Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 215-225 recibido 01-04-2023 revisado 06-05-2023 aceptado 09-05-2023

ALBERTI, a Multilingual Domain Specific Language Model for Poetry Analysis

ALBERTI, *un Modelo de Lenguaje Multilingüe de Dominio Específico para el Análisis de Poesía*

Javier de la Rosa,[1] **Álvaro Pérez Pozo,**[2]
Salvador Ros,[2] **Elena González Blanco**[3]
[1]National Library of Norway, Norway
[2]Universidad Nacional de Educación a Distancia, Spain
[3]IE University, Spain
versae@nb.no

Abstract: The computational analysis of poetry is limited by the scarcity of tools to automatically analyze and scan poems. In a multilingual settings, the problem is exacerbated as scansion and rhyme systems only exist for individual languages, making comparative studies very challenging and time consuming. In this work, we present ALBERTI, the first multilingual pre-trained large language model for poetry. Through domain-specific pre-training (DSP), we further trained multilingual BERT on a corpus of over 12 million verses from 12 languages. We evaluated its performance on two structural poetry tasks: Spanish stanza type classification, and metrical pattern prediction for Spanish, English and German. In both cases, ALBERTI outperforms multilingual BERT and other transformers-based models of similar sizes, and even achieves state-of-the-art results for German when compared to rule-based systems, demonstrating the feasibility and effectiveness of DSP in the poetry domain.
Keywords: Natural Language Processing, Multilingual Language Models, Domain Specific Pre-training, Poetry, Stanzas, Scansion.

Resumen: El análisis computacional de la poesía está limitado por la escasez de herramientas para analizar y escandir automáticamente los poemas. En entornos multilingües, el problema se agrava ya que los sistemas de escansión y rima solo existen para idiomas individuales, lo que hace que los estudios comparativos sean muy difíciles de llevar a cabo y consuman mucho tiempo. En este trabajo, presentamos ALBERTI, el primer modelo de lenguaje multilingüe pre-entrenado para poesía. Usando la técnica de pre-entrenamiento de dominio específico (DSP, de sus siglas en inglés), aumentatmos las capacidades del modelo BERT multilingüe empleando un corpus de más de 12 millones de versos en 12 idiomas. Evaluamos su rendimiento en dos tareas estructurales de poesía: clasificación de tipos de estrofas en español y predicción de patrones métricos para español, inglés y alemán. En ambos casos, ALBERTI supera a BERT multilingüe y a otros modelos basados en transformers de tamaños similares, e incluso logra resultados de estado del arte para el alemán en comparación con los sistemas basados en reglas, lo que demuestra la viabilidad y eficacia del DSP en el dominio de la poesía.
Palabras clave: Procesamiento del Lenguaje Natural, Modelos de Lenguage Multilingües, Pre-entrenamiento de Dominio Específico, Poesía, Estrofas, Escansión.

1 Introduction

Poetry analysis is the process of examining the elements of a poem to understand its meaning. To analyze poetry, readers must examine its words and phrasing from the perspectives of rhythm, sound, images, obvious meaning, and implied meaning. Scansion, a common approach to analyze metrical poetry, is the method or practice of determining and usually graphically representing the

ISSN 1135-5948 DOI 10.26342/2023-71-17

metrical pattern of a line of verse. It breaks down the anatomy of a poem by marking the metrical pattern of a poem by breaking each line of verse up into feet and highlighting the stressed and unstressed syllables (Lennard, 2006).

Having multilingual tools for scansion and analysis of poetic language enables large-scale examinations of poetry traditions, helping researchers identify patterns and trends that may not be apparent through an examination of a single tradition or language (Šeļa, Plecháč, and Lassche, 2022). By using multilingual tools, scholars can compare and contrast different poetic forms, structures, and devices across languages and cultures, allowing them to uncover similarities and differences and gain a more comprehensive understanding of poetic expression.

However, the analysis of multilingual poetry presents significant challenges that must be overcome. It demands a deep understanding of diverse linguistic and cultural traditions, as each language brings its own unique poetic conventions and nuances. Researchers and scholars need expertise in multiple languages to navigate the intricacies of each tradition accurately. Additionally, translation and interpretation pose complex obstacles in multilingual poetry analysis. Figurative language, wordplay, and cultural references deeply rooted in the specific language and culture of the poem make it challenging to convey the intended meaning, emotional impact, and artistic integrity when translating. Cultural contexts, historical references, and subtle language connotations often get lost in translation, making it difficult to fully capture the essence of the original work.

Furthermore, the development of advanced computational tools is crucial for effective analysis and comparison of poetic expression across multiple languages. This requires the application of sophisticated machine learning techniques, natural language processing algorithms, and other emerging technologies. Building models that can accurately capture the unique aesthetic qualities, rhythm, rhyme, and stylistic variations in different languages is an ongoing research endeavor that requires continuous refinement and innovation.

In this work, we investigate whether domain-specific pre-training (DSP) (Gu et al., 2021) in a multilingual poetry setting can be leveraged to mitigate some of these issues. Specifically, we introduce Alberti, a multilingual encoder-only BERT-based language model suited for poetry analysis. We experimentally demonstrate that Alberti exhibits better performance than the base model it was built on, a multilingual BERT (Devlin et al., 2019) which was pre-trained on the 104 languages with the largest Wikipedias. And by reformulating scansion and stanza identification as classification problems, we show that Alberti also outperforms its based model in these downstream tasks. Moreover, we are releasing both Alberti and the dataset used for further training it, which consists of over 12 million verses in multiple languages.

2 Related Work

The transformer architecture (Vaswani et al., 2017) is now pervasive in natural language processing (NLP). In the last five years, context-aware language models have revolutionized the computational modeling of language.

In the humanities, domain specific BERT-based models (Devlin et al., 2019) trained with the goal of predicting masked words are starting to appear. In MacBERTh, (Manjavacas Arevalo and Fonteyn, 2021), the authors present diachronic models for pre-1950 English literature. And a new shared task on historical models for English, French, and Dutch took place last year (Schweter and März, 2020). While pre-training these large language models from scratch is often cost-prohibitive and extremely data demanding, adjusting them to work on other domains and tasks via transfer learning requires less data and fewer resources. For poetry, computational approaches have focused primarily on generation (Lau et al., 2018; Ormazabal et al., 2022) and scansion (Gervás, 2000; Araújo and Mamede, 2002; McAleese, 2007; Ibrahim and Plecháč, 2011; Anttila and Heuser, 2016; Agirrezabal, Alegria, and Hulden, 2017; De Sisto, 2020; De la Rosa et al., 2020), but generally in a monolingual setting. While multilingual systems exist for metrical analysis, they internally work by having different sets of rules for each language (Anttila and Heuser, 2016) or by building ad-hoc neural networks (Agirrezabal, Alegria, and Hulden, 2017). To the best of our knowledge, the only attempt at multilinguality for metrical

pattern prediction was introduced in (De la Rosa et al., 2021) for English, German, and Spanish, where the authors jointly fine-tune different monolingual language models and document some cross-lingual transferability when using multilingual RoBERTa (Liu et al., 2019). Inspired by their good results, in this work we build a domain specific language model trained on a corpus of verses in 12 languages to explore its performance on tasks of poetic nature.

3 Methods and Data

We leverage domain-specific pre-training techniques by fine-tuning the widely used multilingual BERT (mBERT) model with the same base architecture and vocabulary for our specific domain. We adopt the masked language modeling (MLM)[1] objective and further train the model for 40 epochs on a large corpus consisting of 12 million verses, which were sourced from various poetry anthologies. The training was conducted on a Google TPUv3 virtual machine with a batch size of 256, a learning rate of 1.25×10^{-4}, and a weight decay of 0.01. The maximum sequence length was set to 32 since verses with up to 32 tokens using the mBERT tokenizer make up for almost 99 percent of the total. Furthermore, we used a 10,000-step warmup process, which allowed the model to learn the distribution of the corpus gradually. We are naming the resulting model ALBERTI[2]. After training, we evaluate the model on 10% of the corpus held out as a validation set, achieving a final global MLM accuracy of 57.77%.

3.1 PULPO

The training of the model was done over a new corpus we built for the occasion. The Prolific Unannotated Literary Poetry Corpus (PULPO) is a set of multilingual verses and stanzas with over 72 million words. It was created to tackle the needs of scholars interested in poetry from a machine learning perspective. Although poetry is a fundamental aspect of human expression that has been around for millennia, the study of poetry from a machine learning perspective is still in its infancy, largely due to the scarcity of poetic corpora. And while literary corpora are becoming more readily available, multilingual poetic corpora remain elusive. The lack of such corpora presents a major challenge for researchers interested in natural language processing (NLP) and machine learning (ML) applied to poetry.

Language	Verses	Words
English	2,945,882	21,129,934
Czech	1,888,680	10,451,247
German	1,583,504	9,686,032
Arabic	1,388,461	6,539,196
Finnish	1,046,162	3,377,398
Spanish	912,951	5,478,627
Italian	661,526	4,358,541
Russian	628,719	3,458,928
Hungarian	495,167	2,444,775
Chinese	436,384	1,649,711
Portuguese	346,974	2,302,886
French	223,928	1,672,759
Total	12,558,338	72,550,034

Table 1: Number of deduplicated verses and their words per language in PULPO.

The PULPO corpus comprises over 12 million deduplicated metrical verses from 12 different languages in 3 scripts (see Tables 1 and 6). We chose these languages because of the large number of poems freely available on the Internet out of copyright or with a permissive license. The poems date from the 15th-century to contemporary poetry and a number of them also have stanza separations. This makes the corpus a valuable resource for multilingual NLP and machine learning research. In addition, the corpus includes poems from various historical periods and literary traditions, providing a diverse range of poetic styles and forms.

3.2 Stanzas

To further evaluate the performance of the model, we conduct extrinsic evaluations using two different tasks. First, a stanza-type classification task for Spanish poetry. This task aims to assess the ability of the model to distinguish between different stanza types, such as tercet, quatrain, and sestina (see Table 2 for an example).

A stanza, which is considered the fundamental structural unit of a poem, serves to encapsulate themes or ideas (Kirszner and Mandell, 2007). Comprised of verses, stanzas are

[1] MLM is a form of self-supervised learning that involves masking some of the words in a sentence and training the model to predict them based on the surrounding words.

[2] An homage to Spanish poet Rafael Alberti.

Verse	Scheme
Escribí en el aren**a**l	8a
los tres nombres de la vida:	8-
vida, muerte, am**o**r.	6b
Una ráfaga de m**a**r,	8a
tantas claras veces d**a**,	8a
vino y nos borr**ó**.	6b

Table 2: Example of a stanza with its metrical length and rhyme scheme.

influenced by the writing styles and historical preferences of authors. The Spanish tradition boasts a rich abundance of stanza types, rendering their identification a challenging and intricate task. Generally, three factors contribute to the identification of a stanza: metrical length, rhyme type, and rhyme scheme (Domínguez Caparrós, 2014; Jauralde Pou, 2020; Quillis, 2000; Torre, 2000). Consequently, the classification of stanzas can be approached in three stages (Domínguez Caparrós, 2014):

1. Calculation of the metrical length per verse. This process typically involves counting the number of syllables while considering rhetorical devices that may alter this count (e.g., syneresis, synalephas). In some cases, the pattern formed by these verse lengths can assist in determining the stanza type.
2. Determination of the rhyme type. When the sounds after the final stressed syllable of each verse match, it is known as consonance rhyme. Alternatively, assonance rhyme involves the matching of vowel sounds while disregarding consonant sounds. However, there are stanza types where this distinction becomes irrelevant.
3. Extraction of the rhyme scheme. The rhyme scheme is established based on the verses that share a rhyme.

Following (Pérez Pozo et al., 2022), we approached stanza type identification as a classification task. We used their 5,005 Spanish stanzas containing between 12 and 170 examples for each of the 45 different types of stanzas[3], and used the already existing splits of 80% for training, 10% for validation, and 10% for testing.

3.3 Scansion

Second, a multilingual scansion task aimed at testing the ability of the model to predict the metrical pattern of a given verse in different languages. The scanning of a verse relies on assigning stress correctly to the syllables of the words. This process can be influenced by rhetorical figures and individual traditions. The synalepha is a common device in Spanish, English, and German poetry, which combines separate phonological groups into a single unit for metrical purposes. Syneresis and dieresis are two other devices that operate similarly but within the word, either joining or splitting syllables. The meter of a verse can be seen as a sequence of stressed and unstressed syllables, represented by the symbols '+' and '−', respectively. Examples 1, 2, and 3 from (De la Rosa et al., 2021) illustrate verses with metrical lengths of 8, 10, and 7 syllables in Spanish, English, and German, respectively. These examples also demonstrate the resulting metrical pattern after applying (or breaking, as in the case for '*la-her*' in the Spanish verse) synalepha, represented by '‿', and considering the stress of the last word as it may affect the metrical length in Spanish poetry.

(1) *cubra de nieve la hermosa cumbre*[4]
***cu**-bra-de-**nie**-ve-la-her-**mo**-sa-**cum**-bre*
+ − − + − − − + − + − 11
(Garcilaso de la Vega)

(2) *Our foes to conquer on th' embattled plain;*
*Our-**foes**-to-**con**-quer-on-th‿em-**bat**-tled-**pla**in;*
− + − + − − − + −+ 10
(Rhys Prichard)

(3) *Leise lausch' ich an der Thür*[5]
Lei**-se-**la**-schu‿ich-**an**-der-**Thür
+ − + − + − + 7
(Adolf Schults)

In order to measure the performance of Alberti, we follow the experimental design in (De la Rosa et al., 2021) and use their

[3] An extra stanza type 'unknown' was ignored in this study as it does account for anything not recognized as any of the other stanza types

[4] "[It] cover with snow the beautiful summit."

[5] "I quietly listen at the door"

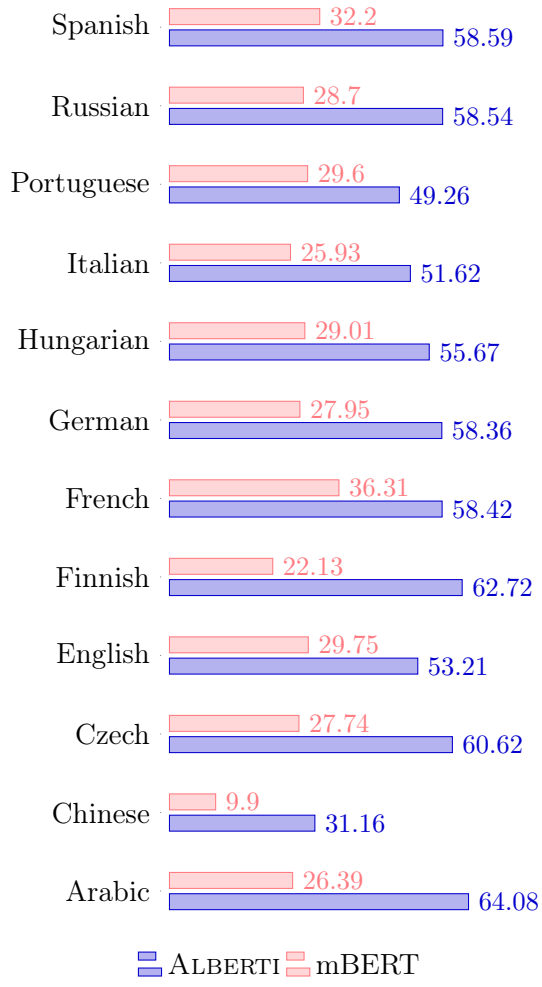

Figure 1: Masked Language Modeling accuracy (%) on the validation set of PULPO for ALBERTI (blue) and mBERT (red). Higher is better.

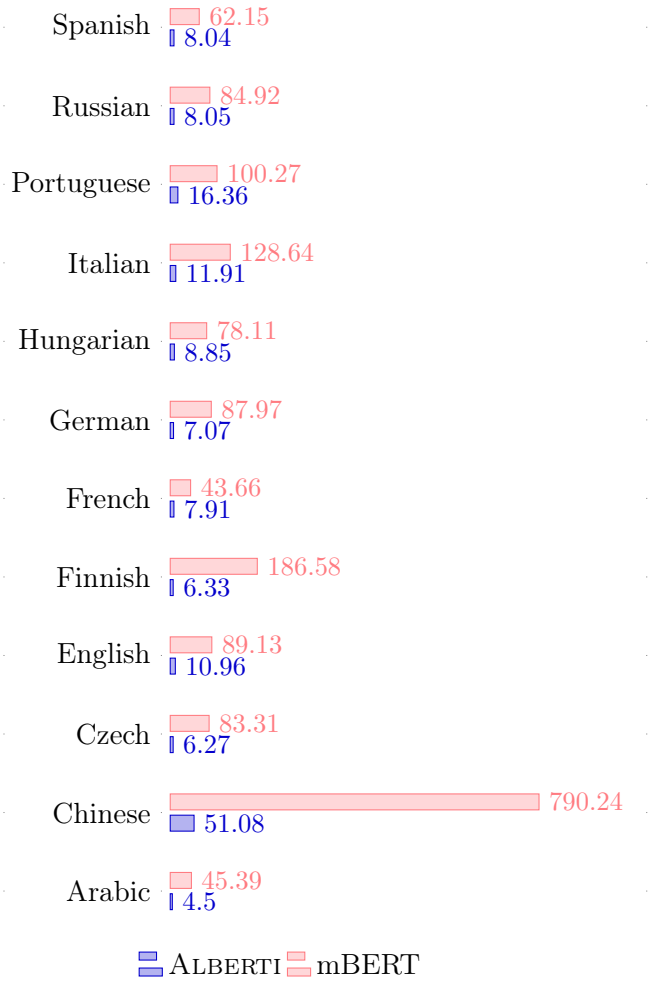

Figure 2: Perplexity proxy scores on the validation set of PULPO for ALBERTI (blue) and mBERT (red). Lower is better.

chosen datasets of verses manually annotated with syllabic stress for English, German, and Spanish. For the Spanish corpus, the *Corpus de Sonetos de Siglo de Oro* (Navarro-Colorado, Lafoz, and Sánchez, 2016) was used. This TEI-XML annotated corpus consists of hendecasyllabic verses from Golden Age Spanish authors. A subset of 100 poems initially used for evaluating the ADSO Scansion system (Navarro-Colorado, 2017) was selected for testing, while the remaining poems were split for training and evaluation.

Unfortunately, suitable annotated corpora of comparable scale were not found for English and German. Instead, an annotated corpus of 103 poems from *For Better For Verse* (Tucker, 2011) was used for English, and a manually annotated corpus from (Haider and Kuhn, 2018; Haider et al., 2020) was used for German. The German corpus contains 158 poems which cover the period from 1575 to 1936. Around 1200 lines have been annotated in terms of syllable stress, foot boundaries, caesuras and line main accent. These corpora were divided into train, evaluation, and test sets, following a 70-15-15 split. Table 4 shows number of verses per language and split.

Model	F1	Accuracy
mBERT	57.51	61.94
ALBERTI	**59.33**	**63.64**
BETO (Cañete et al., 2020)	–	42.12
Rantanplan (De la Rosa et al., 2020)		
+ Expert System (Pérez Pozo et al., 2022)	–	*88.51*

Table 3: F1 scores on stanza classification. Best neural model scores in **bold**. Rule-based systems *italicized*.

	Train	Evaluation	Test
Spanish	7,327	1,421	1,401
English	708	152	153
German	775	167	168

Table 4: Number of verses for each language in the metrical pattern prediction datasets.

4 Evaluation and Results

After training, we evaluated the resulting model ALBERTI on several fronts. For intrinsic evaluation, we used the aforementioned MLM metric as well as a perplexity proxy score based on the predicted token probabilities (see Listing 1). We calculated these metrics for every language on the validation set of PULPO for both ALBERTI and mBERT. As shown in Figure 1, the MLM accuracy of ALBERTI is generally higher than that of mBERT for all languages. The gains of ALBERTI against mBERT range from +19.65 percentage points for Portuguese to +40.59 for Finnish. A similar trend is shown for our perplexity proxy score in Figure 2, with clear gains of ALBERTI over mBERT across the board, ranging from -35.75 for French to staggering -739.16 points for Chinese. The stark difference for Chinese could be a result of differences in the way text is represented in that language in both the pre-training corpus of mBERT and PULPO.

For extrinsic evaluation, we also evaluated ALBERTI against mBERT for stanza classification and metrical pattern prediction. We chose the best performing models on the validation set over a small grid search of learning rates 10^{-5}, 3×10^{-5}, and 5×10^{-5}, for 3, 5, and 10 epochs, and warmup of 0 and 10% of the steps. Figure 3 shows the ROC curves of each stanza type versus the rest for both ALBERTI and mBERT, with higher areas under the curve (AUC) in 29 out of the 45 stanza types for ALBERTI, and in 16 out of 45 for mBERT. Table 3 shows F1 and accuracy macro scores for each model, with ALBERTI outperforming mBERT by a small percentage. Interestingly, our baseline fine-tuned mBERT model scores better than the monolingual Spanish BETO (Cañete et al., 2020) reported in (Pérez Pozo et al., 2022). Nonetheless, the combination of the rule-based system Rantanplan (De la Rosa et al., 2020) with an expert system remains state of the art for stanza classification.

The prediction of metre was approached as a multi-class binary classification task, i.e., one class per syllable where each syllable can be stressed (strong) or unstressed (weak). After a grid search with roughly the same hyperparameters as in (De la Rosa et al., 2021), ALBERTI outperforms mBERT for every language, as shown in Table 5. When compared to other similarly sized models (English RoBERTa (Liu et al., 2019) and multilingual XLM RoBERTa (Conneau et al., 2019)) as reported in (De la Rosa et al., 2021), it still performs better for English and German. Lastly, ALBERTI achieves a new state-of-the-art for German, as it performs better than both the large version of XLM RoBERTa and the rule-based system Metricalizer (Bobenhausen, 2011).

5 Conclusions and Further Work

In this work, we hope to make a significant contribution to the fields of Digital Humanities and NLP by introducing the first multilingual large language model for poetry, ALBERTI. Our model demonstrated substantial improvements over mBERT, indicating its effectiveness in capturing the nuances of poetic language in various languages and demonstrating the feasibility of domain-specific pretraining for poetry. The evaluation of the model on intrinsic and extrinsic metrics highlights its potential for practical applications

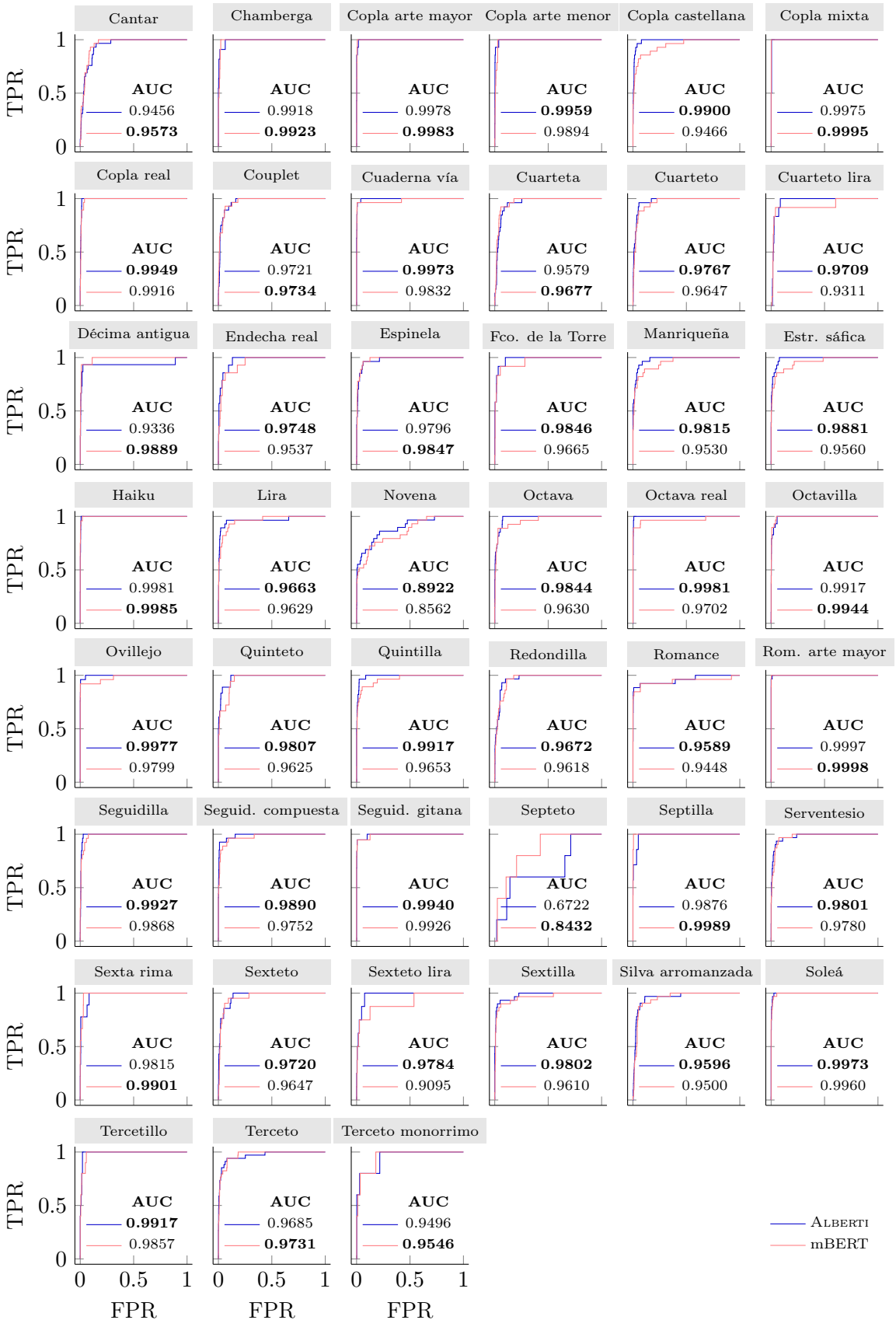

Figure 3: True positive rate (TPR) against false positive rate (FPR) of the receiver operating characteristic (ROC) curves and their corresponding areas (AUC) for the classification of each stanza type versus the rest after fine-tuning Alberti (blue) and mBERT (red). Best AUC score in bold.

Model	Spanish	English	German
mBERT	88.15	35.71	39.52
ALBERTI	91.15	**49.34**	**56.29**
RoBERTa (base) (De la Rosa et al., 2021)	87.37	36.21	43.11
XLM RoBERTa (base) (De la Rosa et al., 2021)	**92.15**	40.79	46.11
Rantanplan (De la Rosa et al., 2020)	*96.23*	–	–
Poesy (Algee-Hewitt et al., 2014)	–	*38.16*	–
Metricalizer (Bobenhausen, 2011)	–	–	*44.91*

Table 5: Accuracy on metrical pattern prediction. Best neural model scores in **bold**. Rule-based systems *italicized*.

in tasks such as stanza-type identification and scansion on a multilingual setting.

The release of our model and accompanying corpus will provide an important resource for researchers in the field, facilitating further investigation into poetry-related tasks. It is our plan to train ALBERTI at the stanza level and compare its performance against the current verse-based model, which presents itself as an exciting avenue for future research, as it could potentially improve the ability of the model to capture the meaning and structure of poetry in a more sophisticated way. Given the good results obtained by ALBERTI, despite its training on an arguably outdated model, future iterations will leverage more powerful and larger pre-trained models, thereby enhancing its performance and versatility.

Moreover, we do believe that the strong accuracy of ALBERTI in the masked language prediction task could pave the way for methods analyzing metaphoric language by leveraging the differences between the predictions of ALBERTI and the predictions of other models trained on more journalistic or encyclopedic type of data.

Overall, the results of this study have the potential to significantly advance our understanding of poetry in various languages and contribute to the development of more sophisticated NLP models that can capture the subtleties of poetic language. We hope that our work will inspire further research and innovation in this field, and we look forward to seeing how our model and corpus will be used in future studies.

Acknowledgments

Research for this paper has been partially supported by the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) obtained by Elena González-Blanco, a project funded by the European Research Council (`https://erc.europa.eu`) (ERC) under the research and innovation program Horizon2020 of the European Union.

References

Agirrezabal, M., I. Alegria, and M. Hulden. 2017. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 18–23, Varna, Bulgaria, September. INCOMA Ltd.

Algee-Hewitt, M., R. Heuser, M. Kraxenberger, J. Porter, J. Sensenbaugh, and J. Tackett. 2014. The stanford literary lab transhistorical poetry project phase ii: Metrical form. In *DH*.

Anttila, A. and R. Heuser. 2016. Phonological and metrical variation across genres. In *Proceedings of the Annual Meetings on Phonology*, volume 3.

Araújo, J. and J. Mamede. 2002. *Classificador de Poemas*. CCTE, Lisbon.

Bobenhausen, K. 2011. The metricalizer2–automated metrical markup of german poetry. *Current Trends in Metrical Analysis, Bern: Peter Lang*, pages 119–131.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

De la Rosa, J., Á. Pérez, M. de Sisto, L. Hernández, A. Díaz, S. Ros, and E. González-Blanco. 2021. Transformers analyzing poetry: multilingual metrical pattern prediction with transformer-based language models. *Neural Computing and Applications*.

De la Rosa, J., Á. Pérez Pozo, L. Hernández, S. Ros, and E. González-Blanco. 2020. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.

De Sisto, M. 2020. *The interaction between phonology and metre: Approaches to Romance and West-Germanic Renaissance metre*. Ph.D. thesis, Radboud University Nijmegen.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Domínguez Caparrós, J. 2014. *Métrica española*. Editorial UNED.

Gervás, P. 2000. A logic programming application for the analysis of spanish verse. In *Computational Logic—CL 2000: First International Conference London, UK, July 24–28, 2000 Proceedings*, pages 1330–1344. Springer.

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct.

Haider, T., S. Eger, E. Kim, R. Klinger, and W. Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. *arXiv preprint arXiv:2003.07723*.

Haider, T. and J. Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 81–86.

Ibrahim, R. and P. Plecháč. 2011. Toward automatic analysis of czech verse. *Formal methods in poetics*, pages 295–305.

Jauralde Pou, P. 2020. Métrica española. *Madrid: Cátedra*.

Kirszner, L. G. and S. R. Mandell. 2007. *Literature: Reading, reacting, writing*. Thomson/Wadsworth.

Lau, J. H., T. Cohn, T. Baldwin, J. Brooke, and A. Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958, Melbourne, Australia, July. Association for Computational Linguistics.

Lennard, J. 2006. *The poetry handbook*. OUP Oxford.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. cite arxiv:1907.11692.

Manjavacas Arevalo, E. and L. Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India, December. NLP Association of India (NLPAI).

McAleese, W. G. M. 2007. Improving scansion with syntax: An investigation into the effectiveness of a syntactic analysis of poetry by computer using phonological scansion theory. Technical report, Department of Computing Faculty of Mathematics, Computing and Technology The Open University.

Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Navarro-Colorado, B., M. R. Lafoz, and N. Sánchez. 2016. Metrical annotation

of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.

Ormazabal, A., M. Artetxe, M. Agirrezabal, A. Soroa, and E. Agirre. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3655–3670, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Pérez Pozo, A., J. de la Rosa, S. Ros, E. González-Blanco, L. Hernández, and M. de Sisto. 2022. A bridge too far for artificial intelligence?: Automatic classification of stanzas in spanish poetry. *Journal of the Association for Information Science and Technology*, 73(2):258–267.

Quillis, A. 2000. *Métrica española*. Grupo Planeta (GBS).

Schweter, S. and L. März. 2020. Triple e - effective ensembling of embeddings and language models for ner of historical german. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece, September. CEUR-WS.org.

Šeļa, A., P. Plecháč, and A. Lassche. 2022. Semantics of european poetry is shaped by conservative forces: The relationship between poetic meter and meaning in accentual-syllabic verse. *Plos one*, 17(4):e0266556.

Torre, E. 2000. *Métrica española comparada*, volume 48. Universidad de Sevilla.

Tucker, H. F. 2011. Poetic data and the news from poems: A" for better for verse" memoir. *Victorian Poetry*, 49(2):267–281.

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

A Appendix: PULPO

PULPO, the Prolific Unannotated Literary Poetry Corpus, is a set of multilingual corpora of verses and stanzas with over 72M words.

The individual corpora were downloaded using the Averell tool, developed by the POSTDATA team, and other sources found on the Internet.

A.1 Averell sources

A.1.1 Spanish

- Disco v3
- Corpus of Spanish Golden-Age Sonnets
- Corpus general de poesía lírica castellana del Siglo de Oro
- Gongocorpus - source

A.1.2 English

- Eighteenth-Century Poetry Archive (ECPA)
- For better for verse

A.1.3 French

- Métrique en Ligne - source

A.1.4 Italian

- Biblioteca italiana - source

A.1.5 Czech

- Corpus of Czech Verse

A.1.6 Portuguese

- Stichotheque

A.2 Internet sources

A.2.1 Spanish

- Poesi.as - source

A.2.2 English

- A Gutenberg Poetry Corpus

A.2.3 Arabic

- Arabic Poetry dataset

A.2.4 Chinese

- THU Chinese Classical Poetry Corpus

A.2.5 Finnish

- SKVR

A.2.6 German

- TextGrid Poetry Corpus - source
- German Rhyme Corpus

[6]The poems as such are not available as lines that "looked like" poetry where extracted from books in the Project Gutenberg. See `https://github.com/aparrish/gutenberg-poetry-corpus`.

Name	Language	Poems	Verses	Words	Period
THU Chinese Classical Poetry Corpus	Chinese	127,682	510,728	2,553,640	$1^{st} - 17^{th}$ C.
TextGrid Poetry Corpus	German	105,849	3,422,223	20,735,344	$15^{th} - 20^{th}$ C.
SKVR	Finnish	89,247	1,340,987	4,290,341	$16^{th} - 20^{th}$ C.
Corpus of Czech Verse	Czech	66,428	2,664,989	12,636,867	$18^{th} - 20^{th}$ C.
Arabic Poetry dataset	Arabic	54,300	54,944	5,328,745	$1^{st} - 16^{th}$ C.
Biblioteca Italiana	Italian	25,341	1,070,717	7,121,246	$15^{th} - 20^{th}$ C.
poesi.as	Spanish	25,300	910,800	5,894,900	$15^{th} - 21^{st}$ C.
19000 Russian poems	Russian	19,315	691,361	6,559,283	$15^{th} - 20^{th}$ C.
Poems in Portuguese	Portuguese	15,543	362,537	2,073,420	$15^{th} - 21^{th}$ C.
ELTE verskorpusz	Hungarian	13,161	594,284	4,606,974	$16^{th} - 19^{th}$ C.
Métrique en Ligne	French	5,081	247,248	1,850,222	$17^{th} - 20^{th}$ C.
Sonetos Siglo de Oro	Spanish	5,078	65,911	466,012	$16^{th} - 17^{th}$ C.
Disco V3	Spanish	4,530	54,066	431,428	$15^{th} - 20^{th}$ C.
Eighteenth C. Poetry Archive	English	3,084	265,683	2,063,668	$18^{th} - 18^{th}$ C.
German Rhyme Corpus	German	1,948	47,900	270,476	$17^{th} - 20^{th}$ C.
Stichotheque	Portuguese	1,702	260,536	168,411	$15^{th} - 20^{th}$ C.
Gongocorpus	Spanish	481	20,621	99,490	$16^{th} - 17^{th}$ C.
Poesía Lírica Castellana Siglo de Oro	Spanish	475	51,219	299,402	$16^{th} - 17^{th}$ C.
For Better For Verse	English	103	1,084	41,749	$15^{th} - 20^{th}$ C.
A Gutenberg Poetry Corpus	English	N/A[6]	3,085,117	22,124,040	$15^{th} - 20^{th}$ C.

Table 6: Number of poems, verses and words, and the approximate coverage period for each corpus in PULPO.

A.2.7 Hungarian

- ELTE verskorpusz

A.2.8 Portuguese

- Poems in Portuguese

A.2.9 Russian

- 19,000 Russian poems

B Appendix: Availability

- ALBERTI: `https://huggingface.co/linhd-postdata/alberti-bert-base-multilingual-cased`
- PULPO: `https://huggingface.co/datasets/linhd-postdata/pulpo`

C Appendix: Perplexity Proxy Score

```python
def score(sentence, model, tokenizer):
    model_inputs = tokenizer(sentence, add_special_tokens=False, return_tensors="pt")
    scores, count = [], 0
    for input_index in range(len(model_inputs["input_ids"][0])):
        masked_token = tokenizer.decode(
            model_inputs["input_ids"][0][input_index], skip_special_tokens=True)
        if len(masked_token) > 0:
            model_inputs["input_ids"][0][input_index] = tokenizer.mask_token_id
            scores.append(fill(model_inputs, targets=[masked_token])[0]["score"])
            model_inputs = tokenizer(
                sentence, add_special_tokens=False, return_tensors="pt")
    return math.pow(math.prod(scores), -1 / len(scores))
```

Listing 1: Perplexity proxy score implementation in Python pseudocode.