Extraction of Stylometric Information from Spanish Documents.

César Espin-Riofrio[1]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador

Abstract
The writing style of individuals is the basis for tasks such as authorship attribution, authorship verification or authorship profile, associated with stylometric analysis. Traditional learning methods based on neural networks use the information encoded in the last encoding layer of a model such as Transformers. In this paper, we describe our thesis project in which we propose to investigate whether a deep neural network encodes the style in any way. To do so, we explore the intermediate layers and embeddings of the initial token encoding of all layers of BERT-based Transformer models, to identify and extract style features to improve stylistic modeling systems, with emphasis on the analysis of documents written in Spanish.

Keywords
Style, Stylometry, Natural Language Processing, Transformers

1. Introduction

Style is defined as a form of expression or way of writing, starting with the choice of words, the combination of various words, punctuation, sentence structure, grammatical patterns and all the elements that an author likes to use [1]. The analysis of authorial style, called stylometry, is based on the assumption that style is quantifiable in order to evaluate its distinctive qualities [2].

The tasks associated with stylometry include authorship attribution, authorship verification or authorship profiling, they are based on the analysis of the writing style of individuals. The problem has been extensively explored, resulting in several traditional methods and tools for extracting stylometric features from a text.

Natural Language Processing (NLP) systems were initially based mainly on learning rules from the extraction of style features from a text. Later, they were replaced by machine learning models. Current deep learning models encode the relationship between words and learn about the final embeddings of their encoding layers, with encouraging results in text classification tasks but, we do not know what information about style is contained along the encoding layers. In this sense, we are exploring what style information is collected in the embeddings throughout the coding layers of the Transformer models, in order to experiment in stylometric analysis tasks such as authorship determination applied mainly to the Spanish language.

In this paper, we describe our thesis project focused on the extraction of stylometric features from Spanish language documents. We highlight the importance of our research, review its origin and related works, state the hypothesis and describe our research along with the methods, experiments and specific research elements proposed.

2. Justification of the proposed research

Stylometric-based Natural Language Processing is an approach that uses style analysis techniques to study and characterize texts. Style is reflected in the words and expressions used in texts in aspects such as syntax and grammar and in other measures such as average number of words used, frequency of word usage, length of paragraphs, etc. How a computer system can represent the style of a text or a set of documents is important. Stylometry includes among its most important tasks authorship attribution, authorship verification and authorship profiling, most of them solved on the basis of the writing style of a text. Text classification is a fundamental task of NLP, where the style of a text is the basis for extracting features.

NLP applies machine learning methods to identify patterns, extract and analyze features related to the writing style of a text. Traditional models obtain features using artificial methods and then classify them with classical machine learning algorithms, the effectiveness of these methods is largely limited by feature extraction. In contrast, deep learning integrates feature engineering into model fitting by learning a set of transformations that map features directly to outputs. Since their emergence, deep learning models treat the issue of style almost blindly, the models are applied to learn about features and relationships between words within text without delving into style.

We consider it important to delve deeper into what a neural network learns in relation to style in order to apply it to new models for solving text classification tasks such as authorship detection.

On the other hand, there are about 496 million people in the world who speak Spanish natively, making it the second most spoken language in the world [3]. Therefore, it is very important to carry out studies related to machine learning methods to extract style features from Spanish language documents to solve different NLP tasks.

3. Related work

The beginning of stylometry dates back to Augustus de Morgan's suggestion to resolve authorship disputes by means of word length frequency, in the year 1851 [4]. His hypothesis was investigated by [5], who published the results of his work by measuring the length of several hundred thousand words from the works of Bacon, Marlowe and Shakespeare. George Zipf discovered, using logarithmic scales, that there was a relationship between the rank and frequency of words, later known as Zipf's Law [6]. [7] measured word frequency for vocabulary richness analysis, now known as the "Yule characteristic". [8] used statistical methods to investigate the authorship of the Federalist Papers. The Federalist problem has been used subsequently as stylometry's 'testing ground' for new techniques. In the late 1980s, John Burrows published a

series of seminal articles in which he regenerated stylometry as a viable tool in authorship attribution [9, 10, 11]. The initial work involving neural networks with stylometry was presented in [12]. [13] achieved results consistent with those of Mosteller and Wallace described earlier, using just eleven of their thirty 'marker' words as input to a neural network.

The stylistic features of a text are present at various levels such as in the vocabulary, the syntax, the grammar, the semantics, and in some cases in the layout, presentation, etc. [14] carried out an exploration of 166 features used for authorship attribution including commonly used stylistic features and several others intended to capture emotional tone. [15] divided authorship attribution features into five groups, on which much work has been don: lexical [16], character [17], syntactic [18], semantic[15] and application-specific features [19].

Simple lexical features, such as word frequencies, word n-grams, function words, word or phrase length, have been widely used since early attribution work [5], functional words were useful features in [8], the usefulness of character n-grams was highlighted in [15, 20]. Bag-of-words (BoW) approaches have also been reported as being useful for authorship attribution [21]. Term Frequency-Inverse Document Frequency (TF-IDF) [22] uses the word frequency and inverses the document frequency to model the text.

Traditional methods are statistics-based models, such as Naïve Bayes (NB) [23], K-Nearest Neighbor (KNN) [24], and Support Vector Machine (SVM) [25]. In a PAN 2013 Competition [26], all participants used a machine learning algorithm for classification, including Decision Trees, Support Vector Machines and Random Forests, [27, 28].

The evolution of better computer hardware like GPUs and word embeddings like Word2Vec [29] and Glove [30] increased the use of deep learning models like CNN [31] and RNN [32]. LSTM (Long short-term memory) [33] attempts to solve the short-term memory problem of RNNs by retaining selected information in long-term memory. Convolutional seq2seq [34] applies convolutional neural networks.

Transformers [35], apply self-attention, captures the weight distribution of words in sentences. The attention mechanism is often used in an encoder-decoder architecture, and there are many variants of attention implementations [36]. A Transformer encoder layer is composed of multi-head self-attention following a position-wise feed-forward network (FFN) with the residual connection [37] and layer normalization [38]. Transformer architectures rely on explicit position encodings in order to preserve a notion of word order. A positional embedding should be considered together with the NLP tasks [39]. The absolute position embedding is used to model how a token at one position attends to another token at a different position [40]

Pre-trained language models [41], became a trend among many NLP tasks. Pre-trained language models effectively learn global semantic representation and significantly boost NLP tasks, including text classification. It generally uses unsupervised methods to mine semantic knowledge automatically and then constructs pre-training targets so that machines can learn to understand semantics [42].

Transformed-based pre-trained language models (T-PTLM) learn universal language representations from large volumes of text data using self-supervised learning and transfer this knowledge to downstream tasks. These models provide good background knowledge to downstream tasks which avoids training of downstream models from scratch [43]. GPT [44] and BERT [45] are the first Transformer-based pretrained language models developed based on transformer decoder and encoder layers respectively.

In general, an encoder-based T-PTLM consists of an embedding layer followed by a stack of encoder layers. For example, the BERT-base model consists of 12 encoder layers while the BERT-large model consists of 24 encoder layers. The output from the last encoder layer is treated as the final contextual representation of the input sequence. In general, encoder-based models like BERT are used in Natural Language Understanding (NLU) tasks.

Transformer-based models can parallelize computation without considering the sequential information suitable for large scale datasets, making it popular for NLP tasks. Thus, some other works are used for text classification tasks and get excellent performance, such as RoBERTa [46], XLNet [47], Bart [48], deBERTa [49], ERNIE [50].

In NLP tasks related to stylometry, some lexicons have been employed as EuroWordNet [51], Spanish Emotion Lexicon (SEL) [52], [53] perform a lexicon-based sentiment analysis of short texts generated on the social network Twitter in Spanish, [54] a lexicon-based approach to extract sentiment from text, Bing Liu English Lexicon or polarity classification [55], Spanish Opinion Lexicon (SOL) [56].

About corpus, there are some publicly available corpora for Stylometrics, they are important for NLP-related research, Autextification dataset [57], Enron[58], IMDB1M reviews [59], Guardian10 corpus [60].There are several important corpora for specific tasks in the Spanish language, such as Spanish-language corpus for researching offensive language OffendES [61], SFU Spanish review corpus [62], PoliCorpus 2020 [63], eSOLHotel. For the shared task on Multi-Author Writing Style Analysis PAN@CLEF2023 [64], PAN22 Style Change Detection [65], PAN21 Profiling Hate Speech Spreaders on Twitter [66], PAN20 Profiling Fake News Spreaders on Twitter [67].

Regarding the main tasks related to stylometry, there are the shared evaluation campaigns of Natural Language Processing (NLP) systems in Spanish and other languages, such as Automatically generated texts identification: Human or Generated, Model Attribution in AuTexTification in IberLEF 2023. Spanish Author Profiling for Political Ideology in IberLEF 2022. PAN CLEF shared tasks; Multi Author Writing Style Analysis PAN23, Style Change Detection PAN22, Profiling Hate Speech Spreaders on Twitter PAN21, Profiling Fake News Spreaders in Twitter PAN20, Celebrity Profiling PAN20, Bots and Gender Profiling PAN19, among others.

4. Research proposal, hypothesis

Neural networks are capable of capturing stylistic information, that information combined with previously known stylistic features such as character-level characteristics, word, phrase, vocabulary richness, lexical complexity, etc., can help solve tasks such as authorship attribution, profiling users based on their writing, differentiating between synthetic text and human-written text, etc.

The question arises; what does a neural network learn that is related to style?

We propose to further investigate the topic and determine what information about style is contained throughout the layers of pre-trained Transformer-based models, and experiment with methods of extracting their embeddings to refine learning models in text classification tasks, especially in Spanish.

5. Methodology and proposed experiments

An exhaustive analysis of the state of the art has been carried out, determining the classical techniques for the extraction of style features from Spanish and English texts, exploring with them in different application domains and tasks.

We participated in the main international forums on NLP tasks such as PAN, IberLEF, Semeval. We use in our experiments the reference datasets proposed in those campaigns, and thus compare our results with those obtained by other researchers.

We are experimenting with current neural network models to determine what they learn about style. To this end, we are currently exploring with the extraction of initial embeddings from all layers of BERT-based Transformers models to fine-tune a learning model for various text classification tasks.

In terms of dissemination of our results, we have published several scientific papers of worldwide impact, and we are also participating in international scientific conferences such as SePLN, Laccei and SmartTech.

6. Specific research elements proposed

We explore the capacity of linguistic features of various kinds that can be extracted from the text to be considered elements of style, such as lexical diversity, lexical complexity, syntactic and semantic features, etc.

We ask whether there are style features in the parameters that a neural network learns, whether style is encoded in any way in a deep neural network such as Transformer-based models, and if so, where and how. In the deeper layers of a Tranformer encoder it is possible that there is information about the style rather than the semantics. In this sense, we are analyzing, based on a series of works, not only the final coding of BERT-based Transformer models, but also the first and intermediate layers of coding in search of style features, so we are exploring ways to analyze and extract that information to improve stylistic modeling systems.

Acknowledgments

References

[1] S. Pinker, The sense of style: The thinking person's guide to writing in the 21st century, Penguin Books, 2015.

[2] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard, Surveying stylometry techniques and applications, ACM Computing Surveys (CSuR) 50 (2017) 1–36.

[3] CVC. Anuario 2022. Informe 2022. El español: una lengua viva. El español en cifras., ???? URL: https://cvc.cervantes.es/lengua/anuario/anuario_22/informes_ic/p01.htm.

[4] S. E. De Morgan, A. De Morgan, Memoir of Augustus De Morgan, Longmans, Green, and Company, 1882.

[5] T. C. Mendenhall, The characteristic curves of composition, Science (1887) 237–246.

[6] G. K. Zipf, Selected studies of the principle of relative frequency in language.(1932) (1932).

[7] G. U. Yule, The statistical study of literary vocabulary, in: Mathematical Proceedings of the Cambridge Philosophical Society, volume 42, ????, pp. b1–b2.

[8] F. Mosteller, D. L. Wallace, Inference and disputed authorship, The Federalist (1964).

[9] J. F. Burrows, Word-patterns and story-shapes: The statistical analysis of narrative style, Literary & Linguistic Computing 2 (1987) 61–70.

[10] J. F. Burrows, 'an ocean where each kind...': Statistical analysis and some major determinants of literary style, Computers and the Humanities 23 (1989) 309–321.

[11] J. F. Burrows, Not unles you ask nicely: The interpretative nexus between analysis and information, Literary and Linguistic Computing 7 (1992) 91–109.

[12] R. A. Matthews, T. V. Merriam, Neural computation in stylometry i: An application to the works of shakespeare and fletcher, Literary and Linguistic computing 8 (1993) 203–209.

[13] F. J. Tweedie, S. Singh, D. I. Holmes, Neural network applications in stylometry: The federalist papers, Computers and the Humanities 30 (1996) 1–10.

[14] D. Guthrie, Unsupervised detection of anomalous text, Ph.D. thesis, Citeseer, 2008.

[15] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for information Science and Technology 60 (2009) 538–556.

[16] J. Houvardas, E. Stamatatos, N-gram feature selection for authorship identification, in: Artificial Intelligence: Methodology, Systems, and Applications: 12th International Conference, AIMSA 2006, Varna, Bulgaria, September 12-15, 2006. Proceedings 12, Springer, 2006, pp. 77–86.

[17] F. P. D. S. V. Keselj, S. Wang, Language independent authorship attribution using character level language models (????).

[18] F. Leuzzi, S. Ferilli, F. Rotella, A relational unsupervised approach to author identification, in: New Frontiers in Mining Complex Patterns: Second International Workshop, NFMCP 2013, Held in Conjunction with ECML-PKDD 2013, Prague, Czech Republic, September 27, 2013, Revised Selected Papers 2, Springer, 2014, pp. 214–228.

[19] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: Writing-style features and classification techniques, Journal of the American society for information science and technology 57 (2006) 378–393.

[20] R. Schwartz, O. Tsur, A. Rappoport, M. Koppel, Authorship attribution of micro-messages, in: Proceedings of the 2013 Conference on empirical methods in natural language processing, 2013, pp. 1880–1891.

[21] M. Koppel, J. Schler, S. Argamon, Authorship attribution in the wild, Language Resources and Evaluation 45 (2011) 83–94.

[22] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern information retrieval, volume 463, ACM press New York, 1999.

[23] M. E. Maron, Automatic indexing: an experimental inquiry, Journal of the ACM (JACM) 8 (1961) 404–417.

[24] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE transactions on information theory 13 (1967) 21–27.

[25] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European conference on machine learning, Springer, 1998, pp. 137–142.

[26] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, Overview of the author profiling task at pan 2013, in: CLEF conference on multilingual and multimodal information access evaluation, CELCT, 2013, pp. 352–365.

[27] T. M. Mitchell, Artificial neural networks, Machine learning 45 (1997) 127.

[28] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[29] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[30] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[31] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, arXiv preprint arXiv:1404.2188 (2014).

[32] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, arXiv preprint arXiv:1605.05101 (2016).

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[34] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: International conference on machine learning, PMLR, 2017, pp. 1243–1252.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[36] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[38] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[39] Y.-A. Wang, Y.-N. Chen, What do position embeddings learn? an empirical study of pre-trained language model positional encoding, arXiv preprint arXiv:2010.04903 (2020).

[40] Z. Huang, D. Liang, P. Xu, B. Xiang, Improve transformer models with better relative position embeddings, arXiv preprint arXiv:2009.13658 (2020).

[41] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences 63 (2020) 1872–1897.

[42] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, L. He, A survey on text classification: From traditional to deep learning, ACM Transactions on Intelligent Systems and Technology (TIST) 13 (2022) 1–41.

[43] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, Ammus: A survey of transformer-based pretrained models in natural language processing, arXiv preprint arXiv:2108.05542 (2021).

[44] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language under-

standing by generative pre-training (2018).

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[47] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[48] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[49] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[50] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, H. Wu, Ernie: Enhanced representation through knowledge integration, arXiv preprint arXiv:1904.09223 (2019).

[51] P. Vossen, A multilingual database with lexical semantic networks, Dordrecht: Kluwer Academic Publishers. doi 10 (1998) 978–94.

[52] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Trevino, J. Gordon, Empirical study of machine learning based approach for opinion mining in tweets, in: Advances in Artificial Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part I 11, Springer, 2013, pp. 1–14.

[53] A. Moreno-Ortiz, C. P. Hernández, Lexicon-based sentiment analysis of twitter messages in spanish, Procesamiento del lenguaje natural 50 (2013) 93–100.

[54] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.

[55] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[56] M. D. Molina-González, E. Martínez-Cámara, M.-T. Martín-Valdivia, J. M. Perea-Ortega, Semantic orientation for polarity classification in spanish reviews, Expert Systems with Applications 40 (2013) 7250–7257.

[57] A. Sarvazyan, J. Ángel González, M. Franco, F. M. Rangel, M. A. Chulvi, P. Rosso, Autextification dataset (full data), 2023. URL: https://doi.org/10.5281/zenodo.7956207. doi:10.5281/zenodo.7956207.

[58] B. Klimt, Y. Yang, The enron corpus: A new dataset for email classification research, in: Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15, Springer, 2004, pp. 217–226.

[59] Y. Seroussi, F. Bohnert, I. Zukerman, Personalised rating prediction for new users using latent factor models, in: Proceedings of the 22nd ACM conference on Hypertext and hypermedia, 2011, pp. 47–56.

[60] E. Stamatatos, On the robustness of authorship attribution based on character n-gram features, JL & Pol'y 21 (2012) 421.

[61] F. M. Plaza-del Arco, A. Montejo-Ráez, L. A. Urena-López, M. Martín-Valdivia, Offendes: A new corpus in spanish for offensive language research, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 1096–1108.

[62] M. Taboada, SFU Review Corpus | Maite Taboada, 2017. URL: https://www.sfu.ca/$\sim$mtaboada/SFU_Review_Corpus.html.

[63] 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.

[64] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Pan23 multi-author writing style analysis, 2023. URL: https://doi.org/10.5281/zenodo.7729178. doi:10.5281/zenodo.7729178.

[65] E. Zangerle, M. Mayerl, M. Tschuggnall, M. Potthast, B. Stein, Pan22 authorship analysis: Style change detection, 2022. URL: https://doi.org/10.5281/zenodo.6334245. doi:10.5281/zenodo.6334245.

[66] F. RANGEL, B. CHULVI, G. L. D. L. PEÑA, E. FERSINI, P. ROSSO, Profiling hate speech spreaders on twitter, 2021. URL: https://doi.org/10.5281/zenodo.4603578. doi:10.5281/zenodo.4603578.

[67] F. RANGEL, P. ROSSO, B. GHANEM, A. GIACHANOU, Profiling fake news spreaders on twitter, 2020. URL: https://doi.org/10.5281/zenodo.4039435. doi:10.5281/zenodo.4039435.