Profiling Poets from Spanish Sonnets

Alysa Meng
University of Washington
menga@uw.edu

Abstract

Every poem contains traces of its invisible author. This project investigates how supervised machine learning techniques (SVM, Bi-LSTM, BERT) can capture authorial metadata from Spanish sonnets. Given characteristics from a poem, we explore how computational models produce information about authorship, including time period, gender and region through a series of text classification tasks. We apply quantitative and qualitative methods for model evaluation to provide a holistic view of the dataset without sacrificing the specific contexts that each poem carries. While each model architecture captures poetic aspects in a different way, they all struggle to perform well on less frequent classes in an uneven dataset. By examining differences in poetic representations and the learned weights, we discover ways that models abstract and map linguistic features to a predicted class. More broadly, this work contributes to the discussion on the limits of artificial intelligence in Hispanic literature and the humanities.

1. Introduction

The question of authorship has long been a subject of scholarly investigation across literature, history, and philology. Authorship attribution concerns itself with identifying the most likely author of a given text, particularly in cases where authorship was disputed or unknown. Over the past several decades, this area of inquiry has been significantly advanced by computational methods that analyze stylistic features of texts to draw inferences about their origins and gained popularity in the latter half of the 20th century, coinciding with the advent of computational linguistics and the increasing availability of digital corpora. Early stylometric approaches focused on quantifying linguistic style through features such as word frequency distributions, syntactic patterns, and semantic tendencies with specialized statistical measures that can be compared across a corpus to support claims of common or divergent authorship. Stylometric methods have found extensive application in textual forensics and historical linguistics research, allowing scholars to infer not only the likely author of a text but also aspects of the author's sociocultural, temporal, and educational background [7]. This broader conception of authorship background inference concerned with extracting latent metadata about an author from stylistic features has been valuable in contexts where direct biographical or contextual information is limited or inaccessible.

The integration of machine learning into stylometric practice has reshaped practices in authorship attribution. Traditional feature engineering techniques are often supplemented by data-driven approaches that leverage large-scale text representations and classification models. This transition aligns with the broader emergence of digital humanities, a field that questions how digital tools and computational methodologies can augment humanistic inquiry. With the recent popularity of large language models (LLMs), more recent works have explored the problem of distinguishing human-authored poetry from that produced by LLMs, with authorship inference methods employed to assess the stylistic alignment of generated texts with established poetic traditions or specific groups of authors [10]. This application illustrates a shift focus from identifying known human authors to tracing stylistic influences embedded within model outputs. These developments call for a reevaluation of the tools for authorship inference, specifically in the context of evolving AI-driven systems that influence how literature is interpreted.

This project explores the extent to which machine learning approaches can help organize, interpret, and derive patterns from Spanish poetry.

In particular, we evaluate different classification models for the following three biographical profiling tasks:

1. Predicting the author's country of origin,
2. Predicting the author's gender, and
3. Predicting the historical period in which the poem was composed.

2. Related Work

While support vector machines (SVMs) and principal component analysis (PCA) have dominated feature-based

stylometric approaches, recent scholarship has begun to explore neural network architectures, such as long short-term memory networks (LSTMs), for capturing syntactic and semantic features relevant to stylistic inference [14]. These approaches increasingly focus on integrating linguistic cues into authorship attribution models, with some early investigations employing transformer-based architectures such as BERT [6]. However, much of this work remains concentrated on English-language prose, typically restricted to texts produced within the last few centuries.

Recent advancements have extended authorship attribution to languages and literary forms beyond Western prose, though such efforts remain limited in scope and number. For instance, Hernández [8] applied stylometric analysis to explore stylistic similarities among three poets from Spain's Golden Age, employing the `stylo` package in R [5]. Plecháč et al. [12] conducted a pilot study assessing the role of poetic versification in SVM-based authorship classification, with an emphasis on feature interpretability across Czech, German, Spanish, and English poetry. BERT-based ensemble models have been applied to Arabic poetry [1] and to Japanese literary texts [9], reflecting a growing interest in the cross-linguistic applicability of transformer models for stylistic inference.

The intersection between Hispanic poetry and state-of-the-art computational methods remains underexplored. This project addresses this gap by examining how machine learning models can (or cannot) be used to classify and infer characteristics about authors based on Spanish sonnets.

Poetry in Spanish presents a challenging case for biographical inference. Spanish has grammatical differences from English, such as gender marking and more flexible subject-verb orderings. These aspects potentially limit the effectiveness of language models trained on more uniform and contemporary datasets with methods derived from English tasks. The historical and regional variation in vocabulary, orthography, and syntax add complexity to the classification process, making it an exciting environment for assessing the limits and capabilities of current computational approaches to literary analysis.

3. Methodology

Preliminary data analysis and visualization to first understand the structure of the dataset and its contents was done in Excel. Data processing was done with Python's pandas package. Then we compare 3 different deep learning approaches for classification, implemented using `pytorch` libraries (e.g., `torch.nn.Module`). Training and evaluation for SVM and LSTM models were done locally. The BERT-architecture model was fine-tuned on Google Colab to take advantage of GPU speedup.

4. Dataset

This study uses the Diachronic Spanish Sonnet Corpus (DISCO), a richly annotated dataset comprising 4,530 Spanish sonnets spanning five centuries [13]. Annotations were generated with computational techniques and verified by domain experts. The authorial metadata labels are used for the supervised classification tasks, and the per-poem rhyme and meter information are used as SVM features.

For the 1,237 authors with associated metadata, The composition of DISCO is skewed towards male poets, Spanish authors, and 19th century works, all summarized in the Appendix.

As there may be ambiguity in mapping authors to a particular century, we take the floor of the `normdate` attribute, which already has standardized the birth date format for the authors. The precise breakdown for poems and their sources are further discussed in [13].

The poems in DISCO are specifically *sonetos*, 14-verse poems where each verse is 11 syllables. The poem into 4 stanzas: 2 quatrains and 2 tercets. Each quatrain has 4 verses, and each tercet has 3 verses. The rhyme scheme for the quatrains and tercets are structured. The stressed or unstressed nature of each syllable in a verse is known as its meter. An example is provided in the Appendix.

4.1. Data Pipeline

For each poem, we obtain its text from the `.txt` file and join it with its associated TEI annotations for rhyme and meter in the `.xml` format. Afterwards, data enrichment is done by joining the authorial metadata with the poem content and prosodic features. As we are interested in overall patterns across gender, countries, and periods, we remove poems by the 20th century author and by authors in a country group with less than 10 total authors. We then sanitize the poems and drop those that do not have rhyme or meter information, since there were some inconsistencies in DISCO's TEI annotations. This includes stripping punctuation and extraneous whitespace, and setting all characters to lowercase in the SVM-based and LSTM-based models. The casing and punctuation is maintained in the BERT-based model since a cased transformer is used for fine-tuning. A visualization of this process is provided in Figure 1.

Stratified sampling is done to split the data into train and test sets for each classification task to maintain the ratios of the category labels and produce a 90-10 train-test split across the poems. After data processing, each classification task had roughly 4050 poems in its training set and 450 poems in its test set. Finally, 10% of the training set was kept aside to form a validation set using the stratified splitting technique.

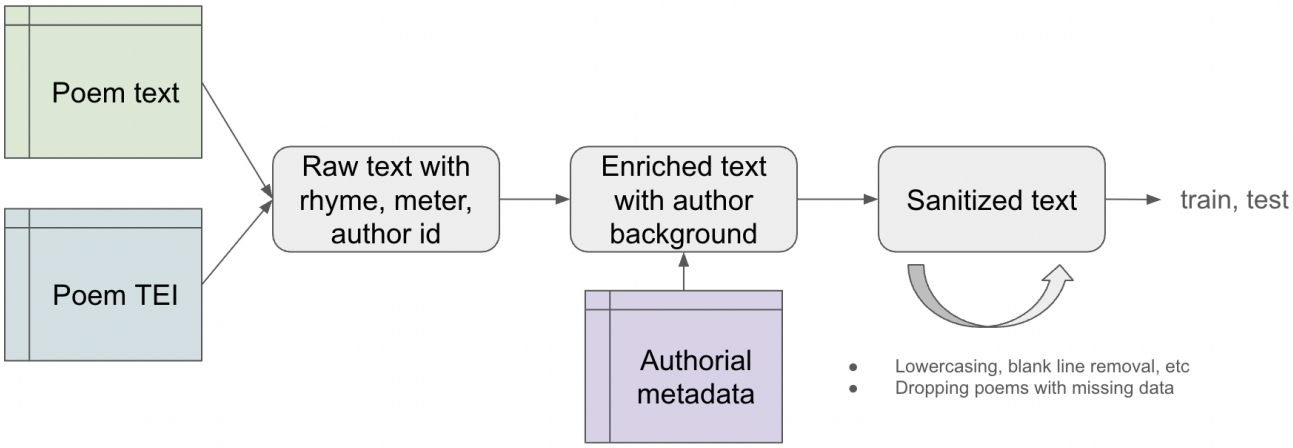

Figure 1. Illustration of data flow for extracting train and test sets

4.2. Support Vector Machine (SVM)

SVMs are a widely used, relatively interpretable machine learning model suitable for high-dimensional text classification tasks so we implement an SVM classifier as a baseline. For features, we use the top 2000 unigrams and top 2000 bigrams using TF-IDF weights, respectively. The TF-IDF weights correspond to the product of the a term frequency and its inverse document frequency.

For a term t and document d in the set of documents D,

$$\mathsf{tf}(t,d) = \frac{\#\text{ times } t \text{ appears in } d}{\sum_{\text{term } t' \in d} \#\text{ times } t' \text{ appears in } d}$$

$$\mathsf{idf}(t,D) = \frac{|D|}{\#\text{ of documents in } D \text{ that contain } t}$$

To incorporate prosodic aspects of poetry, the meter and rhyme annotations from the Text Encoding Initiative (TEI) files are sequentially encoded. The n-gram and prosodic features are concatenated to get a vectorized representation for a poem.

4.3. Long Short-Term Memory (LSTM)

To improve upon the baseline, we evaluate a Bidirectional Long Short-Term Memory (Bi-LSTM) neural network with an attention layer before the classification layer. Bi-LSTMs are more effective at capturing sequential dependencies in text and can incorporate both forward and backward context when processing a sentence. The expectation is that Bi-LSTM models will more effectively capture underlying syntactic and rhythmic patterns that distinguish authors, especially across time and region. We use pre-trained GloVe word embeddings from the Spanish Billion Words Corpus and Embeddings project [2, 11] with a vocabulary size of 10,000 words.

4.4. Transformer

Finally, this study fine-tunes BETO [3], a Spanish-language BERT model pre-trained on a large Spanish corpus, for each classification task. BETO's transformer architecture allows for context-sensitive word embeddings and attention mechanisms that have proven effective in various downstream NLP tasks. This model is adapted to the three classification tasks, with particular attention to fine-tuning strategies suitable for limited, domain-specific training data such as historical poetry.

5. Experiments

Models are evaluated using standard classification metrics, including accuracy, macro F1 score, and confusion matrices. Beyond performance metrics, this study also emphasizes model interpretability. Feature importance analysis is conducted for the SVM classifiers, and attention weight visualization is used to investigate which aspects of the poems most strongly influence the Bi-LSTM and BETO predictions.

The final selected model hyperparameters are reported in each section. See the Appendix for details on hyperparameter tuning.

Model	Gender	Country	Period
MostFreq	0.933	0.613	0.694
SVM	0.883	0.574	0.764
LSTM	0.930	0.605	0.836
BERT	0.928	0.656	0.858

Table 1. Model Accuracies (MostFreq denotes a classifier that always chooses the most frequent label)

Figure 2. Classification model normalized confusion matrices

Model	Gender	Country	Period
SVM	0.597	0.303	0.570
LSTM	0.686	0.278	0.683
BERT	0.582	0.248	0.737

Table 2. Model Macro F1 Scores

5.1. SVM

For all three classification tasks, models consistently predicted the majority class for all elements in the train and test sets, even after hyperparameter selection. This often gave the highest accuracy, but at the cost of always getting the majority class correct and every other class incorrect.

In an effort to mitigate this issue, we implemented a weighted loss function to penalize more for an incorrect prediction for a less frequent class, determined by a normalized inverse proportion. Given a class c in classes C,

$$\text{class-weight}(c) = \frac{\left(\frac{\text{\# samples in train set}}{\text{\# samples in train set that have label } c}\right)}{|C|}$$

With a chosen learning rate 0.005 and weight decay 1e-4, we have the 88.3% accuracy for the gender prediction, 57.4% for birth country prediction, 76.4% accuracy for the century prediction as shown in Table 1. The century classifier was the only model out of the 3 that improved beyond the accuracy evaluated on assigning everything to the majority class (19th century), which was comparatively 69.4%. The accuracy for always predicting the majority for gender (male) was 93.3% and for country (Spain) was 61.3%.

5.2. LSTM

Unlike the SVM-based classifiers, the Bi-LSTM architecture with no class weights did not consistently predicted majority class each time for all three tasks, but there contin-

ued to be a large emphasis on assigning the more frequent labels correctly. The tradeoff between classifying majority classes correctly for higher accuracy versus getting a lower, but balanced performance across classes was sensitive to model hyperparameters. Higher weight decay (e.g., 1e-2) reduced to classifying all examples as the majority class, while lower weight decay led the model to overfit on the training set. Therefore, we also implemented a class weights calculation in the cross entropy loss.

The final version of the LSTM-based models used an embedding dimension of size 300 to align with the GloVe embedding dimensions and a hidden dimension for the Bi-LSTM of size 64 (128 divided by 2 for the bi-directional architecture).

Each of the models had a different best-performing learning rate and weight decay configuration after hyperparameter tuning using the validation set, as shown in Table 3 alongside their respective test accuracies after training for 5 epochs.

Task	Learning rate	Weight decay
Gender	1e-3	1e-4
Country	1e-2	1e-5
Century	1e-2	1e-5

Table 3. LSTM final model hyperparameters

On a high level, the LSTM-based models followed the same pattern as the SVM-based models in the confusion matrix. For all three classification tasks, the LSTM-based models had a higher overall accuracy than the SVM-based models.

5.3. Transformer

Like the other models, the BERT-based model favored the most frequent classes, so class weights were used in the cross entropy loss. Training was done for 5 epochs with a 0.3 dropout rate and a batch size of 32. The results are shown in Table 1. While the accuracy is high for country classification, we observe a low macro F1 score, finding that the model learns to classifying a limited subset of the classes with high accuracy rather than all the classes.

In regards to accuracy, the fine-tuned versions of BETO performs better than the SVM-based models and both the birth country and century models. However, Figure 2 displays lower accuracies in many cases for less frequent classes in the gender and birth country tasks.

6. Discussion

The weighted SVM results demonstrate more balanced classification output between the most common classes and the other classes, but ultimately prefer the majority classes.

Term	Male	Female
valiente (brave)	-	+
señor (sir, lord)	+	-
error (error)	-	+
herido (wounded)	-	+
del amor (of love)	+	-

Table 4. Gender feature importance. Male and female had the same ranking for the top 5 but inverted in sign, ordered most to least significant.

The confusion matrix shows that the model incorrectly predicts male the majority of the time when a poetry was written by a female poet. This is similar for Spain in the birth country classifier and the 19th century in the period classifier.

Taking the magnitude of the learned feature weights, we determine which SVM features are important for gender, country, and century classification. Top features for gender are in Table 4. For birth country, we list each country's most important positive feature to understand what would push towards classification for a particular country (Spanish term, English literal translation): Philippines (misterio, mystery), Cuba (primero, first), Spain (prado, meadow), Portugal (jardín, garden), Mexico (soberbio, proud), Chile (hay un, there's a), Argentina (los que, those that), Venezuela (bronce, bronze), Puerto Rico (y mi, and my), Costa Rica (lluvia, rain), Colombia (cadáver, cadaver). With the century classification, we have the following: 15th and earlier (eterno, eternal), 16th (guerra, war), 17th (voz y, voice and), 18th (de él, of him), and 19th (soneto, sonnet). The n-gram features were generally ranked higher in importance than the prosodic features. This may be due to the rigid structure of a Spanish sonnet that is shared across the corpus of sonnets. Furthermore, the limited vocabulary derived from the training data resulted in inability to fully represent the words in a poem in the test set.

Upon analyzing the LSTM-based model attention layer weights, a similar on focus towards noun phrases can be seen in Figure 4 for a period classification test example with a 16th century true label, which can be compared to the SVM-based 16th century model weights in Figure 3.

Birth country nor century perfectly corresponds to poetic style. For instance, the LSTM-based country classification model misclassified an Argentinian-born poet as Spanish, but the poet had lived in Spain for an extended period of time so it would be reasonable to find lexical traces associated with a dialect from Spain in his poetry. With continuous periods of time being mapped to discrete categories, there are some poets born immediately before the turn of a century that were classified into the later century for all the time period classifiers.

The transformer-based models demonstrate some ability

reina	desotras	flores	fresca	rosa			
-2.05E-01	Removed	-7.53E-01	8.68E-02	1.50E-01			
primero	honor	de	abril	y	de	este	prado
2.68E-01	-1.63E-01	-2.74E-02	9.03E-01	8.58E-01	-2.74E-02	-1.89E-01	1.22E+00
así	te	privilegie	el	cierzo	helado		
9.67E-01	-2.46E-01	Removed	7.72E-02	Removed	1.68E-01		
y	respete	la	helada	rigurosa			
8.58E-01	Removed	-5.74E-01	2.10E-02	Removed			

Figure 3. SVM 16th-century weights for first quatrain of Luis Martín de la Plaza's "Fresca rosa"

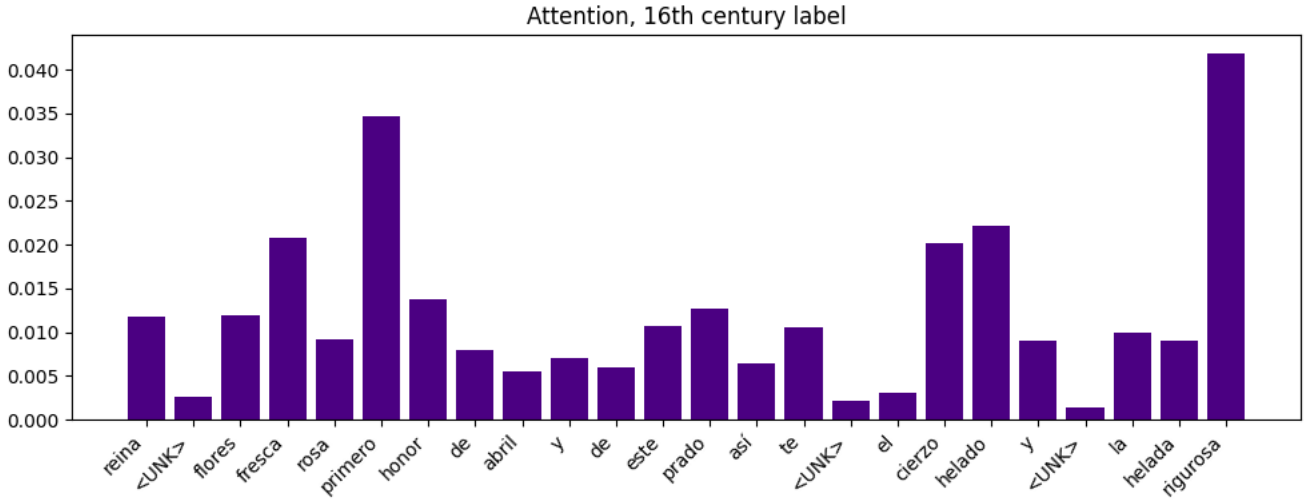

Figure 4. LSTM attention for first quatrain of Luis Martín de la Plaza's "Fresca rosa"

to capture rhythmic features and syntactic structure. Using BertViz [15] to visualize the attention mechanism for the different layers of the model heads, we discover that later layers (8-11) appear to encode higher-level relationships that may correspond to relationships within and between verses. This preliminary observation should be further investigated to better understand the linguistic meaning, if such consistent interpretation can be extracted, from these layers.

7. Limitations

The three approaches investigated do not take into account the full complexity of poetry. Other features to consider include marking more linguistic traits, such as the presence of a synalepha, or tagging named entities, such as places and people. Poetry also has visual cues that standard language modeling approaches do not account for. Punctuation can be critical to the meaning of a poem, even though it was stripped by design for the SVM-based and LSTM-based approaches. All three approaches remove indent and whitespace, but in some cases contribute to the interpreta-

Figure 5. An attention head layer from the first quatrain of Luis Martín de la Plaza's "Fresca rosa"

tion of a poem. Imposing a structured, numerical representation on a poem can unintentionally oversimplify and misrepresent its original meaning and context. As each model produces a different internal representation of a poem, additional work on ensemble models between BERT-based, SVM-based and LSTM-based classifiers. Combining these distinct methods that represent the same poem in different ways can improve overall performance.

It is equally important to acknowledge that the dataset composition is not representative of Hispanic poetry as a whole. The unbalanced nature of the biographical traits posed a challenge for training and evaluating the models. An alternative approach of taking the class with the minimum of examples was considered to sample a subset of data that was balanced across classes, but this would have limited the proportion of the already small dataset that we were working with. A change that we would make given more time in the quarter would be do a primary evaluation across F1 scores rather than accuracies.

In the future, we hope that more diverse datasets are made accessible to train models that optimize across all classes rather than the majority. When handling small, complex and unbalanced humanities datasets, generalized evaluation metrics fail to reveal domain-specific insights about model performance. In each example, we must ask: Was the model incorrect because it didn't capture the key meaningful relationships, or because of an outlier in the corpus? Was the model correct because it was able to abstract patterns from low-level linguistic properties, or because there was a similar, memorized example? Our takeaway is that machine learning models may not always uncover the insightful patterns we anticipate, but nonetheless can underline peculiarities about the input data and direct us towards asking the right questions.

References

[1] Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. BERT-based classical Arabic poetry authorship attribution. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. 2

[2] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. 3

[3] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020. 3

[4] G. de la Vega and J.F. Alcina. *Poesía completa*. Austral : Literatura. Espasa Calpe, 1989. 8

[5] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121, 2016. 2

[6] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. BertAA : BERT fine-tuning for authorship attribution. In Pushpak Bhattacharyya, Dipti Misra Sharma, and Rajeev Sangal, editors, *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India, Dec. 2020. NLP Association of India (NLPAI). 2

[7] D. C. Greetham. Textual forensics. *PMLA/Publications of the Modern Language Association of America*, 111(1):32–51, 1996. 1

[8] Laura Hernández Lorenzo. Digital stylistics applied to golden age spanish poetry is fernando de herrera really a transitional poet between renaissance and baroque? 2024. 2

[9] Taisei Kanda, Mingzhe Jin, and Wataru Zaitsu. Integrated ensemble of bert- and features-based models for authorship attribution in japanese literary works, 2025. 2

[10] Tharindu Kumarage and Huan Liu. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54. IEEE, 2023. 1

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3

[12] Petr Plecháč, Klemens Bobenhausen, and Benjamin Hammerich. Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry. *Studia Metrica et Poetica*, 5(2):29–54, 2018. 2

[13] Pablo Ruiz Fabo, Helena Bermúdez Sabel, and Clara Isabel Martínez Cantón. pruizf/disco: Version 5.0, 2023. 2

[14] Jacques Savoy. *Advanced Models for Stylometric Applications*, pages 153–187. Springer International Publishing, Cham, 2020. 2

[15] Jesse Vig. Visualizing attention in transformer-based language representation models, 2019. 6

Appendix

Hyperparameter tuning

For SVM-based and LSTM-based models, grid search was done for learning rates (1e-2, 1e-3, 1e-4) and weight decay (0, 1e-5, 1e-4, 1e-3); Adam was chosen as the optimizer. For the BERT-based model, we tried various dropout rates (0.2, 0.3, 0.4, 0.5) and batch sizes (8, 16, 32).

DISCO author biographical data distributions

gender	count
Male	1155
Female	81

Table 5. Author gender distribution

country-birth	count
Argentina	29
Bolivia	3
Brasil	2
Chile	13
Colombia	23
Costa Rica	11
Cuba	173
Ecuador	9
España	832
Filipinas	11
Haití	1
Honduras	6
Italia	2
México	35
Nicaragua	2
Panamá	4
Paraguay	2
Perú	9
Portugal	15
Puerto Rico	15
República Dominicana	3
Uruguay	4
Venezuela	10
Unspecified	22

Table 6. Author birth country distribution

century	count
15th or earlier	10
16th	170
17th	309
18th	52
19th	694
20th	1

Table 7. Author period distribution

How to analyze a Spanish sonnet

SONETO XXIII by Garcilaso de la Vega from [4]

En tanto que de rosa y azucena
se muestra la color en vuestro gesto,

y que vuestro mirar ardiente, honesto,
enciende al corazón y lo refrena;

y en tanto que el cabello, que en la vena
del oro se escogió, con vuelo presto,
por el hermoso cuello blanco, enhiesto,
el viento mueve, esparce y desordena;

coged de vuestra alegre primavera
el dulce fruto, antes que el tiempo airado
cubra de nieve la hermosa cumbre.

Marchitará la rosa el viento helado,
todo lo mudará la edad ligera,
por no hacer mudanza en su costumbre.

We observe 14 lines, divided into 2 quatrains and 2 tercets. To analyze the rhyme, we find that the 2 quatrains have consonant rhymes in the pattern "ABBA ABBA" because verses 1, 4, 5, and 8 end in "na" while verses 2, 3, 5, and 6 end in "to". Similarly, the 2 tercets that follow have consonant rhymes in the pattern "CDE DCE". Thus, the rhyme scheme of the poem is "ABBA ABBA CDE DCE". We map each letter to a unique integer for computation.

The meter captures the rhythm of the poem, so to analyze it, we need to analyze which syllables are stressed. Consider the first line, which flows as follows (stressed syllables denoted in capital letters: "en TAN-to que de RO-sa y a-zu-CE-na". While the meter may vary between verse depending on the sonnet, each verse should have a stressed syllable in the 10th position. We can encode this stress pattern with binary variables since each syllable is stressed or unstressed for computation.