www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(3), 2026, 82-97. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i3.1348

Four Lines, Different Deaths: Exploring the Classification of Humor in *Calaveritas* Poems

Victor Manuel Palma-Preciado[1], Liana Ermakova[2], Grigori Sidorov[1,✉], Carolina Palma-Preciado[1]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, (CIC-IPN), Mexico City, Mexico
[2] Université de Bretagne Occidentale, HCTI – EA 4249, Brest, France
E-mails: vpalmap2019@cic.ipn.mx, lermakova@gmail.com, sidorov@cic.ipn.mx, cpalmap2020@cic.ipn.mx

✉Corresponding author: Grigori Sidorov

Abstract. *Calaveritas* are seen as a sort of poem or ode to the dead, since is a Mexican tradition linked to the Day of the death, in which through text you could see the personification of death and popular characters been mock and satirize, this topic is quite interesting because it have humor and the structure of the more serious poem. The classification of this textual genre could help locate text that contain humor with unconventional variants or structures and to preserver in a way this written scheme. To tackle this task, it was decided use a machine learning approach for the baseline, taking quite good results around 94% on the F1-score for the top methods of the baseline, in this case the main approach was to finetune Transformers like BETO or BERT-multilingual obtaining 98% and 97% on de F1-score and analyze the similarities and to observer the characteristic inherent to each class. The classes were quite separable since the calaveritas are more near related to humor than to the classical approach of poem, since the text of these classes contain words that are more easily identifiable. Given the observed degree of separability between classes, we sought to ensure that the classification was not primarily driven by topical information. To this end, we masked the most frequent words in each class as a preliminary control experiment, the produced results were broadly comparable to those obtained through fine-tuning our main models, suggesting that structural features may play a role in the classification process. In a way humour intervenes to create a poem-like structure with a humoristic content, A hybrid, perhaps.

Keywords: Humor, Poems, *Calaveritas*, Deep Learning, Humorous Text, BERT-like

Article Info
Received Ene 26, 2026
Accepted Mar 11, 2026

1 Introduction

In the tradition of Spanish-language poetry, it is common to find a variety of structures with diverse metrical patterns, widely used in songs, couplets, and other poetic compositions. This raises an interesting question: are poetic metrics intended solely for lyrical expression, or can they also be used as a tool to generate humor? Is it a creative adaptation, or merely an imitation of the form that lacks the technical requirements of a proper quartet? What is certain is that even if poetry is filled with melancholic themes, we can also find satire, which plays a significant role in building humor through traditional poetic forms.

Humor is often seen as a response to life's lighter moments, a counterpoint to sorrow and grief. In Mexican culture, however, there is a distinct form of humor that uniquely intertwines with death, giving rise to a vibrant expression known as calaveritas literarias.

These are short, rhymed compositions that blend humor and satire, often featuring popular figures, family members, and well-known characters, making even death a subject of jest (calaveras-literarias.com, 2020).

This form of writing is closely tied to the celebration of the Day of the Dead (Marchi, 2022), sometimes serving as a ritual to honor and remember the deceased. Yet we must not overlook the diversity of humor, the several styles and intentions that it can embody. Beyond remembrance, calaveritas often serve as a medium for social commentary, used to mocking societal norms or raising awareness about pressing issues.

A distinctive element of calaveritas is their use of creative nicknames and representations of death (Romero-González, 2018). This ingenuity is deeply rooted in Mexican culture, where death is personified through playful names such as "the bony one" (la huesuda), "the skeletal one" (la esquelética), and "the skinny one" (la flaca). Additionally, the figure known in some countries as the Elegant Lady or the Dapper Skull is a translation of La Catrina, whose imagery is inspired by historical Mexican attires. These representations not only highlight cultural creativity but also reflect a unique perception of death—one that blends reverence with humor. In the broader context of Mexican culture, this perspective gives rise to mortuary celebrations where calaveritas lighten the sense of loss and commemorate those mentioned in the text—or, in some cases, mock the subjects of the composition. This blend of cultural and social satire is central to this research, which aims to explore whether calaveritas can be detected as a unique form of humor or, given their rhymed structure, whether they align more closely with poetry (Argüelles, 2003).

It is worth noting that, in many aspects, these two forms share compositional ties. As they are short yet similar and employ more complex structures than regular jokes, they present an interesting subject for research using a computational approach. This study aims to leverage deep learning models like transformers to classify calaveritas and non-calaveritas texts that in this case are poems, to explore whether calaveritas exhibit a unique and distinctive form that differentiates them from traditional poetry. Employing computational methods, we seek to uncover the structural, linguistic, and thematic characteristics that define this culturally significant literary form.

One key distinction between calaveritas and poems lies is their approach. Calaveritas use free verse, allowing for varied combinations, whereas traditional poetry often adheres to defined syllabic patterns, such as octosyllabic or decasyllabic structures. In this study, we focus on four-line stanzas without emphasizing syllabic counts, to better capture the flexible and creative nature of calaveritas.

Furthermore, they do not adhere to a rigid rhyme scheme. Their rhymes, whether assonant or consonant, often vary and may even be deliberately forced to fit the intended humor or meaning. Despite this flexibility, their construction typically follows three main requirements, two of which are essential. First, they must address the concept of death or a related theme, often referencing the timeless figure of death. Second, they must incorporate satire or mockery, often targeting a well-known individual or public figure. Lastly, though not strictly necessary, calaveritas typically consist of four or more verses, with the length varying according to the creator's style.

Original Text in Spanish:

"Es calavera el inglés, calavera el italiano,
calavera fue el francés, lo mismo Maximiliano;
el Pontífice romano y todos los cardenales
en la tumba son iguales: Calaveras del montón…
Los ricos por su elegancia…
los pobres por su miseria…
todos en la sepultura…
serán… calaveras del montón."

English Translation:

The Englishman is a skull, the Italian is a skull,
the Frenchman was a skull, the same with Maximilian;
the Roman Pontiff and all the cardinals
in the grave are the same: skulls of the heap…
The rich for their elegance…
the poor for their misery…
all in the grave…
will be… skulls of the heap.

For example, each line above carries a nuance that, within this historical framework, offers a glimpse into the culture behind the calaverita. For example, when it refers to Maximilian, it alludes to Maximilian of Habsburg, Emperor of Mexico imposed by France and executed in 1867. Then it mentions the Roman Pontiff and the cardinals, alluding to the Pope and the high ecclesiastical hierarchy. In its final lines, it concludes with a moral that exposes social disparity and delivers a satirical critique: regardless of nationality, status, or religion, all will end up in the cemetery, for the only certainty is that death comes to all.

In summary, calaveritas literarias represent a rich intersection of humor, cultural expression, and social critique. Through their playful engagement with death and their satirical tone, they provide a unique lens for understanding the creative ingenuity and cultural identity of Mexico. By applying computational methods to classify and analyze these texts, this study aims to uncover the distinctive structural and thematic features of calaveritas, exploring how they differ from other forms of poetry and contributing to a deeper understanding of this culturally significant literary form.

2 Related Works

The use of calaveritas as a poetic or humorous theme can highlight their connection to rhyme and suggests that they may share more characteristics than commonly believed. Following this line of thought, several works in the area can be found, especially those that utilize classifier models to analyze them.

For instance, in (Winters & Delobelle, 2020), the study examines the use of negative examples in humor classification, focusing on the opposite pole of humorous intent. The study points out that current classifiers often rely on specific vocabularies to detect humor, rather than interpreting the semantic message that makes a phrase humorous. This creates difficulties, as the polarity of certain texts complicates the distinction between humorous and non-humorous texts, making the classification task even more challenging.
Tackling poetry as a holistic analysis proves to be quite difficult and time-consuming. In this case, (Deng et al., 2021) proposes the use of similarity models and word embeddings, combined with graph theory, to classify and ultimately analyze the distinctive features obtained and the correlation between rhyme, diction, and metaphor.

Alternatively, (Marco et al., 2021) discusses a framework for classifying poems, using various methods to evaluate the feasibility of classifying poetry with Artificial Intelligence (AI). The study employs techniques such as SVM and Bootstrap in combination with features representation like Tf-idf and Doc2Vec, which achieved the best performance in poetic classification.

Similarly, the approach of using one-class classification methods for detecting hurtful humor in (De La Rosa et al., 2021) found that the main TF-idf and LLM approaches did not achieve good results, and it was discovered that the one-class SVM model performed better for this task. The authors mentioned that this process might not be well-suited for applications in real-world scenarios, as it considers any non-creative or standard methods as humorous. They emphasize the need for a large non-humorous dataset to help improve this classification.

There are also other types of studies where the analysis of performance metrics is the focus, such as the one proposed by (Jhamtani et al., 2019). In this work, the analysis is conducted through two approaches: one called Rantanplan, which manages part-of-speech tagging, syllabification, stress assignment, and metrical adjustment. The other approach, referred to as Jumper, achieves automatic poetic metric analysis by taking a different route. Instead of relying on syllabification, it uses scansion without it. Additionally, within the same line of poetic analysis, (Tang et al., 2023) integrates these tools to enable more comprehensive analyses. This integration allows for interconnection with previously mentioned tools, such as Rantanplan, JolyJumper, and HismeTag, among others, making it easier to analyze poetic corpora.

Studies such as Panda et al. (2025) demonstrate the effectiveness of machine learning approaches, including algorithms such as Naive Bayes and Random Forest, for text classification in complex domains like sentiment analysis in social networks. Their methodological proposal provides a solid foundation for addressing similar tasks, such as the automatic classification of humor.

One challenge often encountered is the need for enough data for classification. the use of a methods such as synonym replacement, place exchange word, deletion word and Data boost to ensure robustness into the data set, have been use but can be quite limiting to use them since they focus on doing lexical and semantic variety which dilute the robustness of the text, in this case in (Winters & Delobelle, 2021) the author opt to use a way method to ensure a correct augmentation, using their own variation of Cognate-based, Antonym-based, and Antipode-based methods for it.

Kolesnikova (2025) explores the use of transformer architectures (e.g., BERT and RoBERTa) to capture complex semantic relationships between words, such as lexical functions. This ability to model meaning in context is fundamental for other tasks, such as humor detection, which often relies on semantic and pragmatic nuances.

The novelty of generating text could improve its understanding in this matter (Kesarwani et al., 2021) focuses on the interpretation of adversarial networks in generating text with rhymes. The method provides insights into the discriminator and, therefore, obtains rhymes in quatrains (four verses) or limericks (five verses) with a certain structure. In dataset selection, they observe an improvement over human evaluation, achieving quite good results with their proposed method.

In a similar way (Inácio & Oliveira, 2023) explores the use of a transformed base model to tackle generative satire with Bert-like models and genetic algorithms for evolving text. The approach involves generating satirical headlines in comparison to human-edited ones, yielding good results. This method has the potential to open pathways in the creative field that are challenging to understand, with humor and poetry being two of the main areas that could benefit from this method.

3 Methodology

As part of the methodology, five key aspects were addressed. The first step involved gathering the necessary data to analyze calaveritas, enabling the creation of a dataset to tackle the classification problem: distinguishing between calaveritas and non-calaveritas (poetic structures of four verses). Following this, the preprocessing and electing embedding techniques for training applicable models. Two primary representations were utilized: one based on Tf-idf and the other leveraging the Universal Sentence Encoder (USE) (Cer et al., 2018).

The third phase focused on selecting machine learning models were chosen to establish a baseline and to explore the most effective approaches for classification. Fourth, transformer-based architecture, including multilingual BERT (Devlin et al., 2018) and BETO (Cañete et al., 2020), were fine-tuned for model training, and finally, the fifth is the same fine-tuned model but with a masked representation.

3.1 Dataset

Before collecting the data is necessary to understand and define what a calaverite is, as mentioned before they are short, satirical, and humorous poetic compositions traditionally written in Mexico during the Día de los Muertos celebrations. Calaveritas are used to playfully mock people, including public figures, friends, or family, by imagining their death or adventures in the afterlife, based on these characteristics, we developed a dataset that reflects such textual properties (Palma Preciado et al., 2024).

The calaverita follows a free verse composition style. Unlike traditional poetry, which typically adheres to a defined structure with a specific syllable count per line and a set rhyme scheme, free verse provides poets with greater flexibility to experiment with line length, rhythm, and language. For instance, there are various rhyme strategies; if a composition contains patterns such as ABBA, AABB, or any of their variations, it indicates the presence of rhyme, whether it is consonant or assonant, appearing constructions of four lines with consonant rhymes:

Original Text in Spanish:
“La muerte llegó brincando (A)
A Robertita le trajo sandía (B)
Una lotería (B)
Y unos dulces que quería.” (B)

English Translation:
"Death came jumping
To little Roberta it brought watermelon,
A lottery ticket
And some candies she wanted."

As it can observe in this calaverita, only the second and last lines rhyme, as the remaining lines are used merely as a preamble to the idea that a person named Robertita will receive a visit from death, to play with it. This composition is a clear example of how

calaveritas use the figure of death to interact with a known character, which only makes sense to the person who created it. However, it remains a strategy to satirize the actions of death, which, for obvious reasons, lacks the ability to play a more down-to-earth stratagem.

In the counterpart class, the negative class in the dataset, the poems are written in the structure of quatrains. These are poetic compositions consisting of exactly four lines. This structure is common in many poetic forms and can follow various rhyme schemes. Depending on the type of poem, the lines may have a specific number of syllables and may be referred to by different names.

Such constructions can express a variety of ideas, ranging from reflections to poetic lyrics. However, this does not mean that calaveritas are far from traditional poetry, as there are also satirical poems and poems with death-related themes. This comparison and classification between them are, therefore, important to analyze.
An example of a poem in this case follows the four-verse scheme but not in a traditional way, as it is constructed with an AABA format:

Original Text in Spanish:
"Tome a mi voz la métrica del persa (A)
A recordar que el tiempo es la diversa (A)
Trama de sueños ávidos que somos (B)
Y que el secreto soñador dispersa." (A)

English Translation:
I took my voice the meter of the Persian
To remember that time is the diverse
Plot of eager dreams that we are
And that the dreaming secret disperses.

Some factors considered include the structure and composition of the texts. This is the primary distinction between the two selected categories. In modern times, calaveritas have transitioned toward a freer, less rigid verse format. While they often consist of four lines, this is not a strict requirement, as they can include additional verses. However, for the purposes of this research, both classes were standardized to four-verse compositions, independent of each other.

For data collection and selection, the material was gathered entirely through web scraping, utilizing sources such as Reddit, Twitter (though little usable content was obtained from Twitter, as the focus was on well-formed structures), Wattpad, and specialized blogs featuring four-line compositions. The data underwent extensive manual curation, as it is easy to overlook instances with more than four lines. The primary focus was on calaveritas, given their tendency to be created in a freer manner.
After manual filtering, a dataset of 1,330 samples across the two categories was compiled:

- 610 non-Calaveritas (poetic constructions of four verses, restrictions on rhyme scheme, to ensure diversity).
- 720 Calaveritas (ranging from traditional to more casual and amateur styles).

The experimental evaluation of XAIS was conducted using eight publicly available biomedical datasets that differ in dimensionality, class distribution, and clinical context. These datasets were selected to examine the behavior of the proposed method under heterogeneous conditions, ranging from moderate to high dimensionality and from balanced to markedly imbalanced class scenarios. The selection covers clinical and diagnostic scenarios related to oncology, dermatology, metabolic disorders and immunology.

3.2 Preprocessing and text representation

Since the collected data comes from various sources, it is necessary to standardize its representation. This includes how each verse is separated, how line breaks are indicated, and addressing cases where the scraper splits a text into multiple lines (e.g., one for each verse). In such instances, it is essential to merge these fragments into a single entry that contains all the verses in one record. However, an interesting aspect arises: the structure of each class benefits from preserving spaces and, most importantly, line breaks. In poetry, these elements carry valuable information about the strategies used to divide the composition while maintaining—or deliberately breaking—the rhyme between verses. Therefore, this feature must be considered, as it is crucial for poetic analysis. A line break, for instance, can indicate a pause or add emphasis.

In the case of text representation, transforming the text from letters into numerical formats for model training was necessary. Data preparation was tailored to different approaches, such as Tf-idf and embeddings. For the Tf-idf representation, tokenization, stop word removal, lemmatization, and the deletion of special characters were performed. Conversely, for Multilingual Universal Sentence Encoder (USE) [14] and the Neural-Net Language Models (NNLM) (Bengio et al., 2003) embeddings, preprocessing beyond tokenization was avoided, as additional processing could result in the loss of vital information. This retained information is critical for accurate classification during training. To accomplish this, preprocessing was performed using the NLTK (Bird & Loper, 2004) library. TensorFlow (Abadi et al., 2015) and Kaggle files were used to implement the embeddings. Since TensorHub was unavailable.

To account for different analytical perspectives and potential edge cases, the most frequent words identified within each class were masked as seen in Table 1, reducing a purely topical interpretation of the classification and enabling the models to capture structural patterns, including quatrains, rhyme schemes, and other poetic features. To gain finer control over the masking process, we used Transformer-based models implemented in PyTorch, applying the following word sets respectively, even though several of these terms do not occur in all samples.

Table 1. Representative lexical sets by class.

Category	Representative Lexicon
Calaveritas	muerte, panteón, calaca, catrina, huesuda, calavera, flaca, parca, muertos, murió, calaveras, difunto, cementerio, tumba, esqueleto, hueso, ánima, guadaña, sepulcro, ataúd, osario
Poems	alma, corazón, amor, vida, sueño, flor, luz, mar, sol, luna, besos, rosa, noche, memoria, pasión, nostalgia, melancolía, ausencia, recuerdo, ternura, suspiro, deseo, pena, tristeza, esperanza

3.3 Baseline

To evaluate the performance of traditional machine learning algorithms for this task, multiple approaches were selected, encompassing probabilistic, function-based, and tree-based models. Six algorithms were chosen:

- SVM (Support Vector Machine);
- RF (Random Forest);
- NB (Naïve Bayes);
- Gradient Boost;
- AdaBoost;
- Decision Tree.

The training of these models was conducted using the Scikit-learn *(Pedregosa et al., 2011)* library. To ensure consistency and comparability, the hyperparameters for each one were kept at their default settings. This approach aimed to provide a baseline understanding of how these models perform without extensive parameter tuning. This initial configuration serves as a reference point to evaluate traditional algorithms against newer and more complex methods like transformers.

3.4 Transformers

In the case of transformers, which are more complex models requiring significant resources but benefiting from pre-training, a fine-tuning process was employed to implement them. Given that strong performance has been obtained using BERT-based models, they were selected for the classification task. Since the original BERT (Devlin et al., 2018) is trained on an English corpus, three variants were considered: a multilingual version (BERT-M) (Devlin et al., 2018) supporting multiple languages, including Spanish, and two Spanish-specific models, BETO (Cañete et al., 2020) and DistilBETO (Cañete et al., 2022), developed by the University of Chile.

To train these models, the Ktrain (Maiya, 2020) wrapper combined with the Colab environment was used to accelerate and simplify the training process. The models were retrieved from the Hugging Face repository, with BETO (Cañete et al., 2020) available as dccuchile/bert-base-spanish-wwm-uncased and DistilBETO (Cañete et al., 2022) as dccuchile/distilbert-base-spanish-uncased.

Various hyperparameters were evaluated, but Table 2 presents the configurations that produced the best results. The maxlen parameter was set based on the longest text in the dataset, which had a length of seventy-two characters. Since only one instance had this length, it was rounded to seventy for the model. Additionally, training for 2 epochs yielded good performance without overfitting, as the dataset was small for fine-tuning. Finally, the Colab environment utilized a Tesla T4 GPU with 15 GB of RAM and CUDA version 12.2.

Table 2. Hyperparameters and settings used for fine-tuning the transformer models.

Model	Learning rate	Epoch	Bach size	Maxlen
BERT-Mult	**2e-5**	**2**	**8**	**70**
BETO	5e-5	2	8	70
DistilBETo	2e-5	2	16	70
BETO Masked	2e-5	3	16	70

4. Dataset analysis

In this section, the analysis of the dataset used for the classification task is present, which consists of two classes: calaveritas literarias in Spanish and a counterpart of poems (Non-calaveritas). Both classes consistently follow a structure of four verses. As presented before there are different rhyme schemes present; thus, if the verses follow patterns such as ABBA, AABB, or any of their variants, it indicates the presence of rhyme, whether consonant or assonant.

In general, constructions with four verses and consonant rhyme are predominant. This structure is typical of calaveritas, where rhyme plays an essential role in the rhythm and tone of the poem. Additionally, calaveritas in the dataset often incorporate playful or humorous themes, which are characteristic of this genre, quite contrary of the scheme that you would find on the poems. These formal and thematic elements contribute to the distinctiveness of both classes in the dataset, offering a rich foundation for classification and comparison (see Table 3).

Table 3. Examples of rhythmic patterns in the dataset for both classes.

Literary Calaverita in Spanish	**Poem-Non Calaverita in Spanish**
La muerte llegó brincando (A	*Brillan al sol las hojas. Sus destellos* (A
A Robertita le trajo sandía (B	*llenan de claridades tus cabellos* (A
Una lotería (B	*de señora silvestre. Ven, permíteme* (B
Y unos dulces que quería. (B	*cubrirte y coronarte como ellos.* (A
Literary Calaverita English translation	**Poem-Non Calaverita English translation**
Death came jumping (A	The leaves shine in the sun. Their gleams (A)
To little Roberta it brought watermelon, (B	fill your hair with brightness (A)
A lottery ticket (B	of a wild lady. Come, allow me (B)
And some candies she wanted. (B	to cover and crown you like them. (A)

Since the dataset text has a defined structure and follows certain aspects in its creation, how they are written and how expressive or lengthy they are could play a key role in studying this task. Therefore, data such as text length, vocabulary size, and term frequency were considered. The vocabulary size of this dataset is relatively small compared to others. The entire dataset contains approximately 5,619 unique words, and when analyzed separately by class, the vocabulary size is as follows: Calaveritas: 3,398 unique words, and non-Calaveritas 2,935 unique words.

In terms of character length, as shown in Figure 1, the texts range between 60–200 characters. calaveritas are more concentrated in the 80–120-character range, while poems have a similar distribution but are slightly more agglomerated in the 120–140-character range.

This distribution of text length provides crucial information that should be considered during the fine-tuning and model-tweaking processes for classification. Understanding these patterns can aid in optimizing the model for better performance.

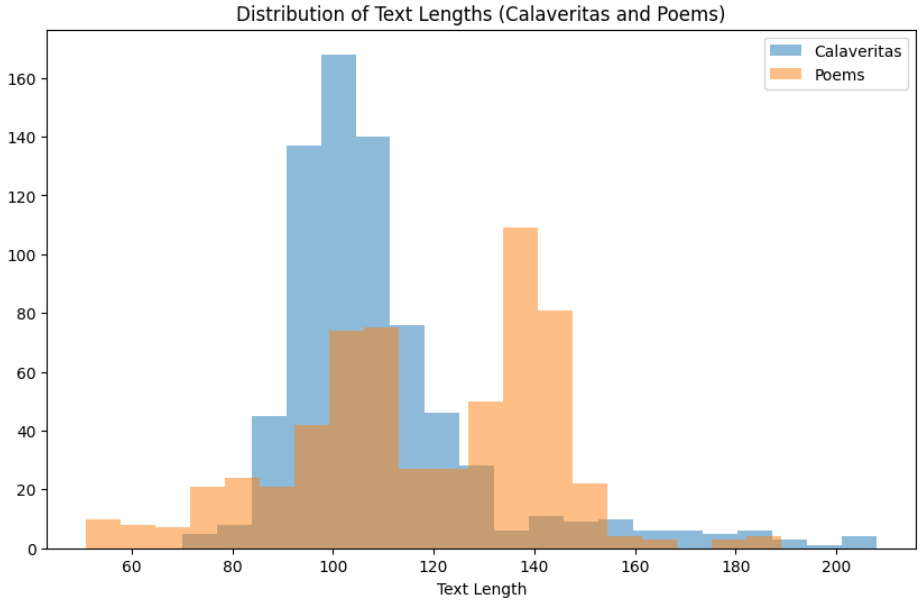

Fig 1. Character frequency by length string in each class of the dataset

The term frequency provides insight into the themes of the dataset, highlighting which words are used more frequently and offering a sense of the content being depicted. In this dataset, it is evident that certain words are more prominent within each class. For example, as shown in Figure 2, terms like death (muerte), skeleton (calaca), and graveyard (panteón) appear more frequently in the Calaveritas class. These words align with the death-related theme of the class, whereas such terms are less common in the poems (Non-Calaveritas) class (see Figure 3).

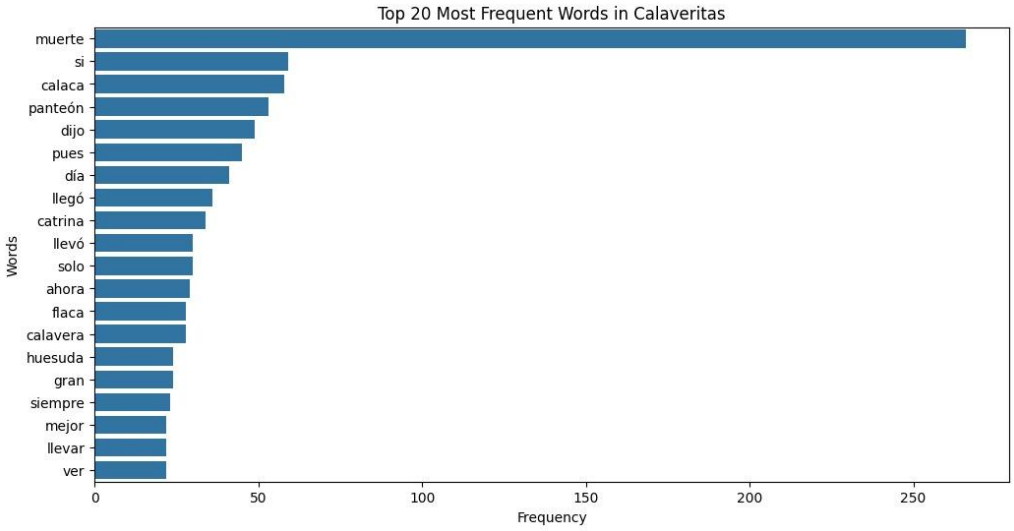

Fig 2. Most frequent terms in the Calaveritas class.

In the case of non-calaveritas, words related to love and life are more frequent. As shown in Figure 3, terms such as "love" (amor), "want" (quiero), "eyes" (ojos), and "life" (vida) appear more often. This class clearly contains terms commonly associated with poetic themes, focusing on life, and emotions. It is evident that the thematic focus of the two classes is quite different. While Calaveritas often take a darker approach with themes of death and the afterlife, poems reflect lighter and introspective topics. In many ways, the two classes could be seen as thematic opposites. Although they are not exclusive to each other.

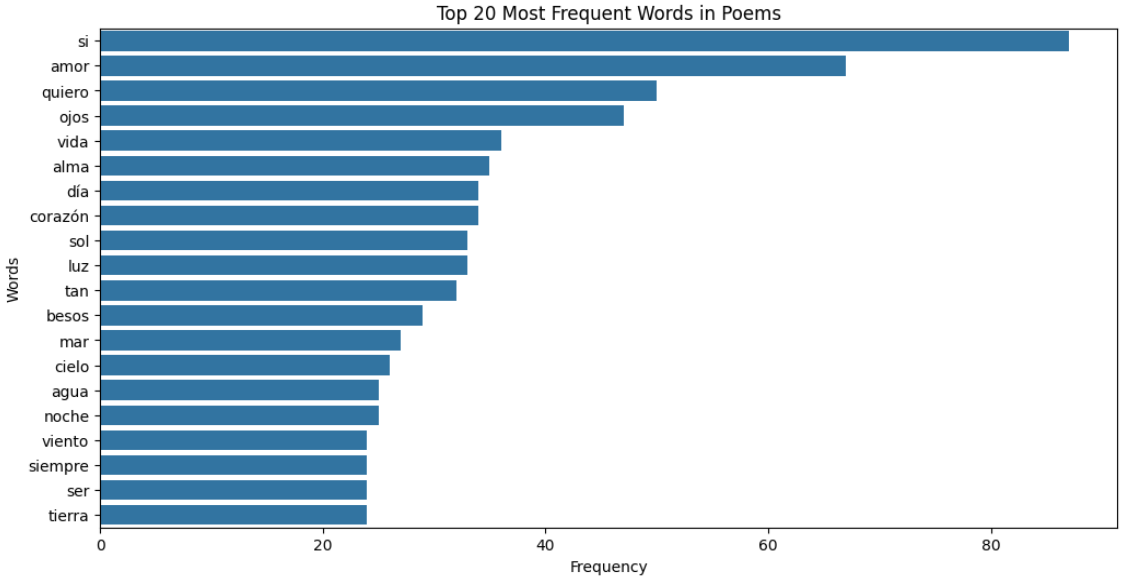

Fig 3. Most frequent terms in the non-calaveritas class.

5. Results

In this section, the results obtained for both the baseline and transformer models are presented. Additionally, an analysis of the data is included to provide a more comprehensive evaluation of the classification task. This analysis also aims to compare each class and examine how similar the calaveritas are to the poems.

5.1 Baseline

It is often observed that traditional machine learning methods perform well in classification tasks. In this case, the baseline set is intended to represent an overview of some useful and lightweight models. Six models and three embedding representations were used as a baseline for the task, resulting in eighteen classifiers being trained to compare.

In Table 4, the F1-score, recall, and precision weighted since it has for each combination of classifier and embedding are presented. Additionally, the values of the confusion matrix are included to show how each class is predicted by the models.

Table 4. Performance metrics (weighted) of baseline models using different embedding representations.

Modelo	**Text representation**	**F1-score**	**Recall**	**Precision**	**TP**	**FN**	**FP**	**TN**
SVM	TFIDF-2gram	0.8268	0.8296	0.8384	200	16	52	131
GaussianNB		0.8181	0.8195	0.8218	191	25	47	136
SVM	TFIDF-3gram	0.5991	0.6591	0.7908	216	0	136	47
GaussianNB		0.6780	0.6992	0.7349	199	17	103	80
SVM	TFIDF-BOW	0.9374	0.9352	0.9395	202	14	13	170
RandomForestClassifier		0.8446	0.9815	0.7413	212	4	74	109
GaussianNB		0.7024	1.0000	0.5414	216	0	183	0

Modelo	Text representation	F1-score	Recall	Precision	TP	FN	FP	TN
GradientBoostingClassifier		0.8777	0.8472	0.9104	183	33	18	165
AdaBoostClassifier		0.8725	0.9028	0.8442	195	21	36	147
DecisionTreeClassifier		0.8475	0.8102	0.8883	175	41	22	161
SVM	NNLM	0.8214	0.8519	0.7931	184	32	48	135
RandomForestClassifier		0.8161	0.8935	0.7510	193	23	64	119
GaussianNB		0.7891	0.8056	0.7733	174	42	51	132
GradientBoostingClassifier		0.7276	0.9028	0.6094	195	21	125	58
AdaBoostClassifier		0.7402	0.7454	0.7352	161	55	58	125
DecisionTreeClassifier		0.6791	0.6759	0.6822	146	70	68	115
SVM	USE	0.9379	0.9444	0.9315	204	12	15	168
RandomForestClassifier		0.9283	**0.9583**	0.9000	207	9	23	160
GaussianNB		**0.9404**	0.9491	**0.9318**	205	11	15	168
GradientBoostingClassifier		0.8387	0.9028	0.7831	195	21	54	129
AdaBoostClassifier		0.9266	0.9352	0.9182	202	14	18	165
DecisionTreeClassifier		0.8037	0.8148	0.7928	176	40	46	137

The setup for these models primarily involved using the Scikit-learn toolkit. Overall, the SVM was the top performer, with the highest scores. There was considerable variance between the metrics across the models, as some performed better than others. The models with the lowest scores were Gaussian Naïve Bayes and Decision Tree.

For the use of TF-IDF, a basic but effective representation, the SVM achieved a score of 93%, while the second-best models, Gradient Boosting and AdaBoost, achieved an F1-score of 87%.

Regarding embeddings, two types of representations were used for each model. The NNLM for Spanish performed the worst between the two embeddings, with its best F1-score being 82% with SVM and the lowest score overall at 68% with Decision Tree. For the use of the multilingual USE embedding, the best overall results were obtained, with a high score of 94% F1-score for Gaussian Naïve Bayes, which tied with SVM when rounded. This was followed by Random Forest and AdaBoost, both achieving 93%. The results of the baseline models were impressive and high enough to make improvement with other models more challenging.

In general, the classifier's ability to predict a class depends on the model. However, most models tended to generate fewer false negatives than false positives. This means that the models had more difficulty detecting the negative class (non-calaveritas) than the positive class (Calaveritas).

5.2 Transformers

After obtaining reliable results with the baseline, the second phase involves moving to the Transformer model paradigm. The most basic yet effective model for this task is the standard BERT. However, since this classification is in Spanish, the multilingual version was chosen as the first approach. The next model selected was BETO, a BERT variant trained specifically on Spanish data. BETO is widely used due to its strong capabilities for certain tasks, and in this case, BETO outperformed the multilingual BERT model as it is more focused on the Spanish language. However, this is not always the case for other tasks. Finally, another variation, DistilBETO, trained on Spanish corpora, was implemented as the final model.

The results achieved using the Transformer models surpassed those of the baseline. The best-performing model, BETO, scored an impressive 98% F1-score, which is comparable to DistilBETO if rounded. Although BERT-M had the lowest score among the

Transformer models at 97%, the difference is minimal, as it is only a matter of tenths. Given the high scores, the model predictions showed few errors, with minimal false positives and false negatives.

The top result in Table 5 is quite notable. The models show almost identical performance when it comes to classification, but they perform well even with the nuances of the *calaveritas* and poems texts. The best and worst examples from the top-performing model in each paradigm could provide insight into what the model struggles to understand.

Table 5. Performance metrics with weighted parameters of transformer models.

Model	F1-score	Recall	Precision	TP	FN	FP	TN
BERT Multilingual	0.9745	0.9768	0.9723	211	5	6	177
BETO	0.9837	0.9814	0.9860	212	4	3	180
DistilBETO	0.9791	0.9768	0.9813	211	5	4	179
Beto Masked	0.9908	0.9954	0.9862	215	1	3	180

6. Discussion

The similarities between both classes in terms of structure may seem apparent, as both have four verses and often feature a corresponding rhyme at the end of each verse. From a human perspective, these features might suggest that they belong to the same class. This leads us to draw some conclusions about the structure of these texts.

Some of the characteristics include number of syllables, verse structure, rhythm, and composition. Both classes share these traits as they are inherent to the selection criteria for each.

The Calaveritas class tends to address humorous dilemmas surrounding the Grim Reaper and death, often involving various actors. They create a satirical relationship between something somber like death and the humorous nature of the composition. On the other hand, the quartets tend to take a more grounded approach, focusing on themes of romanticism and melancholy.

Number of Syllables and Rhythm: The number of syllables in calaveritas varies significantly, as they have a freer structure compared to the poetic quartets. In traditional poetry, it is considered aesthetically pleasing to have specific metrics in the verses, such as octosyllabic, decasyllabic, dodecasyllabic, or even Alexandrian (14 syllables). These metrics are commonly used to maintain rhythm and beauty in the composition. Since these metrics are strict, the poem must follow this formula to some extent. In contrast, calaveritas do not adhere to these rules.

Sometimes, the rhyme may seem off because, in many amateur constructions, the rhyme scheme is often ignored in favor of simplicity, focusing only on rhymes at the end of each verse (either consonant or assonant), which can lack the structure of a well-formed poem. For example, in a well-constructed poem, the verses are usually well-formed, but this is not always the case with calaveritas. If the creator is experienced, they can craft compositions that rival a poem, even replicating its structures. In such cases, it would be harder to detect structural differences, but not in content as poems may focus on melancholic undertones in the representation of death's figure, perhaps leaning more toward memento mori or existential dilemmas.

Structure and Composition: This is the main difference between these two classes. In modern times, calaveritas have shifted toward a freer, less-defined verse composition. While they often consist of four verses, this is not a strict rule, as they can have more verses. However, for the purposes of this experiment, both classes were paired with four verses, independent of each other. Model Outputs: Our best performing model was BETO masked, which achieved a result of 99% in the F1-score. This is quite good, as the model is clearly performing above the 97% threshold. We could gain valuable insights by reviewing the worst-classified texts, particularly the false positives and false negatives.

Table 6 demonstrates that the model is highly confident when predicting that a sentence is not a calaverita. However, it does not show the same level of confidence for poems classified as Calaveritas. In this regard, we can infer that the model misclassifies some calaveritas because they lack certain defining characteristics. For example, while the text references death or bony (huesuda), one of the names representing death as a character, and contains four verses, it lacks rhyme and the proper structure typical of a calaverita.

The other example is a misclassified poem, as it also has a structure quite like that of a calaverita. If a death theme were inserted into it, it could suggest something related to confronting or pursuing death. Additionally, it has a four-verse structure and an ABAB rhyme scheme. In this case, the model is somewhat uncertain about whether the text possesses the characteristics of a calaverita.

Table 6. Examples and prediction probabilities for each class.

True Calaverita	Prediction Score	True Poem	Prediction Score
Como con la boca, camino con los pies pasa la huesuda y me caigo sin querer.	0.99615663 Not a Calaverita	Si no sabes a dónde vas, no entiendo porque te sigo, es mejor dejarme paz, no quiero pasear contigo.	0.62148285 Calaverita

To interpret the model's decision-making process beyond standard performance metrics, we examined the token-level distribution of attention coefficients in representative samples. Both high-confidence and high-uncertainty, enabling a comparison of attentional pattern structures and their relationship to the stability of the model's internal representations.

In this case, the attention coefficients, although numerically bounded, exhibit a more structured and coherent distribution. Functional and linking tokens (e.g., 'En', 'De', 'A') reach values close to 0.011–0.012, while semantically relevant terms within the narrative context, such as 'muerte' (≈0.006), 'Santa Catrina' (tokenized as 'Santa', 'Cat', '##ari', '##na'), and 'camión' (≈0.004–0.006), are consistently integrated into the global attention pattern. This type of organization, together with the presence of a clear diagonal in the attention map, is consistent with behavior in which the model leverages both local dependencies and distributed semantic associations, resulting in a more stable internal representation and a correct prediction (see Figure 6).

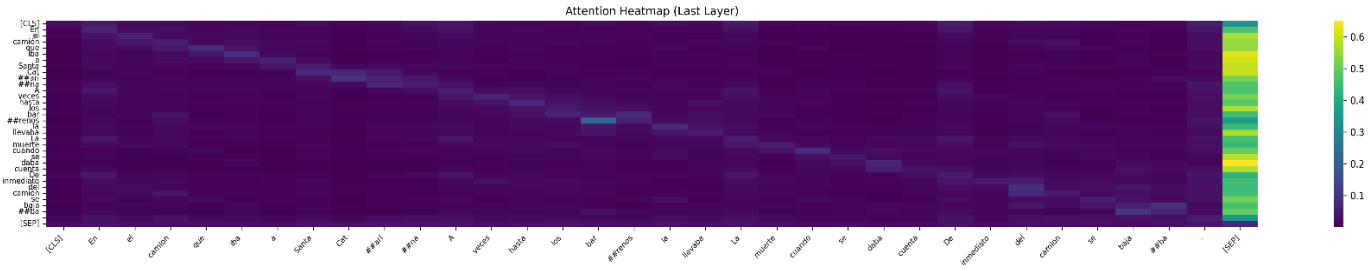

Fig. 4. Attention heatmap on best score Calaverita

The quantitative analysis shows that the attention coefficients assigned to individual tokens are low and relatively homogeneous, with maximum values around 0.041 ('Como') and most tokens concentrated in the approximate range of 0.009–0.017. Even words with potential semantic load, such as the fragmented form of 'huesuda' ('hue', '##su', '##da'), exhibit moderate values (≈0.009–0.015), without clearly standing out from the rest. Although attention weights do not constitute a direct measure of explanatory importance, this pattern is consistent with the visual interpretation of the heatmap, where attention appears diffused and lacks dominant focal points. Such behavior is typically observed in instances where the model fails to identify strongly discriminative lexical cues, which may be associated with higher predictive uncertainty or misclassification (see Figure 5).

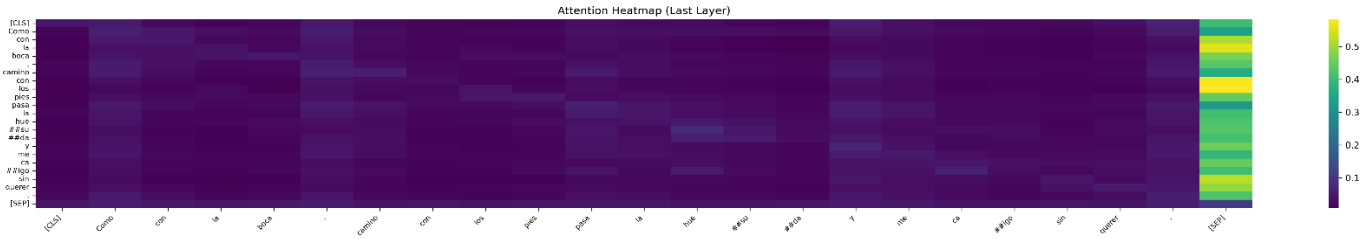

Fig. 5. Attention heatmap on worst score Calaverita

The comparison between the two examples reveals substantial differences at both visual and quantitative levels. In the Figure 5 case, attention coefficients are distributed in a near-uniform manner across tokens, with no single token exhibiting a clearly dominant weight. This pattern suggests that the model fails to identify specific lexical items or structural cues that are particularly informative for the final prediction. By contrast, in Figure 4, although individual attention values remain moderate, their distribution is more coherent and structured, aligning with syntactically and semantically meaningful relationships within the text. This contrast indicates that model performance is not driven by isolated peaks of attention, but rather by the model's ability to consistently organize attention across the sequence, integrating lexical information, narrative context, and higher-level structural patterns. These observations support the use of attention analysis as a qualitative diagnostic tool for distinguishing between reliable predictions and cases where the model's decision-making process may be less stable.

Based on the previous analysis of the model's worst results, an important question arises: Are poems and *calaveritas* distinct enough to be separable from each other? To address this, we compared both classes and observed an interesting phenomenon. *Calaveritas* occupy a similar textual space but are not like one another. In contrast, Poems show a high degree of similarity among themselves (see Figure 6).

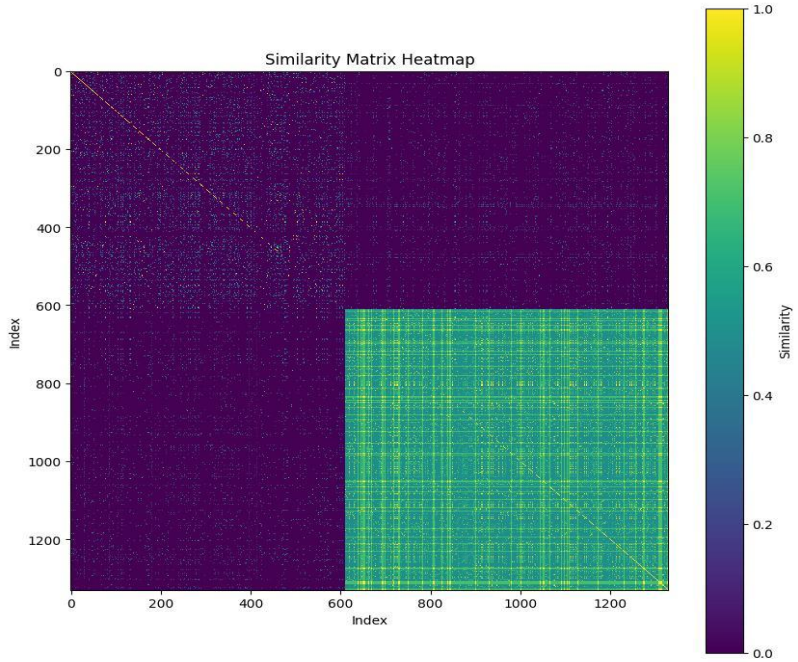

Fig. 6. Cosine similarity heatmap between classes.

This leads to two observations: one class (calaveritas) is more scattered, while the other (poems) is tightly clustered. This suggests that while the classes are indeed distinct from each other, the Poems maintain a stronger internal correlation. This may result from the strict structure that Poems typically follow, whereas calaveritas exhibit a much freer, almost arbitrary structure (see **Figure 7**). The analysis is more inclined to prioritize semantics over syntax for calaveritas. However, the structure, such as the number of syllables, plays a significant role in the proper separation of the classes.

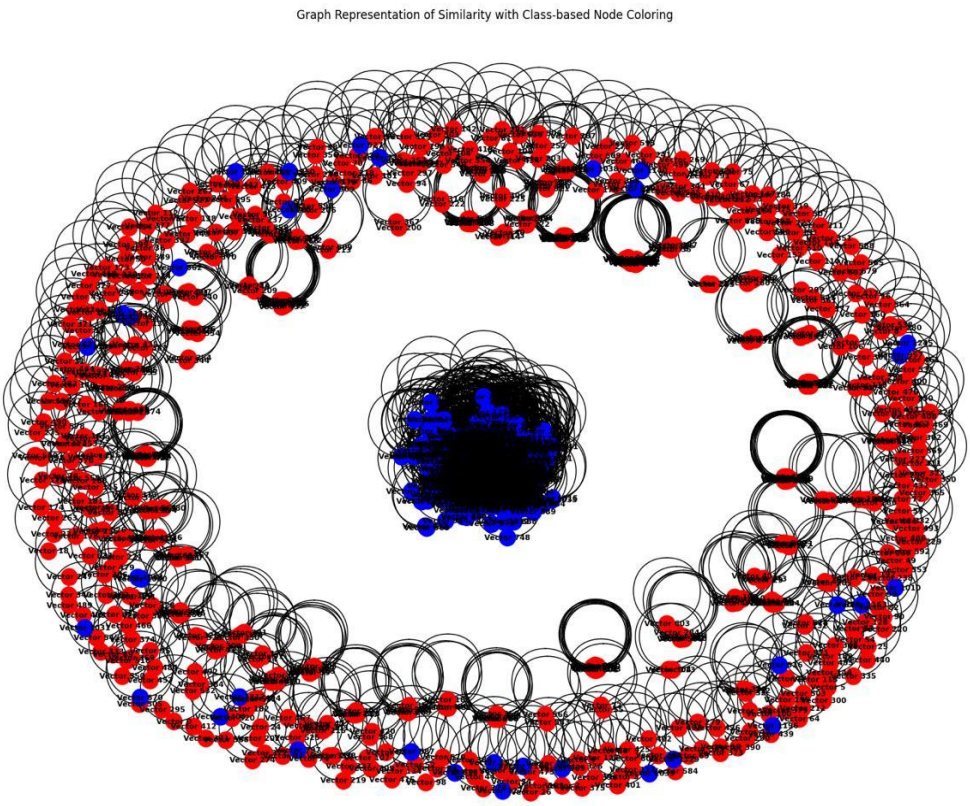

Fig. 7. Graph representation of cosine similarity: Red for Calaveritas, Blue for Non-calaveritas.

7. Conclusion

The task of classifying between Calaveritas and Non-calaveritas yielded strong results in both approaches. The baseline f1-score was relatively high at 94% using Naive Bayes with USE embeddings, setting a solid foundation. The objective was to build on this and achieve even better performance with more advanced models. This improvement threshold ensured that any progress would be meaningful when applying more complex algorithms. Despite the dataset's small size, it proved sufficient for fine-tuning, with BETO achieving an impressive f1-score of 98%.

This robust performance can likely be attributed to the characteristics of the dataset. However, Transformer models ultimately "put the final nail in the coffin," ensuring a high probability of correctly classifying Calaveritas. This score represents a key outcome of the analysis.

From the evaluation, it became evident that the two classes are quite distinct. A calaverita is much closer to a satirical or humorous composition than to a traditional poetic one. Another notable observation during the analysis was the difference in internal similarity between the two classes. poems exhibit greater internal consistency, clustering closely in the representational space. In contrast, calaveritas are more dispersed, reflecting their varied content and structure.

This distinction correlates with the strong performance of the models and suggests that they manage each class differently. For Calaveritas, the models seem to focus more on semantic content, particularly themes related to death, than on structural features. Conversely, for poems, the models rely more heavily on structural elements such as verse organization and rhyme schemes to determine the class.

The treatments applied in this study demonstrated robust results, with Transformer-based models clearly outperforming other approaches. Even multilingual models were effective at discerning nuanced differences in verse construction. However, Spanish-only models, such as BETO, achieved the best outcomes, although multilingual BERT was not far behind.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. En *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (pp. 265–283). USENIX Association. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Argüelles, J. D. (2003, noviembre 2). La adulteración de las calaveras. *La Jornada Semanal*, 452. https://www.jornada.com.mx/2003/11/02/sem-domingo.html

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

Bird, S., & Loper, E. (2004, julio). NLTK: The natural language toolkit. En *Proceedings of the ACL Interactive Poster and Demonstration Sessions* (pp. 214–217). Association for Computational Linguistics. https://aclanthology.org/P04-3031

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. En *Proceedings of the PML4DC Workshop at ICLR 2020*. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

Cañete, J., Donoso, S., Bravo-Márquez, F., Carvallo, A., & Araujo, V. (2022). ALBETO and DistilBETO: Lightweight Spanish language models. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2022)* (pp. 4291–4298). European Language Resources Association. https://aclanthology.org/2022.lrec-1.457/

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)* (pp. 169–174). Association for Computational Linguistics. https://aclanthology.org/D18-2029/

De la Rosa, J., Ros, S., Pérez, Á., Díaz, A., Hernández, L., De Sisto, M., & González-Blanco, E. (2021). PoetryLab as infrastructure for the analysis of Spanish poetry. En *Selected Papers from the CLARIN Annual Conference 2020* (Vol. 180, pp. 75–82). Linköping Electronic Conference Proceedings. https://doi.org/10.3384/ecp1809

Deng, S., Wang, G., Wang, H., & Chang, F. (2021). An artificial intelligence-driven Spanish poetry classification framework. *Big Data and Cognitive Computing, 7*(4), 183. https://doi.org/10.3390/bdcc7040183

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://aclanthology.org/N19-1423/

Inácio, M. L., & Oliveira, H. G. (2023). Attempting to recognize humor via one-class classification. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)* (Vol. 3496). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3496/huhu-paper4.pdf

Jhamtani, H., Mehta, S. V., Carbonell, J., & Berg-Kirkpatrick, T. (2019). Learning rhyming constraints using structured adversaries. En *Proceedings of EMNLP-IJCNLP 2019* (pp. 6025–6031). Association for Computational Linguistics. https://aclanthology.org/D19-1621/

Kesarwani, V., Inkpen, D., & Tănăsescu, C. (2021). #GraphPoem: Automatic classification of rhyme and diction in poetry. *Interférences Littéraires / Literaire Interferenties, 25*, 218–235.

Kolesnikova, O. (2025). Lexical function detection in Spanish collocations using transformer architecture. *Computación y Sistemas, 29*(2). https://doi.org/10.13053/cys-29-2-5620

Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. *Journal of Machine Learning Research, 23*(158), 1–6. https://jmlr.org/papers/v23/21-1259.html

Marchi, R. M. (2022). *Day of the Dead in the USA: The migration and transformation of a cultural phenomenon*. Rutgers University Press.

Marco, G., De la Rosa, J., Gonzalo, J., Ros, S., & González-Blanco, E. (2021). Automated metric analysis of Spanish poetry: Two complementary approaches. *IEEE Access, 9*, 51734–51746. https://doi.org/10.1109/ACCESS.2021.3069635

Palma Preciado, V. M., Palma Preciado, C., & Sidorov, G. (2024, octubre 20). *CalaveritasVsPOEMs* [Dataset]. Hugging Face. https://huggingface.co/datasets/vpalma/CalaveritasVsPOEMs

Panda, B., Sen, R. K., Dash, L., Panigrahi, C. R., & Pati, B. (2025). Machine learning approaches to sentiment analysis in social networks using political tweets. *Computación y Sistemas, 29*(4). https://doi.org/10.13053/cys-29-4-4996

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Romero-González, M. (2018). Día de Muertos – Day of the Dead: A multicultural tradition. En *Text Sets* (pp. 177–184). Brill. https://doi.org/10.1163/9789004368323_016

Tang, H., Kamei, S., & Morimoto, Y. (2023). Data augmentation methods for enhancing robustness in text classification tasks. *Algorithms, 16*(1), 59. https://doi.org/10.3390/a16010059

Winters, T., & Delobelle, P. (2020). Dutch humor detection by generating negative examples. En *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC 2020)*. https://doi.org/10.48550/arXiv.2010.13652

Winters, T., & Delobelle, P. (2021). Survival of the wittiest: Evolving satire with language models. En *Proceedings of the 12th International Conference on Computational Creativity (ICCC 2021)* (pp. 82–86).

Zhou, Q., Li, R., Xu, L., Nallanathan, A., Yang, J., & Fu, A. (2023). Towards explainable meta-learning for DDoS detection. *SN Computer Science, 5*, 115. https://doi.org/10.1007/s42979-023-02383-y