Alejandro Mosquera at PoliticEs 2022: Towards Robust Spanish Author Profiling and Lessons Learned from Adversarial Attacks

Alejandro Mosquera[1]

[1]*Broadcom Corporation, 1320 Ridder Park Drive San Jose, 95131 California, USA*

Abstract

Social media publications can inadvertently reveal a broad amount of potentially sensitive information such as gender, age, ethnicity or political ideology even when are not intentionally disclosed by their authors. For this reason, social media users concerned about the fact that Natural Language Processing techniques can infer some of these traits with a relatively high degree of accuracy, may actively attempt to conceal the information that they do not want to share for profiling purposes. This paper both describes the user profiling system submitted to the IberLEF 2022 task PoliticEs: "Spanish Author Profiling for Political Ideology" and evaluates its robustness by simulating adversarial perturbations. The aforementioned system was ranked 3rd overall in the shared task with a 88.9% F1, just 1.3 percentage points lower than the highest scoring team. Empirical results suggest that some of the avoidance techniques explored in this research are also likely to successfully evade similar NLP-based author profiling approaches.

Keywords

Adversarial ML, NLP, Privacy-preserving, Author Profiling, Spanish

1. Introduction

Natural Language Processing (NLP) techniques have been successfully applied to many author identification tasks using publicly available social network data [1], not only restricted to attribution [2] but also focused on specific traits such as gender and age [3], ethnicity [4], political ideology [5], psychometric [6] and socio-linguistic attributes [7] among others.

While there are potentially legit uses of these inferred author traits such as forensic analysis [8], marketing research and targeted advertising [9], non-consensual and indiscriminate social media profiling can impact privacy [10] and anonymity [11] by revealing sensitive information. In order to address these concerns, there is active research looking for effective ways of disrupting author profiling by purposely changing the textual content and writing style prior publication.

The contribution of this paper is twofold: First, the Spanish author profiling system submitted to the PoliticEs shared task [12] at IberLef 2022 is presented in Section 3. Second, avoidance techniques for disrupting NLP-based social media profiling approaches are reviewed in Section 4. Finally, in Section 5 the author draws the main conclusions and outlines future work.

IberLEF 2022, September 2022, A Coruña, Spain

alejandro.mosquera@broadcom.com (A. Mosquera)

0000-0002-6020-3569 (A. Mosquera)

2. Related Work

The core features of social media author profiling methods relying on NLP techniques are of lexical nature such as words and characters. The use of stylistic information such as readability, ratio of out-of-vocabulary words, part of speech or emoji usage and textual sentiment is also common and consistent across different languages [13] [14] [15].

Most of the previous work attempting to thwart the aforementioned methods are based on content rewriting. From these, the most popular approaches are related to machine translation, either via adversarial retraining [16] or multilingual back-translation [17]. Other research treats this as a lexical substitution task [18] [19] by identifying words that leak sensitive information and finding suitable replacements ranked using several metrics.

Despite the increasing interest in multilingual author profiling, to the best of the authors' knowledge no previous work has tackled the application of adversarial attacks and defenses in this area in other languages other than English. For this reason, this paper also evaluates the robustness of the Spanish author profiling system described in this research by measuring its performance under different obfuscation strategies.

3. System Description

The author profiling dataset [5] provided by the PoliticEs task organizers contained Spanish tweets from 314 authors, either journalists or politicians, with an associated user identifier and annotations for each of the proposed sub-tasks: gender, profession, ideology binary and ideology multi-class. As the main outcome from an initial data analysis step the author decided to address the task at author level and to treat the first 3 sub-tasks as binary classification problems, leaving the last as a multi-class classification challenge rather than regression, which had substantially lower cross-validation score. After testing different models and architectures L2-regularized logistic regression was identified as the best performing classifier for all the sub-tasks when using lexical and stylistic features calculated over the concatenation of the tweets of each user, which are detailed below:

- `Word n-grams:` 1-4 n-gram frequencies extracted in lowercase with a Twitter-aware tokenizer [1], keeping only the top 5000 words, using inverse document frequency (IDF) weighting and logarithmically scaled frequency;
- `Character n-grams:` TF-IDF over character unigrams in lowercase;
- `Readability:` Kincaid, ARI, Coleman-Liau, Flesch reading ease, Gunning-Fog, LIX, SMOG, RIX and Dale-Chall indexes in order to characterize the writing style [20] and textual complexity [21]. It is worth mentioning that some of these indexes were designed for the English language [2], but are still used here due the lack of robust off-the-shelf Spanish-specific equivalent implementations;

A list of the top 10 best features using minimum redundancy and maximum relevance [22] as selection criteria for each sub-task can be found in Table 1.

[1] https://www.nltk.org/_modules/nltk/tokenize/casual.html
[2] https://pypi.org/project/readability

Table 1
Best 10 features (minimal-optimal set) for each sub-task

Gender	Profession	Ideology	Ideology-multiclass
las mujeres	nuestra	los españoles	los españoles
conocido	[political_party] . .	derecho	le
factura	de	de nuestro país	natural
diga	entrevista	de sánchez	la justicia
orgullosa	é	sánchez	rose (emoji)
happy face (emoji)	nuestro	la derecha	progre
vive	RIX (readability)	españoles	sánchez
orgulloso de	_	ciudadanía	alquiler
female sign (emoji)	apoyo	la ultraderecha	españoles
mujeres	periodistas	los españoles .	explica

3.1. Ideology Normalization

There were cases where the predictions of the binary and multi-class ideology classifiers would contradict each other, e.g. the former predicting left and the latter predicting moderate-right. Since the binary classifier had higher cross-validation F1 score, in case of disagreement the multi-class predictions were overridden using the following logic:

- `right-binary and left-multi:` right-multi;
- `right-binary and moderate-left-multi:` moderate-right-multi;
- `left-binary and right-multi:` left-multi;
- `left-binary and moderate-right-multi:` moderate-left-multi;

3.2. Gender Calibration

Due the class imbalance (30% more male than female authors) in the training dataset, 0.4 was estimated via cross-validation as the best probability threshold for the gender sub-task.

3.3. Ablation Analysis

An ablation study on the training data uncovered some interesting insights: On the one hand, word-based features were the strongest overall and across all sub-tasks since their removal would substantially decrease all the cross-validation scores. On the other hand, character-based and readability features showed a decisive contribution for gender classification but were mostly irrelevant for the other sub-tasks. A table summarizing the F1 macro scores for different feature sub-sets can be found in Table 2.

Table 2
Evaluation of author profiling system with different feature sets (10-fold F1 macro)

Features	Overall	Gender	Profession	Ideology	Ideology-multiclass
All	**0.848**	**0.738**	0.882	**0.938**	**0.836**
Without character n-grams	0.842	0.71	**0.889**	0.937	0.831
Without word n-grams	0.695	0.621	0.816	0.76	0.583
Without readability	0.846	0.723	0.888	0.937	0.835

3.4. PoliticEs results

The submitted system improved the baseline by a large margin and ranked third (out of 19 competing teams) in the global and individual sub-categories with the exception of Profession, where it ranked second. The final F1 score was just 1.3 percentage points lower than highest performing team "'LosCalis'. In Table 3 are listed the top 4 results and the baseline including the macro average F1 scores for all the sub-categories.

Table 3
Top-4 PoliticEs classification final results on the test set and baseline (F1 macro)

Team	Overall	Gender	Profession	Ideology	Ideology-multiclass
LosCalis	**0.902**	**0.902**	**0.944**	**0.961**	0.8
NLP-CIMAT-GTO	0.89	0.784	0.921	0.961	**0.896**
Alejandro Mosquera	0.889	0.826	0.933	0.951	0.845
CIMAT_2021	0.879	0.836	0.895	0.941	0.845
Baseline	0.511	0.576	0.432	0.595	0.44

4. Evading Author Profiling

The field of adversarial NLP followed the success of attacks on image classification models using input perturbations. However, while pixel-level modifications to images can be almost imperceptible for humans, the generation of adversarial examples in NLP requires additional effort to ensure the original meaning of the sentence is preserved. Therefore, many adversarial NLP approaches rely on techniques such as synonym replacement, paraphrasing, encoding attacks [23] or back-translation.

Author profiling systems using textual features are in principle also vulnerable to general NLP adversarial attacks leveraging either character or word manipulations. For this reason, the author has simulated via cross-validation the impact of evasion strategies against models trained on the original data but evaluated on adversarial copies of the evaluation set for each of the 4 considered approaches. The list of evaluated adversarial attacks against the author profiling system submitted to PoliticEs are as follows:

- `Encoding`: The insertion of non-printable unicode characters such as unicode zero-width no break space (0xFEFF) can affect how NLP applications pre-process text without

affecting its visual representation. Word and sentence tokenizers by default treat special characters as normal inputs and will produce a large number of out-of-dictionary tokens that cannot be mapped to the input features. This is language independent and will affect both n-gram and embedding approaches;

- `Synonym`: Replacing words with synonyms is another way of rewriting the source text without changing its meaning. This attack is language dependent and the NLPAug [3] library was used in order to replace Spanish words with WordNet synonyms;
- `Back-translation`: Paraphrases can be generated by translating a text to different language and translating it back to the source language. Neural Machine Translation (NMT) was leveraged using MarianNMT [24] and an Spanish-English pair of pre-trained transformers [4];
- `Counterfactual`: Certain Spanish words and suffixes are associated with a particular gender and can reveal too much information to author profiling systems, e.g. "contenta", "nosotras". For this reason, a Seq2Seq pre-trained model [5] was used in order to attempt to rewrite the source text by adding gender counterfactuals;

Table 4
Evaluation of author profiling system before and after adversarial attacks (10-fold F1 macro)

Attack	Overall	Gender	Profession	Ideology	Ideology-multiclass
No attack	0.848	0.738	0.882	0.937	0.835
Encoding	**0.288**	**0.361**	**0.261**	**0.434**	**0.097**
Synonym	0.844	0.723	0.895	0.957	0.811
Back-translation	0.85	0.739	0.888	0.936	0.836
Counterfactual	n/a	0.515	n/a	n/a	n/a

In Table 4 the F1 scores of the reference author profiling system are listed before and after the application of adversarial modifications to the input text. An analysis of these results shows that encoding attacks are highly effective, disrupting the initial process of converting text into a representation suitable for machine learning, thus affecting NLP applications that use off-the-shelf tokenizers. Considering that most social media platforms support unicode this can be seen as a low cost first line of defense against indiscriminate profiling that does not alter the meaning or the visual representation of publications.

Contradicting the literature, the use of approaches such as back-translation and synonym replacement had an unintended effect and improved the profiling accuracy in some sub-tasks rather than lowering it. However, considering the relatively small number of training samples this does not look fully conclusive.

The use of counterfactuals substantially lowered the gender prediction accuracy as expected, but further work would be required in order to better preserve the original meaning of the text, since the sequence to sequence model used for this purpose had several shortcomings and produced not very high quality transformations in many cases.

[3] https://nlpaug.readthedocs.io/en/latest/augmenter/word/synonym.html
[4] https://huggingface.co/Helsinki-NLP/opus-mt-es-en
[5] https://huggingface.co/monsoon-nlp/es-seq2seq-gender-encoder

5. Conclusion

This paper presents a strong Spanish author profiling baseline and raises questions about how to improve its robustness on an adversarial environment. In order to do this, a preliminary set of evasion techniques were evaluated against the author profiling system using lexical and stylistic features submitted to PoliticEs at Iberlef 2022. This analysis returned mixed results: On the one hand, while this research shows that is feasible to evade NLP systems by relying on unicode obfuscations we should also expect these to be easily bypassed by robust tokenizers and text normalization filters. On the other hand, traditionally successful evasion approaches such as back-translations or synonym replacement did not seem to work in this particular experiment, which will require a deeper analysis in a future work.

References

[1] 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.

[2] P. Juola, Authorship attribution, Foundations and Trends® in Information Retrieval 1 (2008) 233–334. URL: http://dx.doi.org/10.1561/1500000005. doi:10.1561/1500000005.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Commun. ACM 52 (2009) 119–123. URL: https://doi.org/10.1145/1461928.1461959. doi:10.1145/1461928.1461959.

[4] P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, Author profiling for abuse detection, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1088–1098. URL: https://aclanthology.org/C18-1093.

[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] K. Luyckx, W. Daelemans, Personae: a corpus for author and personality prediction from text, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/759_paper.pdf.

[7] J. Eisenstein, N. A. Smith, E. P. Xing, Discovering sociolinguistic associations with structured sparsity, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1365–1374. URL: https://aclanthology.org/P11-1137.

[8] W. Daelemans, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Tschuggnall, M. Wiegmann, E. Zangerle, Overview of pan 2019: Bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 402–416.

[9] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, M. Bühner, Personality research and assessment in the era of machine learning, European Journal of Personality 34 (2020) 613–631. URL: https://doi.org/10.1002/per.2257. doi:10.1002/per.2257. arXiv:https://doi.org/10.1002/per.2257.

[10] P. Juola, Authorship studies and the dark side of social media analytics, Journal of Universal Computer Science 26 (2020) 156–170. doi:10.3897/jucs.2020.009.

[11] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, D. Song, On the feasibility of internet-scale author identification, in: 2012 IEEE Symposium on Security and Privacy, 2012, pp. 300–314. doi:10.1109/SP.2012.46.

[12] 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).

[13] M. Fatima, K. Hasan, S. Anwar, R. M. A. Nawab, Multilingual author profiling on facebook, Inf. Process. Manage. 53 (2017) 886–904. URL: https://doi.org/10.1016/j.ipm.2017.03.005. doi:10.1016/j.ipm.2017.03.005.

[14] M. Koppel, Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17 (2002) 401–412. doi:10.1093/llc/17.4.401.

[15] P. Piot, P. Martín-Rodilla, J. Parapar, Gender Classification Models and Feature Impact for Social Media Author Profiling, 2022, pp. 265–287. doi:10.1007/978-3-030-96648-5_12.

[16] R. Shetty, B. Schiele, M. Fritz, A4nt: Author attribute anonymity by adversarial training of neural machine translation, in: USENIX Security Symposium, 2018.

[17] D. Adelani, M. Zhang, X. Shen, A. Davody, T. Kleinbauer, D. Klakow, Preventing author profiling through zero-shot multilingual back-translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8687–8695. URL: https://aclanthology.org/2021.emnlp-main.684. doi:10.18653/v1/2021.emnlp-main.684.

[18] S. Reddy, K. Knight, Obfuscating gender in social media writing, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 17–26. URL: https://aclanthology.org/W16-5603. doi:10.18653/v1/W16-5603.

[19] C. Emmery, Á. Kádár, G. Chrupała, Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2388–2402. URL: https://aclanthology.org/2021.eacl-main.203. doi:10.18653/v1/2021.eacl-main.203.

[20] A. Mosquera, P. Moreda, The use of metrics for measuring informality levels in web 2.0 texts, in: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, STIL 2011, Cuiaba, Brazil, October 24-16, 2011, Brazilian Special Interest Group on Natural Language Processing, 2011. URL: https://aclanthology.org/W11-4523/.

[21] A. Mosquera, Alejandro Mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 554–559. URL: https://aclanthology.org/2021.semeval-1.68.

doi:10.18653/v1/2021.semeval-1.68.
[22] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, volume 3, 2003, pp. 523– 528. doi:10.1109/CSB.2003.1227396.
[23] N. Boucher, I. Shumailov, R. Anderson, N. Papernot, Bad Characters: Imperceptible NLP Attacks, in: 43rd IEEE Symposium on Security and Privacy, IEEE, 2022.
[24] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, A. Birch, Marian: Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 116–121. URL: http://www.aclweb.org/anthology/P18-4020.