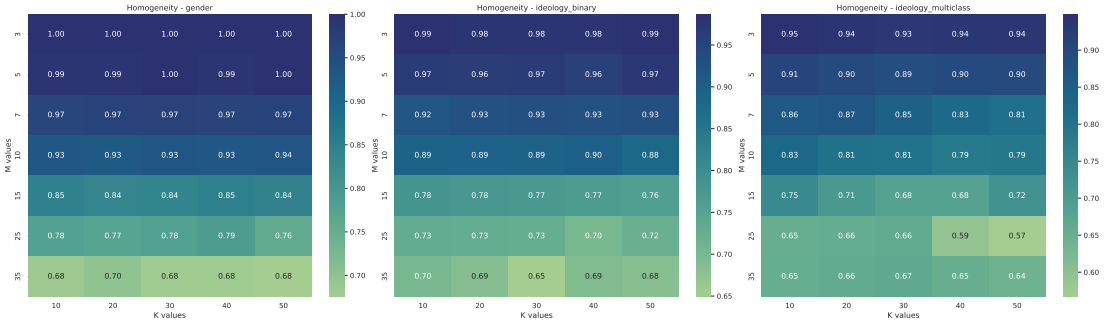

Figure 2: Homogeneity heatmaps for evaluation of low-dimensional clustering.

nels. Through a hyperparameter optimization process and 5-fold cross-validation, we identified the models that exhibited the best performance.

Each model corresponds to a distinct cluster of texts representation defined by parameters k and M. Additionally, we considered three-dimensional representations and the concatenation of the three 3D maps from all tasks, resulting in a 9-dimensional vector space.

Through this comprehensive evaluation, we thoroughly assessed the performance of various models and determined the most effective approaches for each task. The final selection of models was based on the parameter combination that achieved the highest accuracy during cross-validation.

4. Experimental results

In this section, we present the experimental results of our approach for the EVALITA 2023 PoliticIT task. The source code can be found at the following GitHub repository: https://github.com/hiramcp/PoliticIT2023.

Our experiments were conducted on a Windows 10 operating system, utilizing a four-core Laptop with 32 GB of RAM and an Intel Core i7-1165G7 @ 2.80GHz processor. To compute the vector space and perform preprocessing functions, tokenization, and entropy-based weighting, we utilized *TextSearch.jl* Julia package available at https://github.com/sadit/TextSearch.jl. For UMAP projections, we employed the *SimSearchManifoldLearning.jl* Julia package found at https://github.com/sadit/SimSearchManifoldLearning.jl. Lastly, model selection and classification tasks were carried out using the Python scikit-learn package [13].

We explored different values of k and M to find the right balance between preserving local and global structures. Analyzing homogeneity scores and heatmaps, we identified parameter combinations of k and M that yielded reliable and competitive projections for each task. We aimed to ensure that the data representations used for training and classification showed clear and meaningful clusters, increasing the likelihood of achieving competitive results in the training phase.

Table 1 shows each task's best values of k and M. We determined these combinations by evaluating projections using homogeneity scores and affinity propagation.

Task	k	M
Gender	30	7
Ideol. Binary	50	3
Ideol. Multiclass	50	5

Table 1: Optimal combinations of k ($k$nn graph) and M (minimum occurrence) parameters for each task in the challenge.

4.1. Prediction results

For the final selection, we chose the parameter combination for each task that achieved the highest macro F1 score during cross-validation, as described in §3. The best hyperparameters used for each task are summarized in Table 2.

Task	Classifier	Dim.	Hyperparameters
Gender	Linear SVM	3-D	no standardized, C=1000, class_weight=None, dual=False, max_iter=12000, penalty=L2, random_state=42
Ideol. Binary	Linear SVM	9-D	no standardized, C=1, class_weight=balanced, dual=False, max_iter=12000, penalty=L2, random_state=42
Ideol. Multiclass	RBF SVM	9-D	no standardized, C=100, class_weight=balanced, gamma=scale, kernel=rbf

Table 2: Best hyperparameters for the different user profiling model tasks.

We used the top-performing Machine Learning models from our previous tests to analyze the demographic and psychographic characteristics of Tweets in the Test dataset. We specifically aimed to identify self-reported gender as demographic trait, as well as political ideology as a psychographic trait.

Our approach attained an overall score of 0.800788 in the final evaluation, placing us in the 2nd position on the leaderboard. The individual task results are presented in Table 3.

Task	**Macro F1 Score (rank)**
Gender	0.824287 (1)
Ideology Binary	0.860230 (5)
Ideology Multiclass	0.717849 (2)

Table 3: Performance for the most suitable models on the final test dataset.

The macro-average score represents the overall accuracy of our models in predicting the assigned gender and political ideology using the Test dataset.

5. Conclusions

Our framework combines our entropy-based weighting scheme and non-linear dimensional reduction techniques to achieve a practical tradeoff between the model's introspection and high-quality predictions. By experimenting with various combinations of parameters for k and M, we gained valuable insights into task clustering and separation. The results emphasized the significance of considering the neighborhood size and the minimum number of documents required for identifying and distinguishing gender and ideology groups.

Future studies can explore different parameter configurations to assess their effects and a comprehensive error analysis can be conducted to gain valuable insights into the limitations of the models and identify areas for improvement. By identifying specific challenges or patterns in misclassifications, refined models can be developed, and targeted strategies can be implemented to address these issues.

It's also worth exploring other advanced dimensionality reduction techniques to enhance classification performance. Finally, our current approach is limited to lexical features; other representations (e.g., transformer-based ones) can help improve our approach in distinct scenarios.

Acknowledgments

H. Cabrera-Pineda is grateful to CONAHCyT for "Becas Nacionales para Estudios de Posgrado" scholarship No. 899981.

References

[1] C. I. Eke, A. A. Norman, L. Shuib, H. F. Nweke, A survey of user profiling: State-of-the-art, challenges, and solutions, IEEE Access 7 (2019) 144907–144924. doi:10.1109/ACCESS.2019.2944243.

[2] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, B. Stein, Overview of the pan/clef 2015 evaluation lab, in: Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283, CLEF'15, Springer-Verlag, Berlin, Heidelberg, 2015, p. 518–538. URL: https://doi.org/10.1007/978-3-319-24027-5_49. doi:10.1007/978-3-319-24027-5_49.

[3] F. M. R. Pardo, P. Rosso, M. M. y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter, in: CLEF, 2018.

[4] P. Rosso, F. Rangel, Author profiling tracks at fire, SN Computer Science 1 (2020) 72. URL: https://doi.org/10.1007/s42979-020-0073-1. doi:10.1007/s42979-020-0073-1.

[5] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.

[6] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[7] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September, 2018.

[8] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022).

[9] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2023 at IberLEF: Political Ideology Detection in Spanish Texts, Procesamiento del Lenguaje Natural 71 (2023).

[10] H. Cabrera, E. S. Téllez, S. Miranda, Infotec-labd at politices 2022: Low-dimensional stacking model for political ideology profiling, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain, 2022.

[11] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL: https://arxiv.org/abs/1802.03426. doi:10.48550/ARXIV.1802.03426.

[12] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems 14 (2001).

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[14] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] B. J. Frey, D. Dueck, Clustering by passing messages between data points, science 315 (2007) 972–976.

[16] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL), 2007, pp. 410–420.