# Comparative Evaluation and Statistical Analysis of Sentiment Analysis Models for Spanish

Nuvia Beltrán[1,*], Irene Vásquez[2], Darwin Pow-Chon-Long[2], Teresa Samaniego-Cobo[1] and Luis Rodríguez[1]

[1]Universidad Agraria del Ecuador, Milagro, Ecuador
[2]Universidad Agraria del Ecuador, Guayaquil, Ecuador

**Abstract**

Sentiment analysis is a central task in Natural Language Processing (NLP); therefore, this study conducts a comparative evaluation using a paired statistical protocol of four models—PySentimiento, Transformers (BETO), TextBlob and VADER—on twenty Spanish-language datasets from diverse domains: social media, movie reviews, e-commerce products, and tourism. Since the four models were evaluated on the same twenty datasets, the results were subjected to the nonparametric Friedman test and to paired post hoc comparisons using the Nemenyi and Wilcoxon tests with Holm correction, supplemented by the calculation of effect size using paired Cohen's d and rangepair biserial correlation. Transformers achieves the highest average performance, followed by VADER and PySentimiento. The statistical analysis confirms that the differences between model pairs are statistically significant, with paired effect sizes reflecting the systematic separation between qualitatively distinct architectures. Practical implications for the development of Spanish NLP systems are discussed, and operational criteria for model selection based on the application context are proposed.

**Keywords**

Sentiment analysis, Spanish NLP, statistical evaluation

## 1. Introduction

Sentiment analysis aims to identify, extract, and quantify the emotional charge contained in subjective texts [1, 2]. With the exponential growth of user-generated content on digital platforms, the demand for systems capable of processing large volumes of text in languages other than English has intensified considerably [3]. Spanish, with more than 500 million native speakers, is one of the languages with the greatest global reach; however, it remains a language with comparatively scarce NLP resources compared to English [4, 5, 6].

A recurring problem in the literature on NLP system evaluation is the reporting of point metrics without considering the statistical significance of the observed differences; this practice can lead to erroneous conclusions about the superiority of one system over another [7, 8]. This methodological gap is particularly relevant in the context of sentiment analysis in Spanish, where the diversity of domains, registers, and dialectal varieties introduces significant variability in model performance [9].

Although there have been previous comparisons of sentiment analysis tools in Spanish—notably within the TASS competition framework and in multilingual benchmarks focused on social media—these studies rarely combine multiple architectures evaluated under a statistical protocol that demonstrates the significance of the differences observed between models.

This study addresses these limitations through an experimental evaluation of four representative models applied to twenty Spanish-language datasets. In addition to standard metrics, nonparametric statistical tests (Friedman and Nemenyi/Wilcoxon with Holm's correction) are applied, and the effect size (Cohen's paired d) is calculated to determine whether the observed differences are statistically significant. Thus, this study seeks to answer two research questions:

*IWSAI 2026: 1st International Workshop on Systems, Applications, Artificial Intelligence and Innovation, September 16–18, 2026, Piura, Peru*
*Corresponding author.
nbeltran@uagraria.edu.ec (N. Beltrán)

- (PI1) Which model offers the best performance in sentiment analysis in Spanish?
- (PI2) Are the differences in performance between models statistically significant?

## 2. Methodology

### 2.1. Datasets

Twenty Spanish-language datasets from diverse domains were used, sourced from Kaggle, GitHub, and Hugging Face. All corpora were preprocessed through tokenization, orthographic normalization, removal of special characters, and removal of stop words [10, 11, 12, 13, 14].

### 2.2. Dataset Inclusion and Exclusion Criteria

We included datasets that met only the following criteria:

- Language: Spanish.
- Domain: social media, movie reviews, e-commerce products, and tourism.
- File types: .xls, .json, .csv.

### 2.3. Evaluated Models

A software program [15] was developed with four sentiment analysis models configured as follows:

1. TextBlob v0.17.1: applied after automatic translation into English using the Google Translate API.
2. VADER v3.3.2: applied to the texts translated into English.
3. PySentimiento v0.7.0: applied directly to Spanish text.
4. Transformers: BETO-cased model (bert-base-spanish-wwm-cased) fine-tuned for sentiment classification into three classes using supervised fine-tuning with a learning rate of $210^{-5}$, a batch size of 32, 3 epochs, and early stopping with a patience of 2 epochs on the validation set. For each corpus, a stratified split of 70/15/15 (training/validation/test) was used.
   It is important to note two limitations of this experimental study: since TextBlob and VADER are applied to text automatically translated into English— —translation was performed after cleaning and tokenization to enable analysis, which could reduce processing capacity. Another limitation stems from the characteristics of each tool: BETO uses supervised fine-tuning with a corpus, Pysentimiento is pre-trained without fine-tuning, and TextBlob and VADER use unsupervised lexicons; therefore, the comparisons should be interpreted as an evaluation of approaches under realistic usage conditions and not as a controlled comparison of a single architectural variable.

### 2.4. Metrics and Statistical Analysis

For the quantitative evaluation, four standard metrics were used in three-class multiclass classification (positive, negative, neutral): accuracy, macro precision, macro recall, and macro F1-score [16, 17, 18]. These metrics were calculated independently for each class c ∈ positive, negative, neutral and averaged in their macro versions, giving equal weight to each class regardless of its frequency. Next, each metric is formally defined, where VP_c, FP_c, FN_c, and VN_c denote the true positives, false positives, false negatives, and true negatives of class c, respectively, and $C$ is the total number of classes ($C = 3$).

**Accuracy:** The proportion of correctly classified instances out of the total [19]:

$$Accuracy = \frac{\sum VP_c}{N} \tag{1}$$

where $N$ is the total number of instances in the evaluation set.

**Macro-Precision:** The average of the ratio of instances correctly predicted as class c to all instances predicted as c:

$$Precision_{macro} = \frac{1}{C}\sum_{c=1} C\frac{VP_c}{VP_c + FP_c} \tag{2}$$

**Macro-Recall:** The average of the ratio of correctly retrieved instances of class c to all actual instances of c:

$$Recall_{macro} = \frac{1}{C}\sum_{c=1} C\frac{VP_c}{VP_c + FP_c} \tag{3}$$

**Macro-F1-Score:** The harmonic mean of Macro-Precision and Macro-Recall, which penalizes imbalances between the two metrics:

$$F1_{macro} = \frac{1}{C}\sum_{c=1} C\frac{2.VP_c}{2.VP_c + FP_c + FN_c} \tag{4}$$

The statistical analysis followed the protocol below. First, the distribution of the F1-Score values was verified using the Shapiro-Wilk test; the results confirmed deviations from normality in the four evaluated models ($p < 0.05$), justifying the use of nonparametric tests. Second, the Friedman ($\chi^2$) test was applied to the 100 F1-Score values (20 per model) to test the null hypothesis of equal distributions, whose statistic is defined as:

$$\chi^2 = \frac{12}{N.k(k+1)}\sum_{i=1}^{k} R_i^2 - 3N(k+1) \tag{5}$$

where $N$ is the number of blocks (datasets), $k$ is the number of treatments (models), and $R_i$ is the sum of ranks for treatment $i$ across the $N$ blocks. With $N = 20$ and $k = 4$, this formula yields $\chi^2 = 57.90$, with $df = 3$. Multiple post-hoc comparisons were performed using the Nemenyi test (Friedman's canonical post-hoc test) and, complementarily, using the paired Wilcoxon test with Holm correction ($\alpha = 0.05$, applied sequentially) on the ten possible paired comparisons.

The effect size was calculated using Cohen's d for each paired comparison:

$$d_z = \frac{\bar{d}}{s_d} \tag{6}$$

# 3. Results

## 3.1. Descriptive Statistics

Table 1 presents the consolidated descriptive statistics for the four models evaluated across the twenty datasets. The values are sorted from highest to lowest average F1-Score.

The selection of the macro F1-score as the primary evaluation metric is based on three complementary methodological criteria. First, the problem addressed is a three-class multiclass classification (positive, negative, and neutral) with class distributions that may be imbalanced depending on the corpus domain—a context in which accuracy is an unreliable metric because a classifier that labels all instances as the majority class can achieve high accuracy scores without having learned any discriminative structure.

The macro F1-score addresses this limitation by calculating the harmonic mean of precision and recall independently for each class and averaging them without weighting by frequency, thereby giving equal importance to all three categories regardless of their representation in the corpus [20]. Furthermore, the macro F1-score is the dominant benchmark metric in the main Spanish sentiment analysis competitions—TASS [21, 22] —which ensures direct comparability of the results obtained with those reported in the literature on the subject [9]. Finally, since the Friedman test and Cohen's d are paired tests based on the distribution of observed values per model and per corpus, it is methodologically consistent to apply them to a single summary metric per corpus that captures the classifier's overall

**Table 1**
Descriptive statistics by model ($n = 20$ datasets). $\mu$ = mean, $\sigma$ = standard deviation.

| Model | N | F1_Mean | F1_DE | Min_F1 | Max_F1 | IC95_low | IC95_upper | CV_% |
|---|---|---|---|---|---|---|---|---|
| Transformers (BETO) | 20 | 0.8017 | 0.0087 | 0.7896 | 0.8163 | 0.7977 | 0.8058 | 10.802 |
| PySentiment | 20 | 0.7827 | 0.0083 | 0.7715 | 0.7959 | 0.7788 | 0.7866 | 10.565 |
| VADER | 20 | 0.7500 | 0.0093 | 0.7363 | 0.7658 | 0.7457 | 0.7543 | 12.353 |
| TextBlob | 20 | 0.6899 | 0.0105 | 0.6700 | 0.7063 | 0.6850 | 0.6948 | 15.158 |

performance in a balanced manner; the macro F1-score meets this condition by synthesizing precision and recall into a single scalar value that is comparable across models with qualitatively different architectures and learning regimes [21].

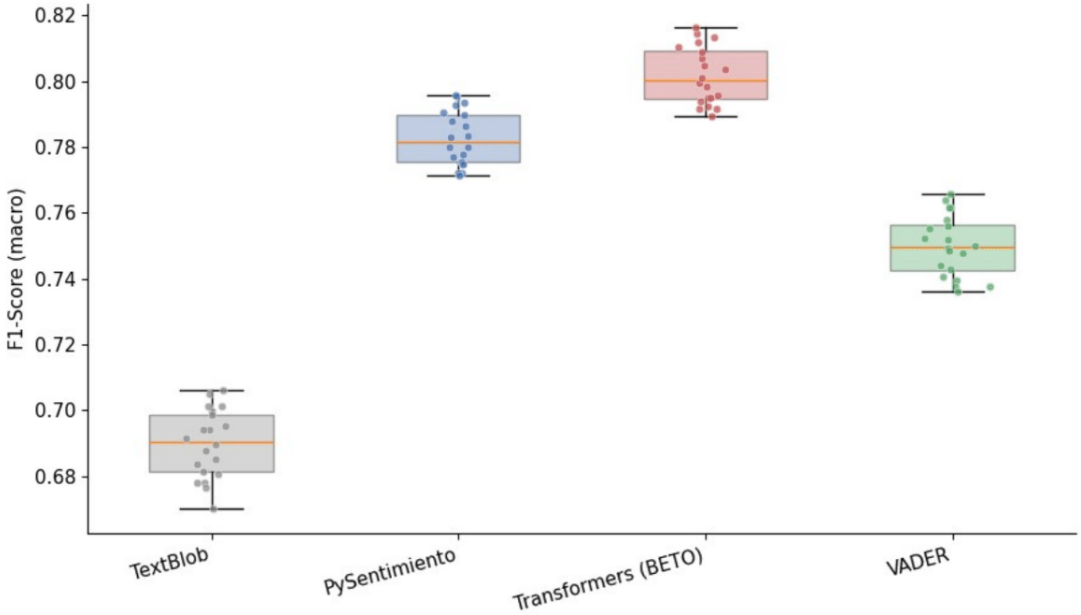


**Figure 1:** Distribution of the F1-Score by model ($n = 20$ datasets).

Transformers (BETO) achieves the highest average F1-Score ($\mu = 80.17\%$), followed by PySentimiento ($\mu = 78.27\%$), VADER ($\mu = 75.00\%$), and TextBlob ($\mu = 68.99\%$), which exhibits the lowest performance and the highest relative variability ($CV = 1.52\%$). The ratio of the highest to the lowest standard deviation across models is just $1.27*$, and Levene's test does not reject the hypothesis of homogeneity of variances ($p = 0.692$), indicating comparable dispersion among the four models.

Figure 2 details the evolution of the F1-scores obtained by each model across each of the twenty datasets.

## 3.2. Assumption Checks

The Shapiro-Wilk test does not reject the assumption of normality for any of the four models ($p > 0.05$ in all cases; see Table 2, which would, in principle, allow for the consideration of parametric alternatives. However, the Friedman test is retained as the primary analysis because it is the most appropriate for the study's paired-blocks design.

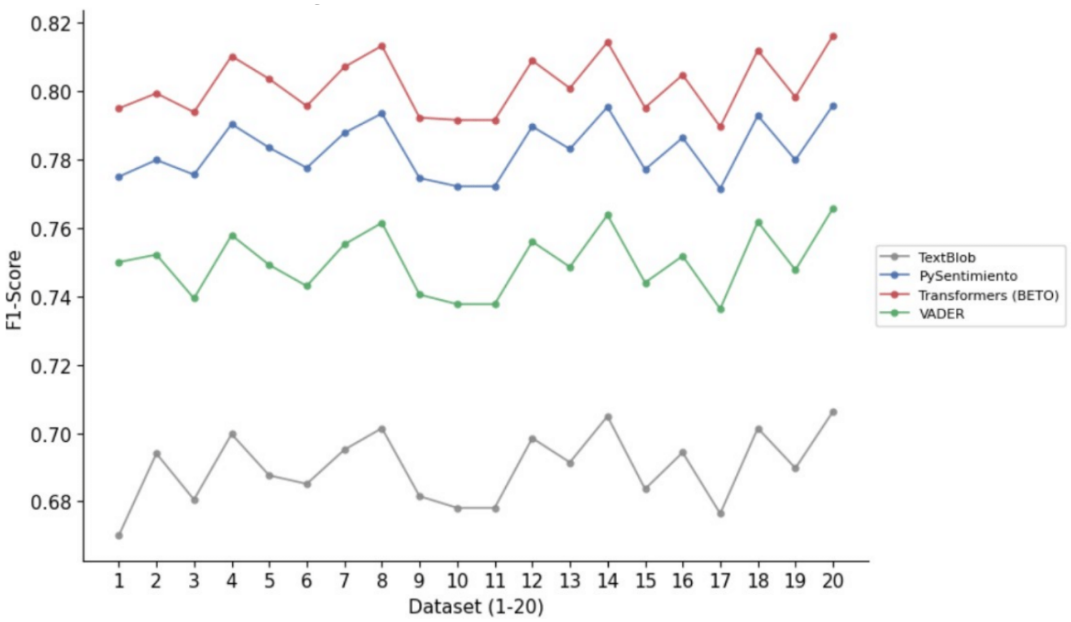


**Figure 2:** Evolution of the F1-Score by model across the 20 datasets.

**Table 2**
Descriptive statistics by model ($n = 20$ datasets). $\mu$ = mean, $\sigma$ = standard deviation.

| Model | W (Shapiro-Wilk) | P-value | Normal ($\alpha = 0.05$) |
|---|---|---|---|
| Transformers (BETO) | 0.9279 | 0.1410 | True |
| PySentiment | 0.9248 | 0.1223 | True |
| VADER | 0.9519 | 0.3975 | True |
| TextBlob | 0.9656 | 0.6601 | True |

### 3.3. Statistical analysis: Friedman

The Friedman test applied to the F1-Score values yielded a $\chi^2$ statistic of $60.00$ with 3 degrees of freedom ($p < 0.001$), which is much greater than the critical value $\chi^2(3, \alpha = 0.001) = 16.27$. This result allows us to reject, with a high level of confidence, the null hypothesis that the distributions across models are equal. Kendall's W statistic reaches its theoretical maximum value ($W = 1.00$), indicating perfect agreement among the twenty datasets regarding the order of the four models: in each of the twenty datasets, as can be seen in Figure 3, without a single exception, the observed performance order was identical (Transformers > PySentimiento > VADER > TextBlob). The mean Friedman ranks ($1 =$ best performance) unambiguously reflect this hierarchy: Transformers ($1.00$), PySentimiento ($2.00$), VADER ($3.00$), and TextBlob ($4.00$).

Table 3 presents the paired post-hoc comparisons (Nemenyi and Wilcoxon with Holm's correction) for the six possible comparisons.

The six p_holm values are identical (0.0005) and well below the conventional significance threshold ($\alpha = 0.05$), even after the most conservative correction for multiple comparisons. As discussed in Section 5, this result should not be interpreted as a favorable coincidence of the Holm adjustment, but rather as a direct consequence of the fact that the six comparisons already started from the minimum possible raw p-value for a Wilcoxon test with $n = 20$ ($p = 0.0001$): in none of the twenty datasets was the direction of the difference between the two compared models reversed, which led the Wilcoxon_W statistic to its theoretical minimum value ($W = 0$) in all six comparisons.

Similarly, the values of r_biserial reach their maximum possible magnitude ($\pm 1.00$) in all six comparisons, well above what is typical in studies comparing classifiers, where values between 0.60 and 0.80 are already considered a consistent advantage.

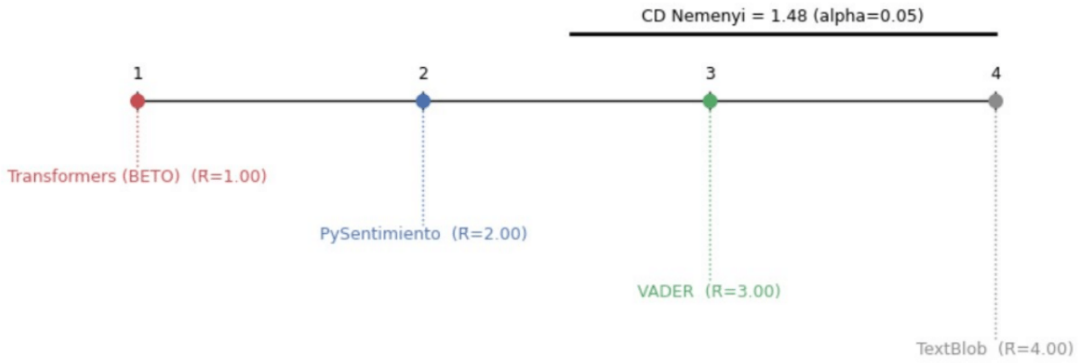


**Figure 3:** Friedman's mean ranks and Nemenyi's critical difference (rank 1 = best performance).

**Table 3**
Pairwise comparisons (paired post-hoc: Nemenyi/Wilcoxon with Holm correction). $\mu$ = mean, $\sigma$ = standard deviation of the F1 difference between pairs; $r$ = rangepair biserial correlation; $d\_z$ = paired Cohen's $d$.

| Comparison | Mean_F1_Dif | DE_diff | Wilcoxon_W | raw_p | r_biserial | d_z | p_holm | significant |
|---|---|---|---|---|---|---|---|---|
| TextBlob vs. PySentimiento | -0.0928 | 0.0037 | 0.0000 | 0.0001 | -10.000 | -25.3981 | 0.0005 | True |
| TextBlob vs. Transformers (BETO) | -0.1118 | 0.0038 | 0.0000 | 0.0001 | -10.000 | -29.3445 | 0.0005 | True |
| TextBlob vs. VADER | -0.0601 | 0.0048 | 0.0000 | 0.0001 | -10.000 | -12.4684 | 0.0005 | True |
| PySentiment vs. Transformers (BETO) | -0.0190 | 0.0008 | 0.0000 | 0.0001 | -10.000 | -22.8449 | 0.0005 | True |
| PySentiment vs VADER | 0.0327 | 0.0027 | 0.0000 | 0.0001 | 10.000 | 12.1916 | 0.0005 | True |
| Transformers (BETO) vs. VADER | 0.0517 | 0.0023 | 0.0000 | 0.0001 | 10.000 | 22.0854 | 0.0005 | True |

### 3.4. Consolidated ranking and model selection criteria

Table 4 summarizes the rankings of the four models based on the four evaluated metrics, showing an F1 score gap of 11.18 percentage points between the highest-performing model (Transformers, F1=80.17%) and the lowest-performing one (TextBlob, F1=68.99%), a difference that remains consistent across the other three metrics (11.53 points in Accuracy, 11.06 in Precision, and 10.8 in Recall). This gap reflects the structural advantage of models based on Transformer architectures with supervised fine-tuning over rule-based lexical approaches, and is corroborated by the Friedman rank, which ranks the four models identically across all four metrics (Transformers=1.00, PySentimiento=2.00, VADER=3.00, TextBlob=4.00), with no reversals in position across metrics.

## 4. Discussion

The results of this study provide affirmative answers to both research questions posed. Regarding RQ1, Transformers (BETO) is the model with the highest average performance in Spanish sentiment analysis (F1 $\mu = 80.17\%$), a finding consistent with recent literature documenting the superiority of pre-trained models in Spanish [23, 24, 25]. This advantage is attributable to BETO's ability to capture bidirectional contextual dependencies and complex semantic nuances —which lexical and probabilistic approaches cannot adequately model [26]).

**Table 4**
Models with performance statistics and operational selection criteria based on application context.

| Model | F1 | Accuracy | Precision | Recall | Friedman Range | Usage Profile |
|---|---|---|---|---|---|---|
| Transformers | 80.17% | 81.44% | 80.57% | 79.7%* | 1.00 | High accuracy; GPU-accelerated training available; labeled corpus for fine-tuning; low error tolerance |
| PySentiment | 78.27% | 79.59% | 78.66% | 77.9%* | 2.00 | Spanish-language production; mixed data; stable performance required; optional GPU |
| VADER | 75.00% | 76.96% | 75.61% | 74.7%* | 3.00 | No labeled corpus; no GPU; short, expressive texts (social media); minimal computational cost |
| TextBlob | 68.99% | 69.91% | 69.51% | 68.9%* | 4.00 | Texts in English or with acceptable machine translation; rapid prototyping; no high-precision requirements |

With regard to PI2, the Friedman test ($\chi^2 = 60.00, p < 0.001$) and paired comparisons (Nemenyi and Wilcoxon with Holm's correction) confirm that the differences between all pairs of models are statistically significant. Unlike previous studies reporting comparisons in which some model pairs did not reach statistical significance, in this dataset all six comparisons are significant by a considerable margin ($p < 0.001$ in all cases), a result of the perfect concordance (Kendall's $W = 1.00$) among the twenty datasets. This finding has relevant practical implications: in contexts where computational cost is a limiting factor, VADER may serve as an equivalent alternative to PySentimiento [27].

## 5. Limitations

The sample size of twenty datasets, while consistent with similar studies comparing classifiers, offers limited statistical power to detect differences of a smaller magnitude than those observed here; future studies could increase the number of blocks to strengthen the generalizability of the findings. Similarly, the evaluation is limited to the aggregated metrics of accuracy, precision, recall, and macro F1 score, without a breakdown by class (positive, negative, neutral) that would allow us to identify whether the differences between models are uniform across all sentiment categories or concentrated in a particular class. Finally, the study is limited to evaluating the four models in their standard or original configuration, without comparing their performance against larger-scale language models (general LLMs) that have recently emerged as an alternative for sentiment classification tasks in Spanish.

## 6. Conclusion

This study conducted a comparative evaluation, using a paired statistical protocol, of four sentiment analysis models in Spanish across twenty datasets. The application of the Friedman test ($\chi^2 = 60.00, p < 0.001$, Kendall's $W = 1.00$), paired post-hoc comparisons (Nemenyi and Wilcoxon with Holm's correction), and the calculation of the paired Cohen's d allowed us to quantify both the existence and the magnitude of the differences between models.

Transformers (BETO) is the highest-performing model (F1 $\mu = 80.17\%$), statistically significantly outperforming the remaining three models. The order of performance (Transformers > PySentimiento > VADER > TextBlob) is identical across all twenty evaluated datasets, without a single exception.

TextBlob shows the lowest performance (F1 $\mu = 68.99\%$) and the highest relative variability, consistent with its reliance on non-specialized lexical resources for Spanish.

As future work, we propose extending the evaluation to larger-scale models, incorporating cross-domain adaptation techniques to reduce the variability observed in corpora with atypical class distributions, and developing a unified and reproducible benchmark for sentiment analysis in Spanish.

## Online Resources

Sentiment analyzer available at https://github.com/MundoCode777/Analisis-De-Sentimientos.

Analyzed datasets:

1. http://tass.sepln.org/2020/?page_id=74
2. https://huggingface.co/datasets/softecapps/spam_ham_spanish/tree/main
3. https://huggingface.co/datasets/jr-garcia/wikipedia_es_qna
4. https://huggingface.co/datasets/alonso02rupa/SpanishEconomyRedditComments/tree/main
5. https://www.kaggle.com/code/philipsanm/sentiment-analysis/input
6. https://www.kaggle.com/datasets/johnurpay/comentarios-instagram-universidades-y-diccionarios?select=ComentariosInstagram.xlsx
7. https://github.com/argilla-io/pln_con_rubrix/blob/main/datos/tweets_en_es.csv
8. https://www.kaggle.com/datasets/anamumaq/opiniones-en-facebook-de-alan-garca-deceso
9. https://www.kaggle.com/datasets/antonioproao/comentarios-de-youtube-de-violencia-institucional
10. https://www.kaggle.com/code/nicolaslav/academia-bintec-nlp/input
11. https://huggingface.co/datasets/mrcsgh/horeca-spanish-reviews/viewer
12. https://huggingface.co/datasets/yabramuvdi/NoticiasColombia/tree/main
13. https://huggingface.co/datasets/Karpacious/hotel-reviews-es/tree/main
14. https://www.kaggle.com/datasets/radimkzl/edit-amazon-reviews-multi
15. https://www.kaggle.com/code/lazaro97/clustering-sentiment-analysis
16. https://www.kaggle.com/code/marcelofab/k-means-clasificaci-n-de-texto/input
17. https://github.com/LeonardoDev24/movie-corn-reviews/tree/main/src/data
18. https://github.com/mrcsgh/horeca-spanish-reviews/blob/main/dataset.txt
19. https://huggingface.co/datasets/SINAI/ALIA-es-cultural-heritage/tree/main
20. https://www.kaggle.com/datasets/lewisdelacruz/text-vph/data

## Declaration on Generative AI

During the preparation of this paper, the first author used Claude Sonnet 5.0 to search for references and DeepLM for translation.

## References

[1] X. Zhang, J. Sun, X. Li, Y. Liu, C. Li, Developing a framework for online review-based health care service quality assessment: text-mining study, Journal of Medical Internet Research 27 (2025) e66141.

[2] M. Yang, B. Jiang, Y. Wang, T. Hao, Y. Liu, News text mining-based business sentiment analysis and its significance in economy, Frontiers in Psychology 13 (2022) 918447.

[3] A. Yadav, D. K. Vishwakarma, Sentiment analysis using deep learning architectures: a review: A. yadav, dk vishwakarma, Artificial intelligence review 53 (2020) 4335–4385.

[4] P. Pakray, A. Gelbukh, S. Bandyopadhyay, Natural language processing applications for low-resource languages, Natural Language Processing 31 (2025) 183–197.

[5] P. Krasadakis, E. Sakkopoulos, V. S. Verykios, A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages, Electronics 13 (2024) 648.
[6] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Urena-López, M. T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Systems with Applications 166 (2021) 114120.
[7] R. Dror, S. Shlomov, R. Reichart, Deep dominance-how to properly compare deep neural models, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 2773–2785.
[8] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, et al., Accounting for variance in machine learning benchmarks, Proceedings of machine learning and systems 3 (2021) 747–769.
[9] J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. E. Anke, F. Liu, E. Martínez-Cámara, et al., Tweetnlp: Cutting-edge natural language processing for social media, in: Proceedings of the 2022 conference on empirical methods in natural language processing: System demonstrations, 2022, pp. 38–49.
[10] S. Minaee, E. Azimi, A. Abdolrashidi, Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models, arXiv preprint arXiv:1904.04206 (2019).
[11] A. Ligthart, C. Catal, B. Tekinerdogan, Systematic reviews in sentiment analysis: a tertiary study, Artificial intelligence review 54 (2021) 4997–5053.
[12] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: Approaches, challenges and trends, Knowledge-based systems 226 (2021) 107134.
[13] M. A. Palomino, F. Aider, Evaluating the effectiveness of text pre-processing in sentiment analysis, Applied Sciences 12 (2022) 8765.
[14] J. Camacho-Collados, M. T. Pilehvar, On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 40–46.
[15] MundoCode777, Sentiment analysis, 2026.
[16] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (2020) 6.
[17] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (2009) 427–437.
[18] D. M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, arXiv preprint arXiv:2010.16061 (2020).
[19] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Scientific reports 14 (2024) 6086.
[20] M. C. Hinojosa Lee, J. Braet, J. Springael, Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores, Applied Sciences 14 (2024) 9863.
[21] M. C. Díaz-Galiano, M. G. Vega, E. Casasola, L. Chiruzzo, M. Á. G. Cumbreras, E. M. Cámara, D. Moctezuma, A. Montejo-Ráez, M. A. S. Cabezudo, E. S. Tellez, et al., Overview of tass 2019: One more further for the global spanish sentiment analysis corpus., in: IberLEF@ SEPLN, 2019, pp. 550–560.
[22] E. M. Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. G. Cumbreras, M. G. Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Munoz, et al., Overview of tass 2018: Opinions, health and emotions., in: TASS@ SEPLN, 2018, pp. 13–27.
[23] H. Gomez-Adorno, G. Bel-Enguix, G. Sierra, J.-C. Barajas, W. Álvarez, Machine learning and deep learning sentiment analysis models: Case study on the sent-covid corpus of tweets in mexican spanish, in: Informatics, volume 11, MDPI, 2024, p. 24.
[24] R. Pan, J. A. García-Díaz, R. Valencia-García, Individual-vs. multiple-objective strategies for targeted sentiment analysis in finances using the spanish mtsa 2023 corpus, Electronics 13 (2024) 717.

[25] R. Pan, J. A. García-Díaz, F. Garcia-Sanchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in spanish, PeerJ Computer Science 9 (2023) e1377.
[26] J. A. Gonzalez, L.-F. Hurtado, F. Pla, Twilbert: Pre-trained deep bidirectional transformers for spanish twitter, Neurocomputing 426 (2021) 58–69.
[27] H. Bashiri, H. Naderi, Comprehensive review and comparative analysis of transformer models in sentiment analysis, Knowledge and Information Systems 66 (2024) 7305–7361.