Machine Learning Techniques for Classifying the Dual Writing System in Palaeohispanic Inscriptions: A Comparative Study of Neural and Non-neural Approaches

Técnicas de Machine Learning para la clasificación del sistema de escritura dual en inscripciones paleohispánicas: Un estudio comparativo de enfoques neuronales y no neuronales.

Gonzalo Martínez-Fernández,[1] **José Francisco Quesada-Moreno**[1]
Francisco José Salguero-Lamillar[2] **Agustín Riscos-Núñez**[1]
[1]Department of Computer Science and Artificial Intelligence,
University of Seville, Avda. Reina Mercedes s/n, 41012 Seville, Spain
[2]Department of Spanish Language, Linguistics and Theory of Literature,
University of Seville, C/ Palos de la Frontera s/n 41004, Seville, Spain
{gmartinez2,jquesada,salguero,ariscosn}@us.es

Abstract: The writing systems of the Celtiberian and Iberian languages —part of the Palaeohispanic languages— share a special characteristic: the dual system. This is crucial for the interpretation of the inscriptions because if it is present, some graphemes must be read differently. But it is not an evident property; therefore, this paper proposes the comparison of seven classification systems based on Machine Learning techniques, both quantitatively and qualitatively, to determine if they can be useful for historical linguists.
Keywords: Machine Learning, Neural Networks, Palaeohispanic languages, Classification.

Resumen: Los sistemas de escritura del celtíbero y el íbero, parte de las lenguas paleohispánicas, comparten una característica especial: el sistema dual. Se trata de una propiedad crucial para la interpretación de las inscripciones porque, de estar presente, altera la lectura de algunos grafemas. Sin embargo, no es una característica evidente; por lo tanto, en este documento se propone la comparación cuantitativa y cualitativa de siete modelos de clasificación basados en técnicas de Machine Learning para determinar si pueden ser de utilidad para los linguistas históricos.
Palabras clave: Aprendizaje automático, Machine Learning, redes neuronales, lenguas paleohispánicas, clasificación.

1 Introduction

The Palaeohispanic languages are a heterogeneous group of languages that were spoken in the Iberian Peninsula before the arrival of the Romans. The best attested languages are Celtiberian, a Celtic language (Beltrán-Lloris and Jordán-Cólera, 2020), and Iberian, considered a language isolate although related to Basque by some authors (Orduña-Aznar, 2022). Even though Celtiberian and the Iberian language are not genetically related, they share very similar writing systems, called semi-syllabaries: systems that employ graphemes for both single phonemes and (occlusive) syllables.

The original script from which both the Iberian and the Celtiberian evolved was probably designed for a language with a different phonological system (de Hoz-Bravo, 1986; Rodríguez-Ramos, 1997; Ferrer-i-Jané and Moncunill-Martí, 2022); thus, the adapted script might not have been well suited for Celtiberian and Iberian. These changes on the writing system make the phonetic transcription of the Celtiberian words more difficult, since it is not possible to detect certain phenomena without prior knowledge, like *muta cum liquida* clusters (Beltrán-Lloris and Jordán-Cólera, 2020).

Another problem with both the Celtiberian and Iberian writing systems is the distinction between voiced and unvoiced occlusive syllables. Excluding *b-*, which does not have the opposition *p-* in Iberian and always uses the same grapheme

as *p-* in Celtiberian, some texts show a written difference between the graphemes representing the voiced syllables *d-*, *g-*, and their voiceless counterparts *t-*, *k-*. However, other texts do not present this distinction. Therefore, one must know the language in order to read the words correctly. This distinction is called the *dual system*, and the pairs of variants of the same sign are called *dualities*. When both variants appear in the same text, they are called *explicit dualities* (Ferrer-i-Jané, 2021). The latter are important in determining whether a text was written using the dual system or not. Ferrer-i-Jané believes that not only the signs with known explicit dualities present such opposition in their voiceness, as did Jordán-Cólera (Jordán-Cólera, 2005; Jordán-Cólera, 2007). Thus, the absence of explicit duality in a text does not mean it was not written taken the dual system into consideration.

The Palaeohispanic languages have not really had many computational approaches in the past. Therefore, the field could benefit from the creation of new automated systems, and the work of linguists could be alleviated with regards to the detection of the dual system in Palaeohispanic inscriptions. A statement by Ferrer-i-Jané (Ferrer-i-Jané, 2021) serves as inspiration and as the starting hypothesis to create and explore different Machine Learning models that detect if an inscription presents the dual system or not: "In Iberian, the procedence, chronology and palaeography can be used as heuristics". Therefore, these attributes will be the base of our experiments, which try to prove the previous hypothesis empirically and provide relevant insights.

The paper is structured in four main sections. The first one includes previous work by authors regarding Palaeohispanic Languages in a computational context, as well as works about poorly attested languages and other computational techniques. The following section describes the sources and the transformations applied to the data that are used in the experiments, along with the research questions that are answered. Afterwards, the results are shown and discussed. Lastly, some conclusions are included.

2 Previous work

Palaeohispanic languages are considered *Trümmersprachen*, that is, languages with no native speakers that have only survived through the records of the people who spoke them (Untermann, 1980). Therefore, working with these languages implies little to no linguistic data depending on the context. Furthermore, the topic of classifying inscriptions by the presence of the dual system is still an unexplored area of research. This linguistic mechanism is applicable only to the Palaeohispanic languages given the unique writing systems developed in the Iberian Peninsula; thus, inspiration must be taken from similar classification projects applied to other languages, either due to the nature of the problem in binary classification, the low-resource nature of the languages, or given the nature of the data.

Some authors have previously worked on other Machine Learning projects applied to the Iberian or the Celtiberian languages, like Luo et al. (Luo et al., 2021), who introduced phonetic geometry into a mixed generative system to extract cognates from unsegmented text in the Iberian language by using Basque as support, among others. However, this field is still vastly unexplored, probably due to the nature of the languages. Low-resource languages are a topic of much interest in the research area, since they are both a challenge and an opportunity to bring Machine Learning advances to smaller communities. The task of binary classification in low-resource languages is growing every year given the goal to bring the models closer to those communities. The living languages of the Iberian Peninsula are receiving more care in this respect, with examples in Galician (Alonso and Gamallo, 2025) and Basque (Goikoetxea et al., 2024), and other ancient languages have received similar approaches, like Phoenician (Rizk et al., 2021), Latin (Kase, Heřmánková, and Sobotková, 2021) and Kannada texts (Soumya and Kumar, 2014).

3 Data

The main source of Palaeohispanic corpora for this project is the Hesperia database, developed under the guidance of Javier de Hoz, in the Department of Greek Philology and Indo-European Linguistics at the Complutense University, in Madrid (Estarán-Tolosa, Orduña-Aznar, and Luján, 2009). This webpage offers three main categories of Palaeohispanic data: epigraphic, onomastic

and numismatic. Despite all of them being valuable, this project employs only the epigraphic data, for these texts are the ones with annotations for the dual system. This work uses the dataset described in (Martínez-Fernández et al., 2026), which in turn has been adjusted for the proposed experiments.

Based on the hypothesis of Ferrer-i-Jané, the main attributes used were *the location where a text was found*, *the approximate dating*, and *the epigraphic text* itself, transcribed into Latin characters. There is no possibility to work directly with the palaeographical properties of the inscriptions yet, so the transcribed texts are used instead. Their combination was tested under six different systems to determine whether one attribute was indispensable for the performance of the system. The location of an inscription is given as the latitude and longitude coordinates of the municipality where the inscription had been found, while the dating attribute is provided as a tuple (*mean*, *width*) based on the approximate interval of an inscription, all as *float* values.

The epigraphic text contains annotations in an adapted form of the Leiden Conventions (Salomies and Bodel, 2001; Moncunill-Martí and Velaza, 2019), where missing or doubtful parts of the texts, as well as corrections or restorations, are indicated. Most of the annotations have been previously simplified or removed, although they require extra steps for the experiments to follow. All the steps that have been carried out in order to obtain a clean, usable text corpus are listed next.

1. FALSA and SUSPECTA entries are removed, since they are likely forged inscriptions, as well as GRAECA and LATINA inscriptions.
2. The dual system is only present in the Iberian Levantino (of Eastern Spain) and Celtiberian signaries, so other entries are filtered out.
3. Listing items are detected and inscriptions are split to obtain smaller, more manageable fragments. Some of these fragments are as long as 449 characters, while the average is about 19.
4. Null values are filled using the mean for the location coordinates, and with the minimum and maximum values for the *dating_min* and *dating_max* features, which are then used to calculate the missing *dating_mean* and *dating_width* features.

Experiments have been carried out with a version of the text with these conventions, as well as with another version where the Leiden annotations were completely removed (e.g. [− − −],+…), so as to test whether the change helps in the classification.

The other major modification of the dataset is performed on the epigraphic texts. They are given as transcriptions of the Palaeohispanic semi-syllabic characters, which may belong to the dual system or to the non-dual system. However, the transcriptions present a contradiction. The non-dual inscriptions use "simple" graphemes for both *t-* and *d-* syllables, as well as for *k-* and *g-*. On the other hand, the dual inscriptions introduce complex variants to differentiate *t-* and *k-*, leaving the simple graphemes for *d-* and *g-*. The non-dual inscriptions (simple graphemes) are always transcribed as *t-* and *k-* while the same graphemes in the dual inscriptions are transcribed as *d-* and *g-*. Therefore, the same graphemes have different transcriptions, and the same transcriptions correspond to different graphemes. To maintain the consistency, all the non-dual transcriptions have their *t-* and *k-* replaced by *d-* and *g-*, as shown in the example in Table 1. The correct reading of the text is not a problem in this scenario because the focus of the task lies with the correct classification within the dual system. With a clean corpus, the texts are split into character n-grams, and mapped into numerical values for neural networks. A TF-IDF representation is also provided for both neural networks and other systems as supporting features.

	NO DUAL	**DUAL**	
Grapheme			
Reading	ta/da	da	ta
Transcription	ta	da	ta
Replacement	da	da	ta

Table 1: Depiction of the transcription contradiction in the inscription dataset and the proposed fix for this document's experiments.

Lastly, the classification attribute, "dual system", is not binary: even though all the texts are either Celtiberian or Iberian, some of the signaries employed for writing Palaeohispanic texts do not present this feature and thus are marked as "NO APLICA" ("DOES NOT APPLY" in English). Since these inscriptions are not relevant and only constitute 1% of all the epigraphic inscriptions, they are ignored. Thus, this attribute's values are either "DUAL" or "NO DUAL". Sadly, there is class imbalance in the corpus, as shown in Table 2. This will be addressed later by assigning class weights to the instances during the training process as $w_c = \frac{|\mathcal{D}|}{|C| \cdot |\mathcal{D}_c|}$ for every class $c \in C$ in the dataset $\mathcal{D}$.

Dual System category	Absolute frequency	Relative frequency
NO DUAL	1390	72.58%
DUAL	525	27.42%

Table 2: Proportion of instances of the dataset presenting the dual system or not.

Other categorical attributes are introduced for posterior experiments: *material*, *medium*, and *word separators*. These values need to be expressed in integers when techniques such as decision trees are to be applied, changing into one-hot vectors for the neural models.

In the end, a dataset comprised of 1915 instances and 8 attributes has been collected and preprocessed. In the experiments that require a non-annotated version of the texts, the inscriptions that end up with less than 3 characters are removed; thus, those experiments are performed with 1789 instances.

4 Experiment Definitions

Stemming from the initial hypothesis, three research questions have been established, with an experiment set up for each one. The first two involve the training and evaluation of different neural and non-neural models against different combinations of attributes to test the initial hypothesis. Meanwhile, the last one focuses on the qualitative aspects of the models.

4.1 Experiment 1. Is Ferrer's hypothesis plausible?

The first experiment tests different combinations of the three main attributes —*location*, *chronology*, and the *text*— with several neural networks architectures, as well as non-neural models. Afterwards, their accuracy and binary F1-score results are collected and compared. Since no previous work has attempted the same classification problem, the baseline will be considered as performing better than always predicting the same, most common value; therefore, an accuracy baseline of 72.58% from the "NO DUAL" class.

The neural architectures considered range from the most basic to more complex approaches. However, LLM models such as Transformers are not used since they require large amounts of data, which is —sadly— not the case. The model with the best metrics scores will become the baseline for further experiments. The neural architectures are listed next, while the accompanying Figure 1 summarises the internal structure of the different models.

1. A multilayered perceptron with 1 hidden layer (MLP).
2. A convolutional network (CNN) to capture n-gram relationships with a window of size 3.
3. A bidirectional GRU neural network (RNN) to capture the sequential nature of the texts.
4. A stacked encoder architecture with attention (ATT) with a Multi-Head Attention layer for text self-attention, in one encoding module.

Each neural layer is composed of 16 neurons for the embedding layers, and 64 neurons for the hidden layers. The Appendix A offers a comparison in the total parameter size of the neural models. The hidden layers use the *ReLU* function, while the output layer is computed with a *sigmoid* function. Training is restricted to 50 epochs and a batch size of 32, with an Adam optimiser and a learning rate set to 10^{-4}. All the parameters have been empirically tested, and a seed with value 42 has been employed for the replicability of the results. All the models except the multilayered perceptron employ a branched architecture: one part of the system

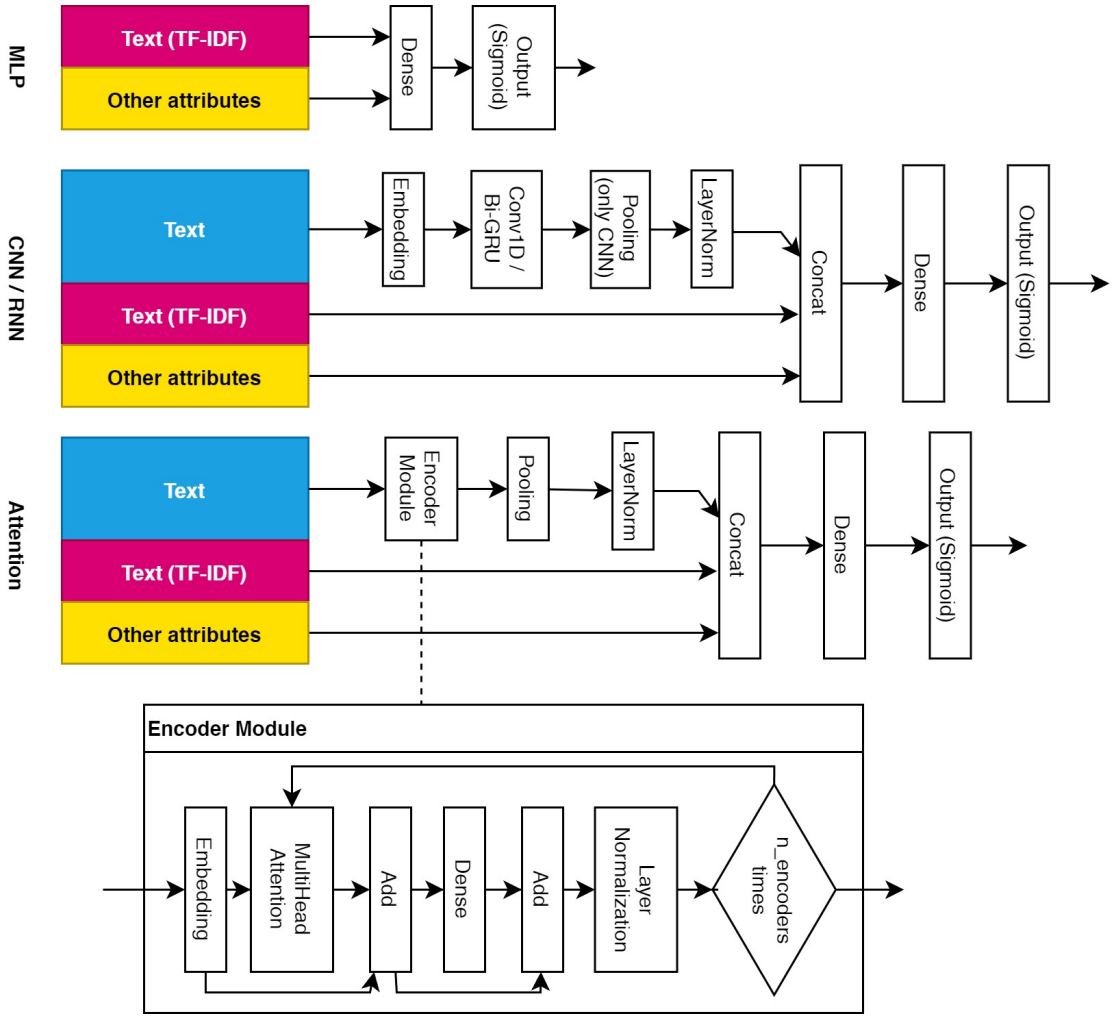

Figure 1: The multilayered perceptron is the simplest architecture with 1 feed-forward layer. The convolutional network and the Bi-GRU model share a similar structure, only differing in one layer. The branched architecture processes the text apart from the other attributes. Lastly, the Encoder architecture with attention (bottom right) allows a more complex combination of layers to process the text. The output of the Encoder module can be re-introduced in an attempt to improve the model's performance.

processes the text, which is concatenated to the other non-textual attributes afterwards, before being passed through a feed-forward layer and the output function.

Additionally, other non-neural Machine Learning models have also been employed in the study, to check whether they provide a better performance than the neural networks.

1. Decision Tree Classifier (DTC)
2. Random Forest Classifier (RFC)
3. XGBoost Classifier (XGB) (Chen and Guestrin, 2016)

The non-neural models are fed the same data as the neural ones, except for the text attribute, which is only introduced as a character TF-IDF vector representing the inscription. Next, details regarding the configuration of the Machine Learning models are provided. First, after empirically analysing the best Decision Tree parameters, a maximum depth of 8 was chosen, leaving the rest of parameters as default. Similarly, the number of estimators for the Random Forest Classifier was set to 100 while maintaining the previous Decision Tree configuration. The XGBoost model is also restricted to a depth of 8.

Finally, data is split into *train*, *validation* and *test* sets in a 0.8/0.1/0.1 proportion. 10-fold Cross-Validation is applied during training.

4.2 Experiment 2. Does performance improve with more attributes?

The second experiment picks up the thread of the previous experiment and aims to check whether adding other unused, categorical attributes such as *separators*, *material*, and *medium* to the input data improves the metrics of the classification systems. The base case will be the best model tested in Section 4.1 for both neural and non-neural versions, which will be both retrained and evaluated. Their results will be compared to the base case in order to determine if there has been an improvement.

4.3 Experiment 3. Are qualitative properties of the inscriptions useful to detect the origin of misclassifications?

The last experiment takes a different approach than the previous ones. Its objective is to observe and study the properties of the misclassified inscriptions to determine the reasoning of the system behind the wrong assignment of the classes. In order to achieve that, the following process is set up. First, for the models created in each fold during Cross Validation, the whole dataset —including all the training and test instances— is fed, and a set of misclassified inscriptions is retrieved. Then, the errors are studied grouped by these categories: dual inscriptions with evidence, dual inscriptions without evidence, and non-dual inscriptions, so the group that poses a bigger problem for the classifiers can be identified. The instances that are always misclassified, independently of the fold, are called *persistent errors*. For each of these errors, their qualitative attributes will be studied to determine why they happen and if it is the fault of the data used. Moreover, an explanation is tried to be given as a plausible explanation for the results obtained. Finally, the importance of the attributes is studied so as to discover which part of the data is more valuable for the model's classification.

5 Results

The evaluation of the experiments described previously are presented in the following section, along a brief discussion.

5.1 Experiment 1. Is Ferrer's hypothesis plausible?

The results of the models of this experiment are averaged from the results of the 10 folds and collected in Table 3 (accuracy) and Table 4 (binary F1-score). The standard deviation is also provided within parentheses. The abbreviations of the tables work as follows. The first column values represent the subdasatet combination used, describing presence of text (T), annotated (AT) or not (NAT); of the location attribute (L), and the chronology attributes (C), as well as their absence ($\neg$T, $\neg$L, $\neg$C). The first row depicts the initials or abbreviation of the model names.

Although Tables 3 and 4 only show 2 decimals, more were taken into account during the evaluation process. The best accuracy and F1-score values are achieved by the XGboost model and the Random Forest architecture, followed closely by the Decision Tree Classifier model, the CNN architecture, and the RNN network, the first two with *non-annotated text* and both numerical features. The best dataset combination is NAT, C, L followed closely by AT, C, L. It is evident from the results that text is crucial for the evaluation of the inscriptions. However, the other two attributes do not seem to have such an impact, except when the input does not contain text.

In general terms, only the MLP model seems to perform worse than the rest. Since there are 1915 entries in the dataset, each of the 10 folds containing around 191 entries, a 1% difference in accuracy is equal to only 2 instances. Therefore, we can treat most of the results of the best models as equivalent. Surprisingly, no model stands out noticeably; the main takeaway is that the shape of the data is crucial for its correct classification. In this sense, Experiment 3 tries to detect why this happens. Apart from that, another question emerges: whether the results improve by adding more features to the training dataset. This is explored in the following experiment.

5.2 Experiment 2. Does performance improve with more attributes?

The second experiment's main goal is to test whether more attributes or their combination provide better results for the classification of Palaeohispanic inscriptions into the dual or non-dual system. Instead of testing the same models from the previous experi-

	MLP	CNN	RNN	ATT	DTC	RFC	XGB
¬T,¬C,L	0.73(0.00)	-	-	-	0.80(0.04)	0.80(0.04)	0.75(0.03)
¬T,C,¬L	0.73(0.00)	-	-	-	0.79(0.02)	0.80(0.02)	0.76(0.02)
¬T,C,L	0.73(0.01)	-	-	-	0.83(0.02)	0.84(0.03)	0.82(0.02)
AT,¬C,¬L	0.92(0.03)	0.94(0.02)	0.94(0.02)	0.94(0.02)	0.93(0.03)	0.94(0.02)	0.94(0.02)
AT,¬C,L	0.92(0.02)	0.94(0.02)	0.94(0.02)	0.94(0.02)	0.94(0.01)	0.95(0.02)	0.95(0.02)
AT,C,¬L	0.92(0.02)	0.94(0.02)	0.94(0.02)	0.94(0.02)	0.93(0.02)	0.94(0.02)	0.94(0.02)
AT,C,L	0.92(0.02)	0.94(0.02)	0.94(0.02)	0.94(0.02)	**0.95(0.01)**	0.95(0.02)	**0.96(0.01)**
NAT,¬C,¬L	**0.93(0.02)**	0.94(0.01)	0.94(0.02)	0.94(0.02)	0.93(0.02)	0.94(0.02)	0.94(0.02)
NAT,¬C,L	0.92(0.03)	**0.95(0.01)**	**0.94(0.01)**	0.94(0.02)	**0.95(0.01)**	0.94(0.02)	0.95(0.01)
NAT,C,¬L	**0.93(0.02)**	**0.95(0.01)**	0.93(0.01)	0.94(0.02)	0.94(0.01)	0.94(0.02)	0.95(0.01)
NAT,C,L	0.91(0.02)	0.94(0.01)	**0.94(0.01)**	0.94(0.02)	0.94(0.01)	**0.96(0.01)**	0.96(0.02)

Table 3: Comparative table of the average accuracy results by model (columns) and subdataset (rows). The standard deviation is shown between parentheses. The best results for each model are highlighted.

	MLP	CNN	RNN	ATT	DTC	RFC	XGB
¬T,¬C,L	0.84(0.00)	-	-	-	0.87(0.02)	0.87(0.02)	0.81(0.03)
¬T,C,¬L	0.84(0.00)	-	-	-	0.86(0.01)	0.87(0.01)	0.83(0.02)
¬T,C,L	0.84(0.00)	-	-	-	0.89(0.01)	0.90(0.02)	0.87(0.02)
AT,¬C,¬L	0.95(0.02)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.95(0.02)	0.96(0.01)	0.96(0.01)
AT,¬C,L	0.95(0.01)	0.96(0.01)	0.96(0.02)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.97(0.01)
AT,C,¬L	0.95(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.95(0.01)	0.96(0.01)	0.96(0.01)
AT,C,L	0.95(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.97(0.01)	0.97(0.01)
NAT,¬C,¬L	0.95(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.95(0.01)	0.96(0.01)	0.96(0.01)
NAT,¬C,L	0.95(0.02)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.97(0.01)
NAT,C,¬L	0.95(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)
NAT,C,L	0.94(0.02)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.97(0.01)	0.97(0.01)

Table 4: Comparative table of the average binary F1-score results by model (columns) and subdataset (rows). The standard deviation is shown between parentheses. The best results for each model are not highlighted due to their similarity.

ment, only one representative of the neural and non-neural models will be used: the CNN and the XGBoost model, respectively. These models were chosen as they had the absolute best results in the previous experiment, taking into account more decimals than shown in the tables. The extra attributes —*material*, *medium*, and *separators*— will be added to the best combination of attributes from before (even if the results are practically equivalent): *non-annotated text*, *chronology*, and *location* values. Tables 5 and 6 show the comparative results in accuracy and binary F1-score for the two models, respectively. The nomenclature of the first column represents the presence of the medium (ME), separators (SE), and material (MA) attributes, as well as their absence (¬ME, ¬SE, ¬MA).

The results obtained are again too close to each other to determine whether one model or dataset is better than the other. A 1% difference still represents about 2 inscriptions, so they are considered practically equivalent. Therefore, the addition of these three attributes has not really helped improve the already high evaluation metrics. Maybe other attributes are needed, so other similar experiments can be set up. However, it might be more fruitful to examine the qualitative aspects of the results obtained in order to determine whether the missing instances for a perfect accuracy have an explanation.

	CNN	XGB
¬ME,¬SE,¬MA	0.94(0.01)	0.96(0.02)
¬ME,¬SE,MA	0.95(0.01)	0.96(0.01)
¬ME,SE,¬MA	0.95(0.01)	0.96(0.02)
¬ME,SE,MA	0.95(0.01)	0.96(0.02)
ME,¬SE,¬MA	0.94(0.01)	0.96(0.02)
ME,¬SE,MA	0.95(0.01)	0.96(0.01)
ME,SE,¬MA	0.95(0.01)	0.96(0.02)
ME,SE,MA	0.95(0.01)	0.96(0.02)

Table 5: Accuracy summary of the CNN and XGBoost models after adding the medium, separators, and material features to the subdataset. All the combinations used also have the base combination NAT,C, L.

	CNN	XGB
¬ME,¬SE,¬MA	0.96(0.01)	0.97(0.01)
¬ME,¬SE,MA	0.96(0.01)	0.97(0.01)
¬ME,SE,¬MA	0.96(0.01)	0.97(0.01)
¬ME,SE,MA	0.96(0.01)	0.97(0.01)
ME,¬SE,¬MA	0.96(0.01)	0.97(0.01)
ME,¬SE,MA	0.96(0.01)	0.97(0.01)
ME,SE,¬MA	0.96(0.01)	0.97(0.01)
ME,SE,MA	0.96(0.01)	0.97(0.01)

Table 6: Binary F1-score summary of the CNN and XGBoost models after adding the medium, separators, and material features to the subdataset. All the combinations used also have the base combination NAT,C, L.

5.3 Experiment 3. Are qualitative properties of the instances useful to detect the origin of misclassifications?

The last experiment takes a more qualitative approach than the previous tests. It starts from the base model of the previous experiment, the CNN and XGBoost models with the data combination NAT, C, L, because adding the other three attributes did not seem to improve the results significantly. The first step consists in examining how many instances are always classified either correctly or incorrectly grouped by three categories: dual instances with evidence (t or k), dual instances without evidence, and non-dual instances. Table 7 offers the absolute and relative count of persistent successes, that is, instances that are correctly classified across all folds. As it can be observed, the classification of dual instances with evidence is trivial. On the other hand, the dual inscriptions without evidence prove to be the most difficult to classify for the models, probably because the texts do not have any clues to differentiate themselves from the non-dual inscriptions. Figure 2 shows the number of instances that are classified incorrectly in at least one fold grouped by the number of folds, whose sum corresponds to the difference with the total in Table 7. The results for each model are practically opposite; while the XGBoost model hardly ever misclassifies a word in more than one fold, when the CNN model misclassifies an inscription, it tends to do it in most folds. Therefore, should a voting system be implemented between the folds of the XGBoost model, the accuracy might improve with the existing data.

Category	CNN	XGB	Both (∩)
Dual with evidence	402 (1.00)	402 (1.00)	402 (1.00)
Dual without evidence	9 (0.08)	62 (0.55)	9 (0.08)
Non-dual	1267 (0.99)	1252 (1.00)	1244 (0.99)
Total	1678 (0.94)	1716 (0.96)	1655 (0.93)

Table 7: Number of instances that are always correctly classified by either and both models. Between parentheses are left the relative count with respect to the total number of instances in each category.

Further evaluations can be performed over the previous results. For instance, the texts that have been classified incorrectly at least once can be examined in order to determine the cause of their misclassification, whether it is the model's fault or the shape of the data is misleading. To achieve the set of inscriptions with at least one error in both models, first, the union of misclassifications between folds is performed, and afterwards, the intersection between the models' sets is done. In the end, the analysis is performed over 50 instances, all belonging to the dual class

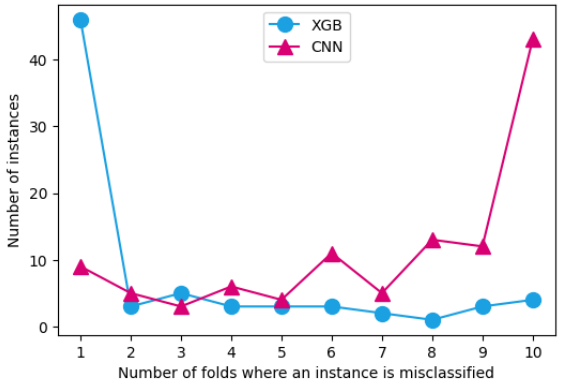

Figure 2: The chart depicts the number of instances that have been classified incorrectly by number of folds in which they have been misclassified. The dots correspond to the convolutional network results while the triangles refer to the XGBoost model.

without evidence. Figure 3 summarises the results, where the inscriptions are categorised by the conclusion obtained after analysing them with extra data from Hesperia. The cross pattern indicates that a category cannot confirm the dualness of the inscription. A full itemisation of the categories is found in Appendix B in the form of Hesperia instance identifiers.

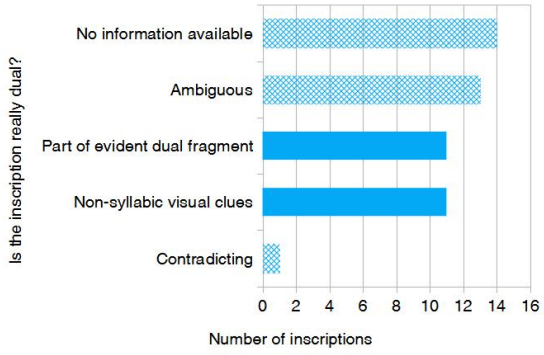

Figure 3: The graph shows the number of misclassified instances categorised by the reasons that determine if they are truly dual. The categories with the cross pattern bar cannot be confidently determined as dual, while the categories with filled bars are definitely dual.

The categorisation of the misclassifications indicates that one of the main problems in the prediction of the instances may be the lack of access to the grapheme realisations within the texts, which usually provides explicit markers of a text belonging to the dual system; for instance, **ŕ** has variants that are more common in dual texts and variants that are more common in non-dual texts, even though the sound is the same. Short fragments of text that belong to a larger dual inscription but lack explicit markers are another major problem in the detection of the dual system. This could be fixed by not splitting inscriptions, although it may generate futher problems if the strings are too long. The third group of instances are said to be dual but without providing any details or illustrations to check it, so we have to trust the authors. The rest of the instances cannot be confidently determined to be dual, so they are either an error on the annotator's side, or the fact that they are naturally ambiguous causes the misclassifications.

The problem of classifying instances without evidences may be due to the fact that the models were trained using more instances with evidence. To test if training with no evidence allows the model to generalise better, a new pair of CNN and XGBoost models have been trained without the instances that show evidence, but all have been used in testing. Table 8 depicts the number of instances that have been correctly classified in every fold, grouped by the type of inscription, as in Table 7. In the end, the CNN model cannot classify any dual instance, whereas the XGBoost classifies 8 and 6 dual instances correctly, at the cost of less non-dual instances.

Category	**CNN**	**XGB**	**Both (∩)**
Dual with evidence	0 (0.00)	8 (0.02)	0 (0.00)
Dual without evidence	0 (0.08)	6 (0.05)	0 (0.00)
Non-dual	1273 (0.99)	1242 (0.97)	1240 (0.97)
Total	1273 (0.71)	1256 (0.70)	1240 (0.69)

Table 8: Number of instances that are always correctly classified by either and both models after being trained with only non-evident instances. The relative count with respect to the total number of instances in each category is left between parentheses.

The final study tries to determine which feature is given more importance by the mod-

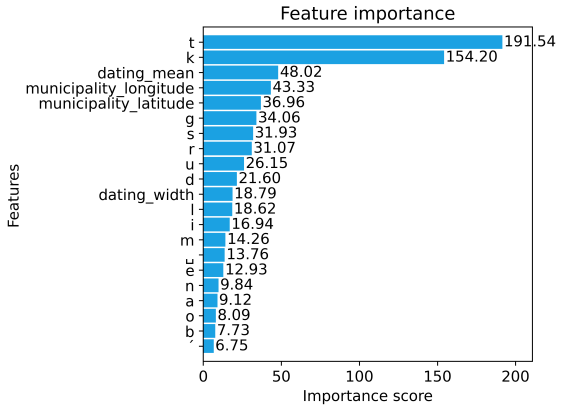

Figure 4: Feature importance of the last tree in the trained XGBoost model.

els in order to classify an instance. The selected model is XGBoost because explainability follows naturally after this type of structure. Figure 4 provides an example graph of the feature importance —by coverage, number of samples affected by the split on said feature— of the last tree of the XGBoost model in one of its folds. All of the folds show similar results. As it was expected, the features representing t and k in the input TF-IDF vector acquire the greatest importance in differencing a dual text from a non-dual one. Remarkably, non-textual attributes are the following most important features.

6 Conclusions

In this work, a series of 7 models, 4 neural and 3 non-neural, whose objective is to classify Celtiberian and Iberian inscriptions in the dual system, are presented. A first experiment compares the results in accuracy and the binary F1-score metrics. Then, more attributes are included as input values so as to improve the classification results of the systems. After a thorough research, given a 10-fold Cross Validation evaluation, the best results come from the XGBoost Classifier and Random Forest Classifier models, with dataset combination $(NAT, C, L, \neg ME, \neg SE, \neg MA)$, 96% accuracy and 97% F1-score averaged over the 10 folds, followed by the Convolutional Neural Network (NAT, C, L, ME, SE, MA) with 95% average accuracy and 96% binary F1-score. Nonetheless, these high metric results mask one critical problem: only non-dual and dual inscriptions with evidence can be confidently classified, and dual inscriptions without markers still entail difficulties for the models. Lastly, a qualitative analysis was performed to detect the probable reason behind the common misclassifications between models and folds. A large proportion of these issues might be solved by working with the grapheme realisations of the characters, although there is no available digitalised corpus and thus it is left as future work. Not splitting the texts from the inscriptions might also help improving the metrics. It is evident that the models learn to rely on the presence of the t and k characters to classify an inscription as dual; however, when there is no explicit evidence, the ambiguity with respect to non-dual texts affects the results negatively.

Limitations were prevalent in the study: The Palaeohispanic languages represent a very small, fragmented corpus, with a high class imbalance between the dual and non-dual inscriptions, and the absence of palaeographical features restricted the evaluation of the last element of Ferrer-i-Jané's hypothesis. Despite the evident difficulty in classifying the dual inscriptions, the trained models can still be useful by establishing an interpretation of possible dual characters in less time. In addition, this project can serve as an essential piece for a Palaeohispanic OCR model, since it provides clues about how to transcribe the texts if instead of training on transcribed inscriptions, the original characters (through an arbitrary mapping) were used.

Acknowledgements

This research work was funded by the Spanish Ministry of Science and Innovation via a FPU 2024 grant, number 00763, awarded to Gonzalo Martínez-Fernández.

This work was partially supported by project PID2025-171542NB-I00, funded by MICIU/AEI/10.13039/501100011033.

References

Alonso, A. and P. Gamallo. 2025. Evaluating Galician language models for sentiment analysis on challenging linguistic phenomena. *Procesamiento del lenguaje natural*, 74:191–205.

Beltrán-Lloris, F. and C. e. a. Jordán-Cólera. 2020. Celtibérico. *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, (20):631–688.

Chen, T. and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cuadrado, D. 1950. El plomo con inscripción ibérica del Cigarraiejo (Mula Murcia). *Cuadernos de historia primitiva*.

de Hoz-Bravo, J. J. 1986. La epigrafía celtibérica. In *Epigrafía hispánica de época romano-republicana.*, pages 43–102. Institución "Fernando el Católico".

de Hoz-Bravo, J. J. 2022. Método y métodos: Estudiar las lenguas paleohispánicas como disciplina. In *Lenguas y epigrafías paleohispánicas*, pages 15–37. Bellaterra.

Estarán-Tolosa, M. J., E. Orduña-Aznar, and E. R. Luján. 2009. El banco de datos Hesperia. *Palaeohispanica*.

Ferrer-i-Jané, J. and N. Moncunill-Martí. 2022. Sistemas de escritura paleohispánicos: clasificación, origen y desarrollo. In *Lenguas y epigrafías paleohispánicas*, pages 97–130. Bellaterra.

Ferrer-i-Jané, J. e. a. 2021. El sistema dual de la escritura celtibérica desde la perspectiva ibérica. *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, 21:399–434.

Goikoetxea, J., M. Etxabe, M. García, E. Guzzi, and M. Alonso. 2024. Multilabel discourse function classification of lexical bundles in Basque and Spanish via transformer-based models. *Procesamiento del Lenguaje Natural*, 73:29–41.

Jordán-Cólera, C. 2005. ¿Sistema dual de escritura en celtibérico? *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, 5:1013–1030.

Jordán-Cólera, C. e. a. 2007. Estudios sobre el sistema dual de escritura en epigrafía no monetal celtibérica. *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, (7):101–142.

Kase, V., P. Heřmánková, and A. Sobotková. 2021. Classifying Latin inscriptions of the Roman empire: A machine-learning approach. In *Computational Humanities Research 2021*, pages 123–135.

Luo, J., F. Hartmann, E. Santus, R. Barzilay, and Y. Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81.

Martínez-Fernández, G., J. F. Quesada, A. Riscos-Núñez, and F. J. Salguero-Lamillar. 2026. Curation of a palaeohispanic dataset for machine learning. *arXiv preprint arXiv:2604.13070*.

Moncunill-Martí, N. and J. Velaza. 2019. *Lexikon der iberischen Inschriften. Léxico de las inscripciones ibéricas*. Ludwig Reichert Verlag, Wiesbaden.

Orduña-Aznar, E. 2022. La teoría vasco-ibérica. In *Lenguas y epigrafías paleohispánicas*, pages 247–268. Bellaterra.

Rizk, R., D. Rizk, F. Rizk, and A. Kumar. 2021. A hybrid capsule network-based deep learning framework for deciphering ancient scripts with scarce annotations: A case study on Phoenician epigraphy. In *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 617–620. IEEE.

Rodríguez-Ramos, J. 1997. Sobre el origen de la escritura celtibérica. *Kalathos: Revista del seminario de arqueología y etnología turolense*, (16):189–198.

Salomies, O. and J. P. Bodel. 2001. Epigraphic evidence: Ancient history from inscriptions.

Soumya, A. and G. H. Kumar. 2014. Classification of ancient epigraphs into different periods using random forests. In *2014 Fifth International Conference on Signal and Image Processing*, pages 171–178. IEEE.

Untermann, J. 1980. Trümmersprachen zwischen Grammatik und Geschichte. In *Trümmersprachen zwischen Grammatik und Geschichte: 245. Sitzung am 16. Januar 1980 in Düsseldorf*. Springer, pages 7–40.

A Parameter size of neural models by dataset combination

	MLP	CNN	RNN	ATT
¬T,¬C,L	257	-	-	-
¬T,C,¬L	257	-	-	-
¬T,C,L	385	-	-	-
AT,¬C,¬L	2753	10785	43361	5873
AT,¬C,L	2881	10913	43489	6001
AT,C,¬L	2881	10913	43489	6001
AT,C,L	3009	11041	43617	6129
NAT,¬C,¬L	1857	9665	42241	4753
NAT,¬C,L	1985	9793	42369	4881
NAT,C,¬L	1985	9793	42369	4881
NAT,C,L	2113	9921	42497	5009

Table 9: Number of trainable parameters for each model and dataset combination. The number of parameters for the RNN is much higher than the rest. With less parameters (neuron units), the results are slightly worse, so maintaining the same number of neurons in every model was preferred.

B Breakdown of misclassified instances by probable motive as Hesperia identifiers

No information available:

- AUD.05.13
- AUD.05.20
- HER.02.016
- HER.02.159
- HER.02.250
- HER.02.283
- HER.02.336
- HER.02.360A
- HER.02.360B
- HER.02.361
- PYO.01.06
- PYO.02.18
- PYO.07.06
- V.07.07

Part of an evident dual inscription:

- AUD.05.02
- AUD.05.38a
- B.23.01
- HER.02.258
- HER.02.301
- PYO.03.02d (.05)
- PYO.03.02f (.07)
- V.06.007
- V.06.012
- V.07.02
- V.21.01

Non-syllabic visual clues:

- V.06.017
- HER.02.254
- AUD.05.05
- PYO.07.33
- PYO.03.02f (.07)

Ambiguous:

- B.22.02
- B.44.02
- GI.10.15
- GI.13.06
- GI.15.21
- GI.15.43
- GI.17.05
- HER.02.030
- HER.02.180
- HER.02.324
- PYO.01.23
- V.06.051
- V.06.078

Contradicting:

- V.02.02