SINAI at SatiSPeech in IberLEF 2025: Detection of Satire in Spanish Texts Using Stylometric and Linguistic Features.

César Espin-Riofrio[1,*], Jenny Ortiz-Zambrano[1] and Arturo Montejo-Ráez[2]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

This study addresses satire detection in Spanish texts within the SatiSPeech 2025 challenge, using a diverse set of supervised classifiers and a voting classifier as an ensemble model. The models were trained on stylometric and linguistic features extracted from the texts, including lexical representations through n-grams evaluated at different quantities, as well as polarity, irony, syntactic structures, and textual complexity measures. In our evaluation, the voting classifier achieved the best performance, reaching perfect scores (1.0) in precision, recall, and F1-score, with optimal results using 2000 n-grams. However, its effectiveness declined when evaluated on the official challenge test set, highlighting generalization issues when facing more diverse data. These findings underscore the value of combining stylometric and linguistic features with n-grams to capture the nuances of satirical language, as well as the need to explore domain adaptation strategies to improve model robustness.

Keywords

Satire detection, Stylometric and linguistic features, Text classification, Natural Language Processing

1. Introduction

The detection of satirical texts is a complex task in natural language processing (NLP) that involves identifying ironic or humorous intent in written content. This challenge arises from the nuanced nature of satire, which often employs figurative language that contradicts literal meanings. While satire is a legitimate form of cultural and political expression, its intentional ambiguity presents particular difficulties for text classification models.

Satire conveys sentiments that are often contrary to literal statements [1], employing wit, irony, or sarcasm [2] to expose absurdity and critique social issues through comedic devices [3]. It is a rhetorical strategy commonly found in informal online content [4], where humor serves as a vehicle for social commentary and indirect criticism [5].

Satirical texts often lack explicit markers, making it difficult for models to discern the author's true intent [6]. Moreover, satire heavily relies on context, which can mislead detection algorithms if not properly accounted for [3].

In this work, we propose an approach based on supervised learning techniques that incorporates stylometric features for the detection of satire in written texts. Stylometry offers a set of quantifiable measures—such as the distribution of function words, syntactic complexity, and lexical variation—that can capture subtle signals of the distinctive discursive style of satire. In contrast to approaches focused solely on thematic content or semantic embeddings, our proposal emphasizes the formal aspects of language as a source of discriminative information.

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
cesar.espinr@ug.edu.ec (C. Espin-Riofrio); jenny.ortizz@ug.edu.ec (J. Ortiz-Zambrano); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0001-8864-756X (C. Espin-Riofrio); 0000-0001-6708-4470 (J. Ortiz-Zambrano); 0000-0002-8643-2714 (A. Montejo-Ráez)

2. Related work

The detection of satirical texts is a complex challenge in NLP due to the nuanced use of language that often contradicts literal meanings. Several studies have addressed automatic satire detection through linguistic, statistical, and deep learning approaches, highlighting the complexity of the task due to satire's figurative and ambiguous nature.

[7] applied traditional classifiers such as Random Forest, Naive Bayes, and Logistic Regression to linguistic features, achieving good performance in distinguishing satire from other content types. In a follow-up study, concluded that features related to text composition—such as writing style, paragraph structure, and readability—were more effective than other types. Similarly, [8] compared linguistic features with term-based metrics, finding comparable performance (84% and 83.5% accuracy, respectively), with a slight advantage for linguistic features.

[3] emphasized that satire poses a challenge for NLP and sentiment analysis systems due to its frequent use of figurative language, which can lead to ambiguity. In this context, [9] revealed limitations in automatic satire detection tools, which achieved under 73% accuracy, compared to 87% for human detection. They suggested enhancing model performance by incorporating related features such as irony and exaggeration.

Other studies have explored combining multiple feature types. [10] showed that Random Forest, when used with linguistic, psychological, and punctuation features, achieved 96.92% accuracy. [11] reported that SVM-based models outperformed other methods—such as Random Forest, Boosting Trees, and Naive Bayes—achieving over 95% accuracy.

Neural network-based models have also proven effective. [12] implemented an RNN with LSTM cells and attention mechanisms, achieving F1-scores of 0.82 (irony) and 0.76 (satire), further improved by incorporating emotional features. [4] proposed BiSAT, a model combining bidirectional memory with attention mechanisms to focus on words or phrases that implicitly convey satire.

Transformer-based models have gained prominence. [13] trained a DistilBERT-based model to distinguish satire from fake news, reporting performance gains of 5.2% in F1-score and 6.4% in accuracy compared to traditional methods. [14], using their Spanish-language corpus SatiCorpus, showed that combining linguistic features with BERT boosted accuracy from 85.15% to 97.40%.

Finally, some studies proposed hybrid strategies. [15] suggested using two separate linguistic models—one for satirical and one for authentic content—leading to improved classification by capturing the specific features of each text type. [16] approached satire detection with a multimodal model, arguing that satirical texts are often accompanied by images that reinforce figurative meanings and help infer the intended message.

3. Method

For the classification task, several supervised models were employed, including a Linear Support Vector Machine (SVC), a probabilistic Support Vector Machine with a linear kernel (SVM), a Random Forest classifier (RF), a Multilayer Perceptron neural network (MLP), and an Extreme Gradient Boosting classifier (XGB). Additionally, a Voting Classifier (VC) was used to combine the predictions of the base models through soft voting, i.e., by averaging the predicted class probabilities from each individual model. All models were trained using stylometric and linguistic features extracted from the input texts, capturing both surface-level patterns and deeper structural and semantic cues relevant to satire detection.

3.1. Data

The dataset for the SatiSPeech Task at IberLEF 2025 [17] was specifically created to tackle the challenge of detecting satire in a multimodal context, incorporating both text and audio. The data was collected from a diverse selection of YouTube channels, encompassing satirical programs like El Intermedio, Zapeando, Homo-Zapping, and El Mundo Today, as well as non-satirical news programs such as Antena

3 Noticias, El Mundo, and BBC News. This approach ensures a wide representation of the various Spanish dialects spoken across different regions.

3.2. Pre-processing

The text data underwent several preprocessing steps to prepare it for analysis. This included cleaning the text by removing special characters, numbers, and irrelevant symbols. Stopwords were removed to reduce noise, and lemmatization was applied to standardize words to their root forms. This preprocessing ensured that the models would focus on the meaningful content of the text, enhancing the effectiveness of the satire detection task.

3.3. Text features

As mentioned above, the features extracted were stylometric and linguistic in nature and are described in Table 1.

Table 1
Description of stylometric and linguistic features for satire detection

Feature	Description
num_words	Total number of words in the text.
num_chars	Total number of characters in the text.
exclamations	Number of exclamation marks.
uppercase_ratio	Ratio of uppercase letters to total letters.
polarity	Average sentiment polarity (positive/negative).
subjectivity	Degree of subjectivity in the text.
vader_polarity	Sentiment polarity using the VADER analyzer.
irony_score	Estimated irony score.
PoS (ADV, NOUN, VERB, ADJ)	Proportion of key part-of-speech tags.
rhetorical_questions	Number of rhetorical questions.
metaphors	Estimated number of metaphors.
satire_words_density	Density of satire-related words.
total_satire_words	Total count of satire-related words.
MeanWordLen	Average word length in characters.
LexicalDiversity	Lexical diversity (unique words / total words).
MeanSentenceLen	Average sentence length in words.
StdevSentenceLen	Standard deviation of sentence lengths.
MeanParagraphLen	Average paragraph length in sentences.
DocumentLen	Total document length in words.
WordsPerText	Average number of words per text.
SentencesPerText	Average number of sentences per text.
MeanDifferenceSentenceLengths	Mean difference between consecutive sentence lengths.
Average Syntactic Depth	Average depth of syntactic trees.
Average Dependency Length	Average length of syntactic dependencies.
Flesch Score	Flesch readability score (adapted for Spanish).
Lexical Entropy	Lexical entropy (vocabulary unpredictability).
Syntactic Repetition	Degree of repetition in syntactic structures.
Unusual Word Frequency	Frequency of rare or unusual words.
n-grams	Frequency distribution of words (unigrams and bigrams).
TF-IDF	Term frequency–inverse document frequency scores.

4. Experiment

The experimental phase was conducted using the classification models previously described, including both individual classifiers and a voting-based ensemble. Prior to training, we applied feature selection

methods to reduce dimensionality and retain the most relevant stylometric and linguistic features. This step aimed to enhance model generalization and mitigate the risk of overfitting.

As part of the feature engineering process, we also incorporated n-gram representations (unigrams and bigrams) and explored the impact of feature space size by experimenting with different vocabulary sizes: 500, 1000, 2000, 3000, and 5000 most frequent n-grams based on term frequency. This allowed us to evaluate how the quantity of lexical features influences model performance.

Each model was trained on the selected features and evaluated using a stratified train-test split. We assessed classification performance using standard metrics: accuracy, precision, recall, and F1-score. The experiments were designed to determine not only the overall effectiveness of each model but also the contribution of different feature types and quantities to the task of satire detection.

5. Results

The evaluation phase was conducted using an 80/20 training-validation split and an independent test set to assess model generalization. As presented in Table 2, the Voting Classifier (VC) consistently outperformed all individual models across the standard evaluation metrics—accuracy, precision, recall, and F1-score. Its optimal configuration, obtained using 2000 TF-IDF features, yielded a performance of approximately 92.33% across all metrics, indicating a high degree of reliability and balance between sensitivity and specificity. Notably, the VC model exhibited stable behavior across different feature set sizes (1000, 3000, 5000), with only marginal variations in performance, demonstrating its robustness with respect to the dimensionality of the input representation. In contrast, the XGBoost (XGB) classifier, although competitive, achieved slightly inferior results, with a maximum validation F1-score of 0.9097 under the same feature configuration. To provide a detailed insight into the classifier's performance, Figure 1 displays the confusion matrix corresponding to the best VC model. The model correctly classified 577 non-satirical (true negatives) and 529 satirical instances (true positives), while producing 50 false positives and 44 false negatives. This balanced distribution of classification errors reinforces the model's high precision (0.9231) and recall (0.9234) scores, confirming its capability to accurately distinguish between the two classes without exhibiting skew or overfitting to a particular label. In summary, the results underscore the effectiveness of the Voting Classifier in the satire detection task. When configured with 2000 TF-IDF features, it achieves a well-balanced trade-off between precision and recall, with a low misclassification rate, thereby positioning it as a robust and reliable approach for binary text classification in this domain.

Table 2
Validation metrics (80/20 train split) for top-performing models with varying `max_features`

Model	Accuracy	Precision	Recall	F1-score	max_features
Voting Classifier (VC)	0.9233	0.9231	0.9234	0.9232	2000
Voting Classifier (VC)	0.9200	0.9197	0.9202	0.9199	3000
Voting Classifier (VC)	0.9167	0.9164	0.9168	0.9165	1000
Voting Classifier (VC)	0.9167	0.9165	0.9165	0.9165	5000
XGBoost (XGB)	0.9100	0.9105	0.9092	0.9097	2000

When evaluated on the independent test set (Table 3), the Voting Classifier achieved perfect scores (1.0) in accuracy, recall, and F1-score across all tested feature configurations. By contrast, XGBoost yielded a still-high F1-score of 0.9792, suggesting that while it remains highly effective, the ensemble approach of the Voting Classifier is more adept at capturing the subtle stylistic and linguistic cues characteristic of satirical texts when trained on the selected feature set.

In the context of task 1 from the SatiSPeech 2025 workshop [18], the system submitted by the author, identified as **cespinr (Sinai)**, achieved an F1-score of 0.794787, resulting in a 10th place ranking in the official leaderboard (Table 4). While this score did not place the system among the top three, the performance remains competitive, with a performance gap of less than 6% compared to the leading

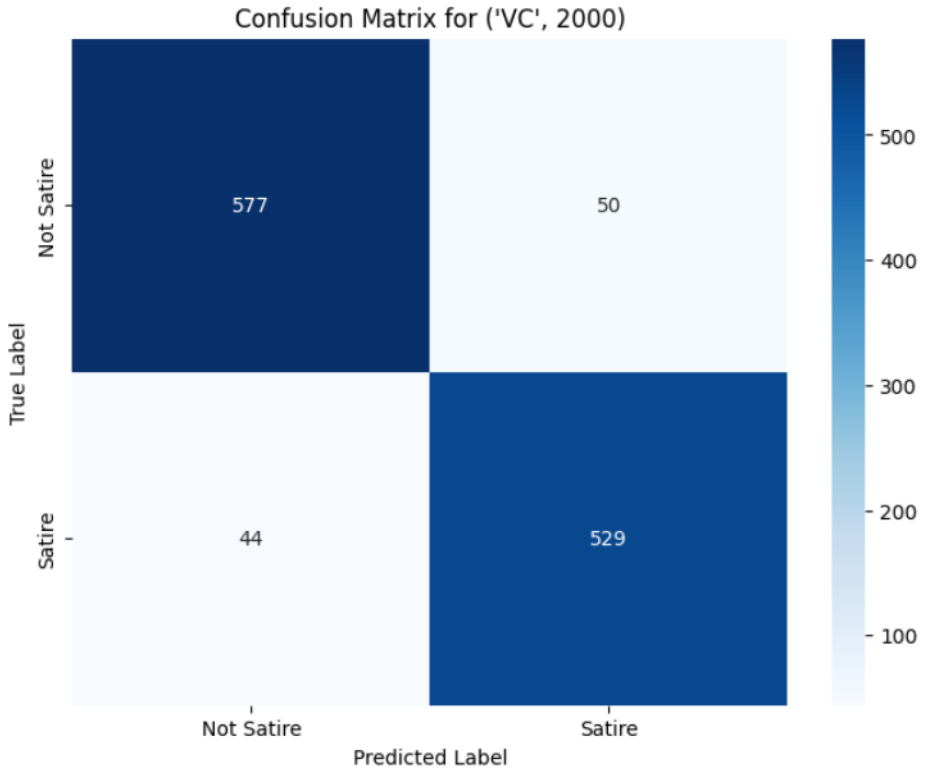

Figure 1: Confusion matrix of the best Voting Classifier model with 2000 TF-IDF

Table 3
Test set performance metrics for top models using different `max_features`

Model	Accuracy	Recall	F1-score	max_features
Voting Classifier (VC)	1.0000	1.0000	1.0000	2000
Voting Classifier (VC)	1.0000	1.0000	1.0000	3000
Voting Classifier (VC)	1.0000	1.0000	1.0000	1000
Voting Classifier (VC)	1.0000	1.0000	1.0000	5000
XGBoost (XGB)	0.9792	0.9792	0.9792	2000

submission *mcastro* (F1-score of 0.856376). These results are particularly noteworthy given the inherent complexity of satire detection in spoken language, a task that requires the system to capture subtle linguistic cues and prosodic variations. The model demonstrated a strong capacity for generalization and outperformed several other participants, which underscores the robustness of the methodological choices made. Overall, this outcome constitutes a meaningful contribution to the task and provides a solid foundation for future improvements aimed at narrowing the remaining performance gap.

Table 4
F1-scores for Task 1 by user

#	User	F1 Task 1
1	mcastro	0.856376
2	ITST	0.845458
3	edu_valero	0.844550
...	...	...
10	**cespinr (Sinai)**	**0.794787**
11	JoseAGD	0.793733
12	deniscedeno	0.779364

6. Conclusions

The Voting Classifier proved to be the most effective model for satire classification in Spanish texts, achieving perfect scores on the independent test set and demonstrating high predictive power and stability. The use of stylometric and linguistic features—particularly through n-gram representations—was essential for capturing the distinctive stylistic and lexical cues of satirical language. Different configurations of feature quantity (max_features = 1000, 2000, 3000, 5000) were evaluated, with the best performance observed when using 2000 features.

XGBoost also delivered competitive results but consistently fell slightly short of the ensemble model, suggesting that individual approaches may struggle to fully grasp the nuances of satirical expression.

Despite strong results in validation and internal testing, the system's performance dropped notably on the unlabeled test set from the SatiSPeech 2025 competition. This gap highlights the challenges of generalizing satire detection to more diverse, real-world data and underscores the importance of further refining feature selection and exploring domain adaptation techniques to enhance model robustness. Future improvements to the proposed satire detection system could include the integration of pretrained acoustic models and multimodal representations that combine textual and prosodic features. Additionally, the use of data augmentation and self-supervised learning may enhance generalization in low-resource settings. Conducting a thorough error analysis could also inform more targeted refinements in future iterations.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government. Computational resources provided by Red Española de Supercomputación (activity FI-2025-1-0003) have been used to run our experiments.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 for grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Reganti, T. Maheshwari, A. Das, E. Cambria, Open secrets and wrong rights: automatic satire detection in english text, in: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 291–294.

[2] A. N. Reganti, T. Maheshwari, U. Kumar, A. Das, R. Bajpai, Modeling satire in english text for automatic detection, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE, 2016, pp. 970–977.

[3] G. Casalino, A. Cuzzocrea, G. L. Bosco, M. Maiorana, G. Pilato, D. Schicchi, A novel approach for supporting italian satire detection through deep learning, in: International Conference on Flexible Query Answering Systems, Springer, 2021, pp. 170–181.

[4] A. Kamal, M. Abulaish, et al., Contextualized satire detection in short texts using deep learning techniques, Journal of Web Engineering 23 (2024) 27–52.

[5] D. Goldwasser, X. Zhang, Understanding satirical articles using common-sense, Transactions of the Association for Computational Linguistics 4 (2016) 537–549.

[6] C. Shi, Research advanced in sarcastic detection based on deep learning, Theoretical and Natural Science 79 (2025) 23–27. URL: https://www.ewadirect.com/proceedings/tns/article/view/19930. doi:10.54254/2753-8818/2025.19930.

[7] A. Gaeta, F. Orciuoli, A. Pascuzzo, Satiric content detection through linguistic features, in: Machine Learning and Artificial Intelligence, IOS Press, 2023, pp. 114–119.
[8] Ó. Apolinario-Arzube, J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, R. Valencia-García, Comparing deep-learning architectures and traditional machine-learning approaches for satire identification in spanish tweets, Mathematics 8 (2020) 2075.
[9] A.-C. Rogoz, M. Gaman, R. T. Ionescu, Saroco: Detecting satire in a novel romanian corpus of news articles, arXiv preprint arXiv:2105.06456 (2021).
[10] A. Onan, M. A. Toçoğlu, Satire identification in turkish news articles based on ensemble of classifiers, Turkish Journal of Electrical Engineering and Computer Sciences 28 (2020) 1086–1106.
[11] N. Mafla, M. Flores, S. Castillo, R. Andrade, Automatic detection of fake news in spanish: Ecuadorian political satire, Revista Politécnica 50 (2022) 7–16.
[12] R. Ortega-Bueno, P. Rosso, J. E. M. Pagola, Multi-view informed attention-based model for irony and satire detection in spanish variants, Knowledge-Based Systems 235 (2022) 107597.
[13] J. F. Low, B. C. Fung, F. Iqbal, S.-C. Huang, Distinguishing between fake news and satire with transformers, Expert Systems with Applications 187 (2022) 115824.
[14] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems 8 (2022) 1723–1736.
[15] Y. Zhang, F. Yang, Y. Zhang, E. Dragut, A. Mukherjee, Birds of a feather flock together: Satirical news detection via language model differentiation, arXiv preprint arXiv:2007.02164 (2020).
[16] L. Li, O. Levi, P. Hosseini, D. A. Broniatowski, A multi-modal method for satire detection using textual and visual cues, arXiv preprint arXiv:2010.06671 (2020).
[17] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.
[18] R. Pan, J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, R. Valencia-García, Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish, Procesamiento del Lenguaje Natural 75 (2025).