ISSN 2007-9737

LyricScraper: A Dataset of Spanish Song Lyrics Created via Web Scraping and Dual-labeling for LLM Classification

Tania Alcántara[1], Omar García-Vázquez[1], Mayte Hernandez[1], Hiram Calvo[1,*], Alan Desiderio[2]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

[2] Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica, Mexico City, Mexico

{talcantaram2020, ogarciav2024, mhernandezl2021, hcalvo}@cic.ipn.mx, afeliped1900@alumno.ipn.mx

Abstract. Songs represent a powerful means of expressing emotions through melody and lyrics. This study focuses on understanding and classifying emotions present in songs, ranging from positive and negative to neutral emotions. This classification and understanding would not be possible without data, which was gathered using a proprietary web scraping algorithm to collect lyrics data online. Subsequently, a pseudo-labeling approach based on BERT was employed to assign sentiment labels to these lyrics, leveraging BERT's ability to comprehend context and semantic relationships in language. This process enhanced the dataset's quality and contributed to the success of sentiment analysis in songs. The new dataset addressed challenges related to sentence length by providing examples of song lyrics of varying lengths, facilitating more effective model training. Additionally, data imbalance was addressed through careful sample selection, representing a wide range of emotions in songs. This new dataset underwent classification using large-scale language models, achieving promising results. The accuracy metric reached an impressive 97.66% for DistilBERT and 97.83% for the F1 metric, highlighting the effectiveness of this approach in song sentiment analysis. This study underscores the importance of understanding emotions in songs and offers practical solutions to enhance the capabilities of language models in this task.

Keywords. NLP, pseudo labeling, web scraping, deep learning.

1 Introduction

Someone in life hears or reads a lyric of a song, and some emotion travels through your mind. This emotion is associated with social elements, cultural aspects, and education. Each emotion corresponds to neurobiological functions, based on stimuli and responses [10]. Music has the ability to activate specific areas in our brains.

When talking about music, we often only think of the rhythm; however, it is not the only way to stimulate it, lyrics can do that too. Composers carefully select words that can evoke different things, and this is a substantial part of a musical piece. Based on this, it is particularly important to recognize the emotion labels.

Music Emotion Recognition (MER) is a computer method to extract and analyze music features [20]. MER generally works with different types of data, such as audio signals, lyric text, and even through EEG (Electroencephalogram). MER can work with various disciplines, such as psychology, audio signal processing, and Natural Language Processing (NLP) [20, 5].

The recognition and classification of emotion have been an important area in NLP [14]. This task has been widely studied in different jobs in the English language; however, it has been little

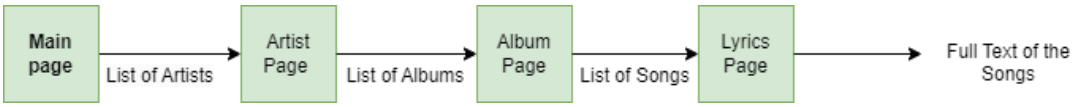

Fig. 1. Page structure

studied in others like Spanish. In addition to this, Spanish represents other challenges, such as variations by regions. To name a few, the Spanish in Mexico is not the same as in Chile or Spain. Additionally, the length and diversity of the text samples analyzed play a fundamental role in the accuracy of the categorization.

This article specifically focuses on polarity classification through Large Language Models (LLM) into three emotional categories: positive, negative, and neutral. Given the limited availability of information on song lyrics in Spanish, a dataset was generated using algorithms and web scraping Ùtechniques. Additionally, two labeling approaches, one manual and one automatic, were implemented to further enrich and diversify the dataset used in this analysis. The rest of this article is structured as follows:

In Section 2, the motivation behind this work and why it is relevant are addressed; Section 3 shows the theorical Framework behind the work; Section 4 analyzes previous work related to songs; Section 5 focuses on datasets and how the new dataset was built; Section 6 details how Large Language Models used it; Section 7 details the method to hyperparameter tuning finally, in Section 8, the results obtained in this work are presented and compared with previous research in the same field.

2 Motivation

Spanish is recognized as one of the most widely spoken languages globally, with approximately 21 countries officially using it, including Spain, Mexico, and various countries in Latin America. The linguistic richness of Spanish, particularly in Latin America, arises from the interaction with indigenous languages, making it valuable for Natural Language Processing (NLP) tasks when dealing with diverse datasets.

Music, being a powerful medium for conveying emotions, can exhibit various sentiments within a single song, adding complexity to emotion analysis, an area gaining prominence, especially with the advent of advanced models like GPT. Despite the rise in emotion analysis, there hasn't been significant exploration into language variations, especially within the Spanish-speaking world. Notably, differences in lexicon and linguistic nuances exist between Spain and Latin American regions, impacting emotional expression.

Emotion analysis typically involves identifying positive, negative, and neutral sentiments, with various approaches available for comparison. It's noteworthy that the datasets used in this context are in Spanish, emphasizing the importance of testing large language models trained on diverse datasets, including Spanish and multilingual ones.

Using web scraping to increase datasets are a very useful tool because it is a fast way to improve the metrics of the dataset, combining this technique with pseudo labeling. Large Language Models (LLMs) are pivotal in NLP, built on transformational structures to understand contextual cues and syntactic patterns. DistilBERT stands out as a significant contender, trained extensively on textual data to grasp contextual relationships within passages.

Its bidirectional approach, considering both left-to-right and right-to-left perspectives, enables it to cater to diverse languages and regions. DistilBERT, enhances speed and practicality with fewer parameters, while still discerning between English and other languages.

3 Theorical Framework

3.1 LLMs

LLMs, powered by the principles of transformer architectures (as described in [9]), are trained extensively to master language nuances

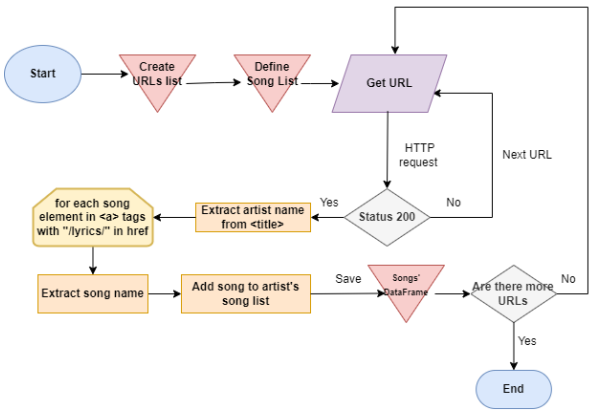

Fig. 2. General extraction process

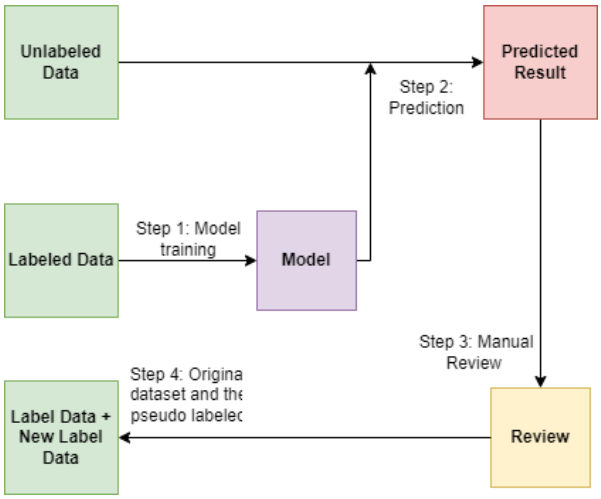

Fig. 3. Pseudo labeling diagram

by capturing contextual cues and syntactic patterns. Among language-specific models, BERT (Bidirectional Encoder Representations from Transformers) shines. This pre-trained model, developed by Google [1], ingests massive amounts of text data to learn contextual relationships between words within large passages and individual sentences.

Unlike other systems, BERT's unique bidirectional approach (analyzing from both left-to-right and right-to-left) makes it adaptable to diverse languages and regions. Multilingual BERT expands this by incorporating training in multiple languages [13]. Multilingual capabilities offer several benefits [13]:

- Resource Efficiency: One model tackles multiple languages, eliminating individual models for each.
- Language Representation: Models share and leverage knowledge across languages, enhancing performance in others.
- Performance: While language-dependent, multilingual models generally perform well.
- Mixed Set: Mental representations of various languages aid in learning unique patterns and structures.

Multilingual BERT: This impressive model pre-trains in 102 languages [3]. It analyzes sentence pairs to understand language structure and relationships, ultimately building an internal representation. DistilBERT: This "distilled" version of the multilingual BERT, also capable of language identification, has 6 layers, 768 dimensions, and 12 heads (totaling 134 million parameters). Notably, it runs twice as fast as BERT-base, making it a practical choice.

3.2 Pseudo-labeling

Pseudo-labeling is a machine learning technique used to train supervised learning models on unlabeled datasets. The main idea is to employ a pre-trained model to generate pseudo-labels for unlabeled data, and then use these pseudo-labels to train a new model [8].

How does it work? Pre-training, An initial model is trained using a dataset that already has labels; Pseudo-label generation, The pre-trained model is used to generate labels for the unlabeled data; Training the final model, The pseudo-labels are used to train the new model.

Web scraping, also known as web scraping, has become a fundamental tool for various applications in the digital world. This technique takes advantage of computer programs to extract valuable information from websites, emulating human browsing through the HTTP protocol or integrating a browser within an application. This approach offers multiple advantages. Firstly, it allows the efficient collection of large amounts of data, saving time and human resources.

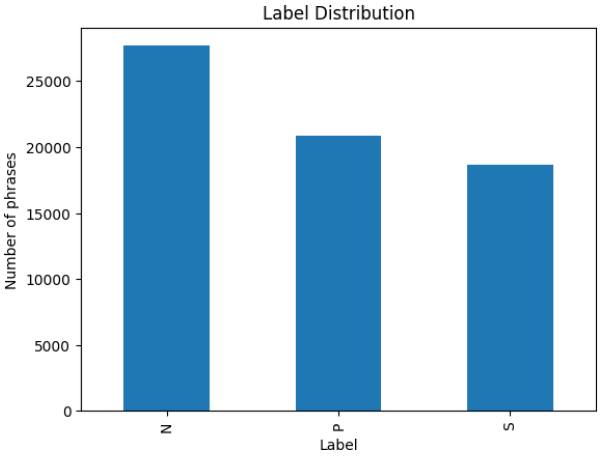

Fig. 4. Dataset distribution

Furthermore, it is highly scalable and capable of collecting information from thousands or even millions of web pages effortlessly. Its flexibility is also notable, as it can adapt to a wide range of websites and extract diverse data, from product prices to news and corporate information [17]. However, despite all the advantages it may have, sometimes data exchange is not always available [12].

3.2.1 Web Scraping and its Ethical Considerations

As mentioned, web scraping has multiple advantages, but there are also difficulties, especially some “scrapers” as they are commonly called, use it to duplicate the website, in addition to obtaining other types of income. Another point that is important to highlight is the protection of personal data [16], especially in scrappers where photographs are used and then charged for access to them.

Matthew Creed, in Kansas City, who obtained photographs of people in his town who had problems with the law and extorted them so that in order to delete their data, they had to provide money [11]. We can conclude that we need a “copy ethic” [17]. The copying of information should not be considered something completely “bad”, since in some cases it works to provide access to information, rather, special emphasis should be placed on how it is done and the purposes for which it is used, and how it fits on the topics of the philosophy of art, technology, law and ethics [11]. In this case and for the purposes of this research, web scrapping is used and considered a source of information for songs in Spanish, since the use of some APIS did not generate good results.

4 State of the Art

An early exploration into the study of emotions is presented in [6], where an investigation involves a collection of 1000 English songs. The research employs a categorical approach to classify emotions, including categories such as happy, sad, angry, and relaxed, as well as their complementary counterparts: not happy, not sad, not angry, and not relaxed. The main emphasis of their work revolves around applying a conventional method using Support Vector Machine (SVM) for the purpose of emotion classification.

The researchers in [4] investigate different methodologies utilizing the Edmonds Dance dataset, which is written in English. In their research, they utilize BERT to train the system on a song-by-song basis and attain significant outcomes, particularly in the classification of the emotion “anger”, where they achieve an accuracy metric of 0.88. In the domain of Emotion Classification, considerable efforts have been expended.

As highlighted in [18], a polar classification strategy was employed for a collection of Thai songs, focusing solely on the lyrics. The study proposed utilizing a lexicon alongside traditional machine learning methods. Regarding the song’s components (title, verse, chorus, pre-chorus, and bridge), emphasis was placed on solely using the chorus and verses as the corpus.

This selection was influenced by the likelihood that these two sections are most indicative of the song’s theme. On the other hand, the rise of social media has led to an increase in data and public discourse, unfortunately accompanied by offensive content. This issue is exacerbated by the diversity of languages on platforms and the various formats used to share offensive content (images, GIFs, videos, etc.).

La mentira que a ti te dijieron	N
Nos a separado	N
Te lo juro por mi madre santa	P
Que no soy casado	N
Esa gente que platica tanto	P
Pierde la amistad	P
La mentira que a ti te dijieron	N
Juro que no es verdad	N
Tu te fuiste no dijiste nada	N
Eso es sin razon	S
Me dejaste muy enamorado	P
A mi corazon	S
Algun dia quiera Dios que sepas	S
Toda la verdad	S
La mentira que a ti te dijieron	N
Juro que no es verdad	N
Tu te fuiste no dijiste nada.....	N
Me gustas por coqueta y altanera	P
me gusta tu mirada insinuante	P
me gusta todo lo que llevas puesta	P
y por tu bella forma de menearte	P
Me gusta ver cómo te vuela el pelo	P
cuando vas caminando por la calle	S
y tu aroma se queda por el viento	N
dándome valor para acompañarte	P
Me gustas por coqueta y altanera	P

Fig. 5. Dataset extract

Table 1. Modified hyperparameters proposed for the RoBERTa LLM

Hyperparameters	Values
Maximum sentence length	300
Minimum learning rate	0.00004
Batch size	24
Evaluation Batch size	24

The article “Ensemble of Multilingual Language Models with Pseudo Labeling for Offense Detection in Dravidian Languages” [4] proposes a multilingual ensemble-based model to identify offensive content targeting individuals or groups in resource-scarce Dravidian languages. The model handles data with a mix of code and scripts (e.g., Tamil and Latin).

The solution secured the first position in the Malayalam dataset and the fourth and fifth positions in Tamil and Kannada, respectively. There’s a big worry that explicit music can harm kids. Current systems used to label songs aren’t perfect, missing many tracks. But new research offers a clever solution: a machine learning model that uses song lyrics to tell if they’re explicit.

This “smart sorter” collects lyrics from the internet, analyzes them, and sorts them into “clean” or “explicit” piles. It’s surprisingly good at its job, catching 96% of explicit songs and almost never making mistakes.

The study also found that nearly 40% of songs are explicit, especially in Hip-Hop/Rap. This new tool not only helps protect kids but also gives us a clearer picture of how common explicit content is in different music styles [19]. The article “Pseudoetiquetado para el Análisis de Polaridad en Tweets: Un Primer Acercamiento” [7] introduces a pseudo-labeling technique for sentiment analysis in tweets.

This method involves creating a pseudo-labeled dataset from an original dataset with known sentiment labels. The pseudo-labeled dataset is then used to train a machine learning model for sentiment classification in tweets.

The authors experimented with two pseudo-labeling techniques: Similarity-based Pseudo-labeling: This technique assigns sentiment labels to unlabeled tweets based on their similarity to tweets from the original dataset with known sentiment labels. Machine Learning-based Pseudo-labeling: This technique employs a machine learning model trained on the original dataset to assign sentiment labels to unlabeled tweets.

The study found that the machine learning-based pseudo-labeling technique was more effective than the similarity-based approach. Additionally, using a pseudo-labeled dataset was observed to enhance the performance of a machine learning model in sentiment classification for tweets. Finally, in [2] this study delves into the potential of large language models (LLMs) like BERT and RoBERTa to classify emotions in songs.

Researchers faced hurdles with existing LLMs due to inconsistent sentence lengths and uneven distribution of emotions within song datasets. To address this, they crafted a new dataset featuring diverse sentence structures and balanced representation of various emotions. Utilizing this novel approach, the models achieved an impressive accuracy of 96.34%, showcasing the effectiveness of LLMs in analyzing song sentiment.

Table 2. Modified hyperparameters proposed for the DistilBERT LLM

Hyperparameters	Values
Maximum sentence length	300
Minimum learning rate	0.00004
Batch size	48
Evaluation Batch size	48

Table 3. BERT, Roberta and DestilBERT results with text of spanish songs for the accuracy metric and F1 metric

Model	Accuracy Results	F1 Results
RoBERTa	97.09%	97.32%
DistilBERT	97.66%	97.83%

This research not only deepens our understanding of emotional expression in music, but also proposes a valuable dataset design for future LLM-based song sentiment analysis. Moreover, it opens exciting doors for further exploration of LLMs in music analysis and unlocking the mysteries of human emotions through song.

5 LyricScraper

Data Usage Statement: The data used in this study were obtained exclusively for academic and research purposes. No profit was made from the extraction or use of this data.

Any further use of the data beyond the scope of this study will require appropriate consent and authorization from the data owners. The method called LyricScraper by us has steps which will be described, each subsection corresponds to a step.

5.1 Azlyrics Analysis

The page follows a hierarchy from the list of artists to the specific lyrics, described as follows:

1. **Artist List A-Z:** The main page displays a list of artists organized alphabetically, from letter A to Z. Users can click on the initial of the artist's name to access the list of artists starting with that letter.
2. **Artist Selection:** Clicking on a specific artist leads to a new page that displays information related to that particular artist.
3. **Album Listing:** On the artist's page, there is a list of albums published by that artist. Each album usually has its own link or section.
4. **Song Titles by Album:** Within each album, a list of songs associated with that particular album is shown. Each song is generally presented with its title.
5. **Access to Song Lyrics:** Clicking on the title of a specific song allows users to access the complete lyrics of that song on a separate page.

When entering each section, the URL for a specific song is constructed as follows[1].

5.2 URL Construction and Song Extract

URL Construction Part of a base[2]. With this in mind, work began on the program for constructing the URLs using a .csv file, where the data completing this URL includes the artist and the song title.

When creating the algorithm to extract the artist's name and the song title from the AZlyrics page, a difference was found in constructing the URLs, as songs after 2021 are built with the artist's name and the song title, but with an important change: the accented vowel is incorporated into the URL. Here is an example of this change.

– Example 1: URL of a song prior to 2021[3]. In this example, it can be observed how the accented vowel "ó" has been removed from the word "adiós."

– Example 2: URL of a song after 2020[4]

In this example, it is observed that the characteristic of removing the accented vowel has been eliminated, primarily due to the type of encoding used. The encoding in these types of files is crucial because we are working with text;

[1] www.azlyrics.com/lyrics/artist/songtitle.html
[2] www.azlyrics.com/lyrics/
[3] www.azlyrics.com/lyrics/christiannodal/adisamor.html
[4] www.azlyrics.com/lyrics/carinleon/enpeligrodeextincion.html

Table 4. Comparative table of the previous experiments and results obtained

	Alcántara T. et al. [2]		Our Work	
Model	**Acc**	**F1**	**Acc**	**F1**
RoBERTa	96.34%	91.94%	97.09%	97.32%
DistilBERT	95.88%	91.37	97.66 %	97.83%

it is necessary to choose an appropriate encoding, in this case, UTF-8. The song lyrics were stored in an Excel file, as it is useful for labeling the songs. The song lyrics extraction process begins with the alphabetical organization of the website, where the artist is selected based on the first letter of their name.

Subsequently, a web scraping process is carried out to extract the song titles of said artist, as well as the artist's name. The information is saved in a .csv file with two columns, one for the author's name and another for the song name. This .csv file is then used to generate the corresponding URLs that will be used to collect the lyrics of each song.

The implementation of this file is integrated into the Python code, resulting in the creation of a .xlsx file that contains the lyrics of the songs along with the name of each song. This process facilitates the management and subsequent analysis of the extracted information. This process is visualized in Figure 2.

6 Labeling

Pseudo-Labeling is a Key Technique in Artificial Intelligence and Machine Learning Pseudo-labeling is a fundamental procedure in the field of artificial intelligence and machine learning, designed to address the scarcity of labeled data in datasets. This process involves using two main datasets: one previously labeled and another unlabeled. Figure 3 visually presents the key elements and the flow associated with pseudo-labeling.

6.1 Labeled Data

The first stage of the process employs a dataset that has been accurately and reliably labeled (referred to as "labeled data").

This set serves as a knowledge base on which pseudo-labeling will be based. In our case, we used the dataset called "Textos de canciones en español mexicano" [1]. This dataset consists of Spanish song lyrics, with 5958 instances, classified into three polarities: positive, negative, and neutral.

The labeling was performed manually by a group of 3 men and 3 women, between the ages of 18 and 27, assigning the labels through the average of their evaluations.

The second set, of the same type for correct pseudo-labeling, must lack labels (referred to as "unlabeled data"). This set contains 61,323 data instances, extracted using the so-called LyricScraper 5.

6.2 Generating Pseudo-labels

The second stage involves applying the trained model to the unlabeled dataset. As the model makes predictions on this set, "pseudo-labels" are generated for the unlabeled instances.

These pseudo-labels, which are model predictions, are used as a substitute for real labels in the unlabeled set. The model used for pseudo-labeling is BERT, selected due to its superior performance as reported in [2].

6.3 Predicted Result

The third stage (labeled "Predicted Result" in light pink) represents the result of the model's predictions on the unlabeled dataset.

6.4 Manual Review

The fourth stage (labeled "Review" in yellow) requires a manual review of the labels in the pseudo-labeled set. Despite the usefulness of pseudo-labeling, this method has limitations and risks.

Manual review mitigates possible errors in predictions and provides confidence in identifying and filtering low-confidence instances. Additionally, this review addresses conceptual deviations, where the model might incorrectly interpret the context or semantics.

6.5 Final Labeled Data

In the final stage (labeled “Label Data + New Label Data” in green), the datasets are combined to create the final dataset. The final dataset comprises a total of 67,280 song lyrics, segmented into verses. The ultimate class distribution consists of 27,678 negative (N) instances, 20,900 positive (P) instances, and 18,701 neutral (S) instances. The figure 4 shows this information. To exemplify the type of data at hand, figure 5 provides a visualization of an excerpt from the dataset.

7 Classification Methodology

For classification, it was decided to use the 80/20 split data set for training and validation, with a maximum sentence length of 300 characters. For greater consistency and comparison, the RoBERTa-base and DistilBERT-base-multilingual-cased models were chosen, which were used in the previous experiments.

Also, these models were chosen based on their ability to process Spanish texts better, since they were trained in a ”Multilingual” way. To avoid overfitting the data, BERT was not used for classification, as it was the LLM used for pseudo labeling. To configure **RoBERTa** and solve challenges such as local minima and small data sets, variable learning rates and small batches were used.

A dynamic learning rate and batch size of 24 were applied, aligning with both the literature and the capabilities of the team. Details are summarized in Table 1. In the case of **DistilBERT-base-multilingual** this lightweight version of BERT was chosen for its efficiency and capitalization support. Delivers accurate predictions with lower resource demands.

The table 2 presents the custom hyperparameters for this job. For hyperparameter optimization, grid search was used for each model. The hyperparameter exploration began with the base configuration and progressed to a grid search based on sets recommended by the model authors. Eighteen training runs were performed for each combination.

Additionally, well-performing configurations from previous studies were included in BERT [3], RoBERTa [9] and DistilBERT [15].

8 Results and Comparisons

The results obtained from the evaluation of the data sets using the RoBERTa and DistilBERT models are presented in the table 3, where the precision and F1 values are shown for validation. These results reflect a meticulous optimization process, which included hyperparameter tuning and word size refinement for each model.

Crucially, the high accuracy achieved is not only the result of fixing previously identified bugs, but also of the considerable increase in the number of examples provided compared to previous iterations, supported by extensive additional manual review.

Despite the achievements obtained, it is important to recognize that this evaluation would be more complete with an exhaustive analysis with other methods and different sets. The currently presented comparison focuses on the data set similar to the one used and the models same as the ones we chose.

It is important to highlight that the texts show an inherent subjectivity to the texts and the complexity of the task for the models, which increase the need for this comparative analysis for a more complete and robust evaluation. It is important to note that the comparison is not 100% direct, as the state of the art does not use the same datasets. However, as mentioned, the work cited in reference [2] did use a part of the same dataset. Therefore, in Table 4, a comparison of the results.

9 Conclusions

This study has focused on analyzing and automating the categorization of song lyrics across different variants of the Spanish language. To achieve this objective, models like RoBERTa and DistilBERT Multilingual were utilized, sharing notable similarities in their approach and functionality.

The findings indicate that employing web scraping and pseudo labeling enhances all metrics. By exposing the model to more examples of the problem, it can generalize better and make more accurate predictions.

It's evident that models tailored with specific Spanish language datasets, such as RoBERTa, achieve significant and competitive results compared to multilingual models like BERT and DistilBERT.

Another aspect was the exploration of hyperparameters, which, although quite similar among themselves, produced satisfactory results and should be considered for future implementations of similar projects. It could be interesting to use this labeled dataset as training for some traditional and novel classifiers to be a benchmark in the field of Spanish lyrics polarity classification.

Speaking of the limitations of the approach, it can be concluded that although the labeling was done automatically and although it also went through an exhaustive review process, there is always a risk that the reviewer's personal experiences may affect the polarity of the labels.

Finally, it should be clarified that only musical genres that in a certain way respect the grammatical rules of Spanish and try to use neutral language with few idioms were addressed.

In order to include genres such as reggaeton or rap in Spanish, a deeper analysis of the phrases would have to be included in order to homologate them. In conclusion, this study has provided insights into the categorization of Spanish song lyrics through various model implementations, paving the way for future research to refine and improve this technique.

Acknowledgments

The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20240610) and the Mexican Government (CONAHCyT, SNI).

References

1. **Alcántara, T., Desiderio, A., Garcia-Vazquez, O., Calvo, H. (2023).** Corpus "textos de canciones en español mexicano v1". Laboratorio de Ciencias Cognitivas Computacionales, Centro de Investigación en Computación.
2. **Alcántara, T., García-Vázquez, O., Calvo, H., Sidorov, G. (2024).** Classification of songs in spanish with LLMs: An analysis of the construction of a dataset, through classification. Special Issue of the Journal of Intelligent and Fuzzy Systems.
3. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. Vol. 1, pp. 4171–4186. DOI: 10 .18653/v1/N19-1423.
4. **Edmonds, D., Sedoc, J. (2021).** Multi-emotion classification for song lyrics. Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 221–235.
5. **Han, D., Kong, Y., Jiayi, H., Wang, G. (2022).** A survey of music emotion recognition. Frontiers of Computer Science, Vol. 16. DOI: 10.1007/s11704-021-0569-4.
6. **He, H., Jin, J., Xiong, Y., Chen, B., Sun, W., Zhao, L. (2008).** Language feature mining for music emotion classification via supervised learning from lyrics. Advances in Computation and Intelligence: Third International Symposium, pp. 426–435. DOI: 10.1007/978-3-540-92137-0_47.
7. **Jimenez, D., Cardoso-Moreno, M. A., Macias, C., Calvo, H. (2023).** Pseudoetiquetado para el análisis de polaridad en tuits: Un primer acercamiento. Research in Computing Science, Vol. 152, No. 8, pp. 289–299.
8. **Lee, D. H. (2013).** Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. International

Conference on Machine Learning Workshop: Challenges in Representation Learning.

9. **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019).** RoBERTa: A robustly optimized BERT pretraining approach. International Conference on Learning Representations, pp. 1–15. DOI: 10.48550/arXiv.1907.11692.
10. **Maureira-Cid, F. (2010).** Ser humano: Emociones y lenguaje. Revista Electrónica de Psicología Iztacala, Vol. 11, No. 2.
11. **Pagallo, U. (2011).** The trouble with digital copies: A short km phenomenology. Ethical Issues and Social Dilemmas in Knowledge Management: Organizational Innovation, IGI Global, pp. 97–112.
12. **Pagallo, U., Ciani, J. (2023).** Anatomy of web data scraping: Ethics, standards, and the troubles of the law. DOI: 10.2139/ssrn.4707651.
13. **Pires, T., Schlinger, E., Garrette, D. (2019).** How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493.
14. **Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L. P. (2017).** Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 873–883. DOI: 10.18653/v1/P17-1081.
15. **Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019).** DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv.
16. **Sellars, A. (2018).** Twenty years of web scraping and the computer fraud and abuse act. BU Journal of Science and Technology Law, Vol. 24, pp. 372.
17. **Sitelabs (2017).** Web scraping: Introducción y herramientas. sitelabs.es/web-scraping-introduccion-y-herramientas/.
18. **Srinilta, C., Sunhem, W., Tungjitnob, S., Thasanthiah, S. (2017).** Lyric-based sentiment polarity classification of thai songs. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1.
19. **Tula, D., Potluri, P., Ms, S., Doddapaneni, S., Sahu, P., Sukumaran, R., Patwa, P. (2021).** Bitions @ DravidianLangTech-EACL2021: Ensemble of multilingual language models with pseudo labeling for offence detection in dravidian languages. Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 291–299.
20. **Yang, X., Dong, Y., Li, J. (2017).** Review of data features-based music emotion recognition methods. Multimedia Systems, Vol. 24, pp. 365–389. DOI: 10.1007/s00530-017-0559-4.

Article received on 15/04/2024; accepted on 22/06/2024.
**Corresponding author is Hiram Calvo.*