Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 135-148 recibido 05-12-2025 revisado 02-02-2026 aceptado 18-02-2026

Hybrid Annotation Framework for Spanish Subjectivity Detection

Marco de anotación híbrido para la detección de subjetividad en español

Alba Pérez-Montero,[1] **Elena Lloret,**[2] **Paloma Moreda**[2]
[1]University Institute for Computing Research (IUII), University of Alicante
[2]Dept. of Software and Computing Systems, University of Alicante
alba.perezm@ua.es, {elloret,moreda}@dlsi.ua.es

Abstract: Subjectivity Detection (SD) is a fundamental NLP task that distinguishes between factual content and opinion. Its analysis provides profound linguistic knowledge, useful for tasks such as disinformation detection. Our objective is to develop and validate a hybrid annotation framework for SD in Spanish. Our methodology includes: (1) a pilot study comparing human annotation against Machine Learning (ML) and Large Language Models (LLMs); (2) an error analysis to refine the framework; and (3) the application of the consolidated framework. Results show ML achieving higher precision, but overlooking complex structures (e.g., subjunctive). The consolidated hybrid methodology is tested in three annotation setups (human [A1], unrevised automatic [A2], consolidated semi-automatic [A3]). A3 achieves the best balance between quality and time consumption (only 19% of the manual annotation effort).
Keywords: Subjectivity Detection (SD), Natural Language Processing (NLP), hybrid annotation framework, Large Language Models (LLMs).

Resumen: La Detección de Subjetividad (SD) es una tarea fundamental del PLN, que distingue entre contenido factual y opinión. Su análisis proporciona conocimiento lingüístico profundo, útil para tareas como la detección de desinformación. Nuestro objetivo es desarrollar y validar un marco de anotación híbrido para la SD en español. Nuestra metodología incluye: (1) un estudio piloto que compara la anotación humana con el aprendizaje automático (ML) y grandes modelos de lenguaje (LLMs); (2) un análisis de errores para refinar el marco; y (3) la aplicación del marco consolidado. Los resultados muestran que ML logra mayor precisión, pero ignora estructuras complejas (ej., subjuntivo). La metodología híbrida se prueba en tres configuraciones de anotación (humana [A1], automática no revisada [A2], semi-automática consolidada [A3]). La A3 logra el mejor equilibrio entre calidad y consumo de tiempo (solo 19% de la anotación manual).
Palabras clave: Detección de subjetividad (SD), Procesamiento del Lenguaje Natural (PLN), marco de anotación híbrido, Grandes Modelos de Lenguaje (LLMs).

1 Introduction

Subjectivity is an intrinsic element of everyday communication. Since language is fundamental human act, every linguistic performance is a unique manifestation that carries individual opinions and, therefore, may reflect the individuality of the speaker (Bajtin, 1982). This linguistic individuality not only reflects the author's viewpoint (whether explicit, e.g., using the first person singular, or implicit, e.g., selecting evaluative adverbs or adjectives) but also the underlying intentionality. As noted by Hyland (2005), "we project ourselves into our texts to manage our communicative intentions". This intention is reflected through linguistic markers of subjectivity, making the analysis of subjective features critical in contexts where discourse is used as a weapon; this includes attempts to persuade the public, deceive readers with false information, or generate hate speech targeting specific individuals ((Saquete et al., 2020); (Ott et al., 2011)).

The development of robust systems for

ISSN 1135-5948 DOI 10.26342/2026-76-10

detecting subjectivity depends on creating high-quality annotated resources containing accurate representations of subjectivity linguistic markers. The creation of such resources is crucial, as they enable the training of models adjusted for this specific task. However, manual corpus creation is costly and time-consuming, making the integration of Natural Language Processing (NLP) methodologies necessary to automate this process. The methodological focus of this research centers on developing a hybrid methodology that combines Machine Learning (ML) efficiency and Large Language Models (LLMs) advanced reasoning capabilities. This approach aims to find a balance between capacity and quality, with the human expert remaining crucial throughout the process for defining guidelines and evaluating model outputs to optimize the final annotation pipeline. To achieve this goal, the proposed methodology is structured around three steps: (1) a initial pilot study used to establish preliminary results, and (2) a subsequent exhaustive error analysis that leads directly to the design of (3) the definitive hybrid annotation framework. Based on the results derived from this optimization process, this paper presents three main contributions:

1. We propose a novel **hybrid annotation framework** for SD in Spanish, which strategically combines the efficiency of traditional ML with the advanced reasoning of LLMs.
2. We present a comprehensive set of **annotation guidelines**, refined through an iterative error analysis process.
3. We develop a new **annotated dataset**. This dataset comprises 626 sentences extracted from journalistic sources.

The paper is structured to reflect the development and validation of this framework. Section 2 presents the related work by reviewing relevant studies about subjectivity detection, hybrid approaches on annotation, and the use of LLMs as annotators. Section 3 details the design of the pilot study. Section 4 covers the iterative experimental design and the comprehensive error analysis derived from the pilot study results, which leads to the consolidation of the final hybrid annotation framework. Section 5 presents the main experiment results, evaluating the consolidated hybrid framework's performance and efficiency gains. Finally, we present the discussion in Section 5.4 and the final conclusions and future work in Section 6.

2 Related Work

2.1 Subjectivity Detection and Annotation

Subjectivity Detection (SD) is a task closely linked to sentiment analysis and opinion mining, reason why researchers often treat these areas as a "unified body of work" ((Krüger et al., 2017); (Pang, Lee, and others, 2008)). Subjective content itself covers a wide spectrum of human expression, encompassing feelings, emotions, evaluations, and personal opinions (Kasmuri and Basiron, 2017). Every NLP task requires quality annotated data for developing accurate and reliable resources and models. At this point, a crucial challenge arises in the field of SD: annotating subjective language is a task inherently subjective, what makes achieving consistent annotation extremely difficult. This inconsistency appears because accurately interpreting subjective language requires complex linguistic analysis, forcing annotators to deeply evaluate complex language structures and pragmatic intent. Such analysis demands an elevated degree of reflexivity about the linguistic elements and their relationship to the real world, that is, a high metadiscoursive or metalinguistc ability.

The concept of *metadiscourse* refers to linguistic markers that presents the author's stance and organize the discourse structure, operating outside the text's core propositional content (Dynel, 2023). These markers enhance understanding by providing the audience with a guide for interpretation. Following Hyland (2005) well-established conceptualization, metadiscourse transmits speaker involvement by conveying the author's communicative intention, attitude, positioning, and commitment to both the discourse and the audience. Specifically, linguistic markers like hedges (e.g., *possibly*) and boosters (e.g., *unquestionable*) reflect the author's degree of certainty and selfmention, managing audience interaction and contributing to the author's credibility.

Previous studies exemplify some of the key areas of application for SD and present diverse efforts required to annotate subjectiv-

ity: for instance, (Antici et al., 2021) developed a specialized corpora for tagging subjective language in Italian, while (Vieira et al., 2020) developed five subjectivity lexicons and applied them to distinguish verified news from disinformation in Portuguese. Recent studies also focus on integrating LLMs into SD. For instance, Shokri et al. (2024) explore fine-tuning on models like BERT and Llama-2 and the use of Chain-of-Thought (CoT) methods, guided by annotation guidelines, to assess the LLMs' capacity for the complex linguistic reasoning required for English subjectivity analysis.

2.2 Large Language Models as Annotators

High-quality, reliable data annotation is foundational to NLP, but complex manual annotation tasks are often costly. This constant challenge has motivated the development of annotation approaches that automatize the process, involving both quality and reproducibility. Within these methodologies, LLMs have established themselves as effective annotators, resulting in novel annotation methods (Akkurt et al., 2024). Their utility stems from their contextual reasoning and their strong performance across various tasks, presenting a significant opportunity to enhance annotation workflows.

The integration of LLMs as annotators directly addresses the challenge of costly and difficult to scale traditional human annotation (Da San Martino et al., 2019). Comprehensive reviews confirm the practical advantages of this approach, presenting efficiency gains of up to an 80% reduction in annotation time ((Nasution and Onan, 2024); (Jahan et al., 2024)). LLMs often exhibit high recall but low precision, frequently due to overlooking pragmatic aspects or being influenced by model biases ((Stureborg, Alikaniotis, and Suhara, 2024); (Wang et al., 2025); (Koo et al., 2024)). Nevertheless, their annotation quality is frequently perceived as "good enough" for high-complexity tasks, given that such tasks are inherently challenging even for human annotators (Kasner et al., 2025). Despite these benefits, the quality trade-off requires refined methodologies, such as structured Human-in-the-Loop (HITL) collaboration. This is essential to prevent human annotators from being biased by LLM suggestions (Schroeder, Roy, and Kabbara, 2025). Techniques like the SynCode framework (Xia et al., 2025) and role-playing prompting (Voutsa, Tsapatsoulis, and Djouvas, 2025) are employed to enhance output consistency and iterative refinement.

Consequently, the evaluation of LLMs as annotators is an active research topic, leading to alternative metrics like the Alt-Test (Calderon, Reichart, and Dror, 2025). This test moves beyond traditional performance match assessment to justify the LLM's utility by comparing its output against a human gold-standard group, providing a definitive *yes/no* result on the model's efficacy.

3 Methodology

We present a three-stage methodology designed to develop and evaluate a consolidated hybrid framework for Spanish subjectivity annotation. This approach balances computational efficiency with linguistic depth by integrating traditional ML with LLMs. This process begins with a pilot study (described in Section 4.1), using a pilot dataset and preliminary annotation guidelines to establish a baseline hybrid framework and evaluate the results by performing a critical error analysis. Findings on this stage drive the refinement of the system into a consolidated hybrid framework and revised annotation guidelines. This methodological framework is represented in Figure 1, progressing from the pilot study (left) to the central error analysis and framework refinement stage, and concluding with the main experiment (right).

3.1 Pilot Dataset and Annotation Guidelines

To ensure a robust methodology by developing consolidated quality resources we use previously created datasets as seed sources, specifically (Bonet-Jover et al., 2023) and (Posadas-Durán et al., 2019). These datasets represent different varieties of the Spanish language (Spain, México, Colombia, etc.) and also different topics (health, politics, society, and sports), this way we can obtain samples of a wide range of linguistic manifestations typologies. These resources are focused on disinformation detection task, which is a particular use case where SD plays a significant role by identifying the author's degree of involvement.

Despite previous studies employ text-level

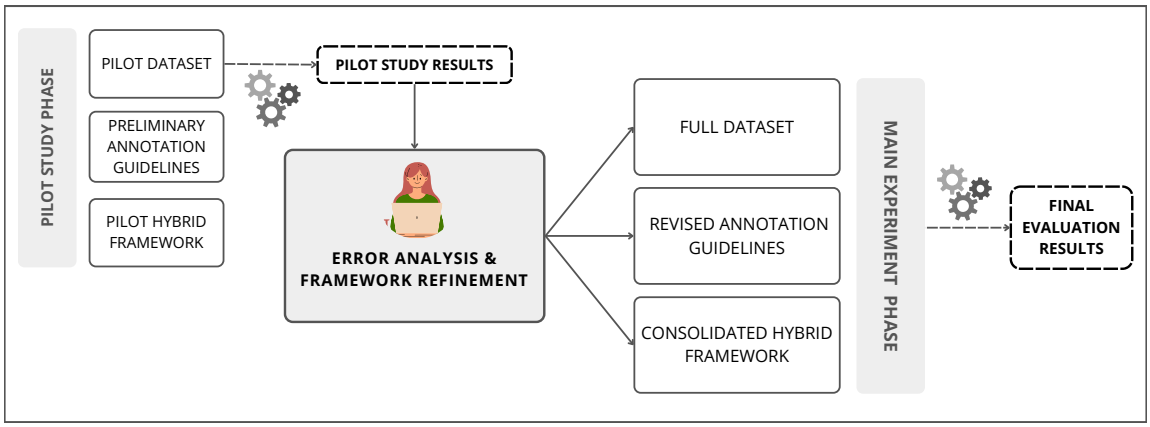

Figure 1: Diagram presenting the methodological steps on our experimental design and hybrid framework development process.

annotation, we use sentences as the unit of analysis, providing a level of granularity that efficiently captures intra-sentence linguistic markers of subjectivity.

Metric	Value
Sentences	339
Tokens	10,011
Avg. Tokens / Sentence	29.53

Table 1: Pilot dataset statistics presenting sentences, tokens and average tokens per sentence.

The pilot dataset provides the foundation for the subsequent subjectivity analysis. To ensure quality and consistency, our methodology employs an dynamic workflow, drawing on the quality management framework proposed by (Klie, Castilho, and Gurevych, 2024). This workflow is characterized by batch annotation and pilot revisions, where annotation validation is carried through exhaustive review of subsets. This validation is key to addressing ambiguities and conducting continuous guideline refinement. The initial phase of this iterative approach involves establishing precise criteria. We start by developing a set of annotation guidelines, which define specific subjectivity categories grounded in foundational linguistic studies (specifically those by (Hyland, 2005) and (Zorraquino, 1999)). These categories are conveyed in language through linguistic markers and express different degrees of author commitment, explained as follows:

- **High author commitment categories**:
  - **Certainty:** Expresses factual affirmation. This category is conveyed through verbs of knowledge (e.g., *confirmar* [to confirm]) or the future simple tense.
  - **Specificity:** Emphasizes precision and clarity of reference. This category is presented using named entities and precise numbers.
  - **Participation:** Underlines the presence or involvement of the author. This involvement is marked by first-person pronouns and verbs, or second-person pronouns and verbs.
- **Low author commitment categories**:
  - **Doubt:** Signals uncertainty or hedging. It is marked by linguistic elements that express conditionality or questioning, such as the conditional tense, the subjunctive mood, or verbs like *dudar* (to doubt).
  - **Non-specificity:** Transmits ambiguity. This category is realized through indefinite pronouns (e.g., *algunos* [some]), or generalizing terms.
  - **Distancing:** Creates detachment between the author and the discourse. This separation is reflected by the passive voice, the impersonal reflexive passive (*se dice* [it is said]), or the use of third-person pronouns.

The analysis extends beyond superficial morphological parsing. This is because words

inherently carry both semantic (meaning) and pragmatic (contextual) content (Bybee and Fleischman, 1995), specially through elements like pronouns, verbal tense variations, and specific terminology. Specifically, the author's precise choice in every word used directly reflects this complex content, which conveys their underlying opinion or commitment to the presented statement.

3.2 Pilot Hybrid Framework

As the initial step in our analysis, this pilot experiment is designed to systematically measure and compare the annotation performance of a traditional ML pipeline and LLMs against expert human annotations to determine the best methodological approach for subjectivity analysis.

Traditional Machine Learning Pipeline. We developed a three-stage annotation pipeline using the spaCy library (Honnibal et al., 2020). The process started with morphological pre-processing via PoS tagging, followed by the application of lexicon-based annotations for specific semantic markers (such as verbs of knowledge and belief, extracted from the Database of Verbs developed by the University of Vigo (Vaamonde Dos Santos et al., 2010)) and specific rule-based extraction for complex syntactic structures like the particular impersonal reflexive passive in Spanish. The output was subsequently cleaned to isolate essential linguistic data for the final analysis.

Large Language Models Selection and Prompt Strategies. We selected three LLM (GPT-5, Gemini 2.5 Flash, and Llama 4 Maverick Instruct) to evaluate subjectivity annotation using default parameters [1]. To address the need of linguistically oriented prompts, we adopted three strategies derived from Liu et al. (2023): Decomposition (direct questioning), Composition (chain-of-thought), and Augmentation (few-shot learning). Each strategy employed Spanish prompts that established the model's role as a linguistic annotator to test performance under different levels of provided information and instructional conditions, as detailed in Appendix A.1.

[1]This pilot study was performed using the web interfaces of the models and manually collecting the answers.

4 Experimental Design and Error Analysis

This section outlines the transition from the pilot study to the main experiment. By analyzing error patterns from initial findings, the annotation guidelines and hybrid framework were iteratively refined, resulting in a consolidated methodology.

4.1 Pilot Study

The primary goal of the pilot study is to establish a performance baseline for the initial hybrid framework (defined in Section 3.2) and to identify weaknesses in the preliminary annotation guidelines and LLM prompt strategies. The experiment involved applying the pilot hybrid framework to the pilot dataset (Section 3.1) before scaling the approach to the main experiment. To assess the linguistic reasoning capabilities of the models, the framework was executed across 9 different configurations, generated by fully crossing our three selected LLMs with the three distinct prompting techniques employed. For evaluation, the system's output was measured against the human annotations using three key metrics: Precision, Recall, and F1-Score.

4.1.1 Pilot Study Results

As shown in Table 2, Gemini achieved the best performance among the LLMs with an F1 of 0.81, though its precision was 0.73. However, our traditional ML pipeline (spaCy) surpassed the language models, achieving a higher F1-score (0.83) while maintaining superior precision (0.87). This difference is attributed to a key factor in error patterns: spaCy primarily presents omission errors (false negatives), whereas LLMs commit both omission and addition errors (false positives).

4.2 Error Analysis and Framework Refinement

Following the pilot study, we conducted a systematic error analysis. These insights guided an iterative refinement phase, where both the annotation guidelines and the hybrid architecture were optimized to produce the final consolidated framework used in the main experiment.

The unit of analysis consists of text spans, including words and multi-word expressions. For instance, in the sentence:

Model	Precision	Recall	F1
Human	1.00	1.00	1.00
GPT	0.72	0.48	0.57
Gemini	**0.73**	**0.90**	**0.81**
Llama	0.57	0.70	0.63
spaCy	**0.87**	**0.80**	**0.83**

Table 2: Performance Metrics for the different models used in the pilot study, using *Human* as reference. The presented scores reflect the average performance across diverse prompt techniques, given that none of the individual templates achieved relevant statistical improvements.

(1) *El contrato (...) ha* ***hecho*** *saltar las alarmas.* [The contract (...) has **made** alarm bells ring.]

The pipeline incorrectly tagged *"hecho"* as a generalizing term. In Spanish this is an homograph *"hecho"*, which can function as the participle of the verb *hacer* [to do/make] or the noun *fact*. Conversely, in the following example:

(2) *Los* ***hechos*** *han sucedido en Estados Unidos(...)* [The **facts** have occurred in the United States(...)]

The system correctly identifies *"hechos"* [facts] as a generalizing term.

Results presented in Table 3 enable us to draw key conclusions. Traditional ML (spaCy) proved more effective for categories conveying lower author involvement, while Gemini was superior for categories presenting higher author involvement. This performance disparity is attributed to the frequency of the word classes: spaCy captures common linguistic structures, such as generalizing terms (Non-specificity) and third-person pronouns (Distancing), while Gemini is needed for less frequent, complex forms like verbs of knowledge (Certainty) or precise numerical expressions (Specificity). Consequently, this analysis validated the need of the hybrid approach by demonstrating methodological complementarity.

Refinements to the traditional ML pipeline focused on mitigating systematic errors in lexical and morphological tags. The linguistic analysis processes were enhanced to detect and process the full range of conjugated forms for specific verbs and variational forms within the designed lexicons, resulting in a significant reduction in omission errors. Furthermore, rules for complex periphrastic structures (like the impersonal reflexive passive and passive voice) were expanded using windowing techniques to accurately capture multi-word constructions.

Category	Best Model	Prec.	Rec.	F1
Low Author Commitment				
Doubt	spaCy	0.69	0.84	0.75
Non-specificity		0.76	0.84	0.80
Distancing		0.88	0.88	0.88
High Author Commitment				
Certainty	Gemini	0.87	0.90	0.88
Specificity		0.94	0.99	0.96
Participation		0.95	0.95	0.95

Table 3: Performance metrics by subjectivity category, presenting Low Author Commitment in the first section and High Author Commitment in the second section.

The revision of the pilot study outputs also conducted to the integration of novel specific linguistic elements during the refinement of the annotation pipeline. First, we incorporated the modal periphrasis of the infinitive (e.g., forms expressing obligation like *haber de* + infinitive [should] or *tener que* + infinitive [must]); which was essential for capturing the author's intention as these structures explicitly convey obligation. Second, we performed specific lexical augmentation aimed at refining the existing lexicons, such as expanding the initial linguistic analysis processes for the Doubt category to ensure the accurate detection of conditional adverbs (e.g. *(quizás*, *tal vez* [maybe], etc.).

After implementing the initial refinement measures, we conducted a manual qualitative analysis to find the potential gaps in our annotation methodology. Despite the improvements, our error analysis confirmed that four complex linguistic phenomena remained challenging for our pilot hybrid framework, presented in Table 4.

Due to these persistent errors, a more curated approach using LLMs (Gemini, in this case) became necessary to achieve resolution. Furthermore, given that no single prompt template provided a statistically significant advantage over the others, a prompt ensembling strategy was adopted.

4.3 Consolidated Framework and Full Dataset

Following the refinements driven by the comprehensive error analysis, the consolidated

Linguistic Phenomenon	Description and Examples
Subjunctive Mood	Analysis reveals conditional verbs incorrectly tagged as subjunctive (e.g., *podría* [could]). Additionally, false positives occur due to morphological similarities between non-verb endings and subjunctive inflections (e.g., *Leverkusen*) or errors during lemmatization (e.g., assigning *quedarar**, which does not exist, as the lemma for *quedaremos* [we will stay]).
Imperative Mood	This mood is frequently misclassified due to its ambiguous forms, which often overlap with the third-person singular indicative (e.g., *piensa* [he/she thinks]).
First-person Verbs	These forms exhibit ambiguity due to their morphological overlap with third-person forms in certain tenses, such as the imperfect or present subjunctive (e.g., *hubiera* [had], *fuera* [were/was], *tenga* [have]).
Named Entities	Initial capitalized words in sentences are erroneously tagged as named entities despite of their semantic function (e.g., *Habrá* [there will be], *Riesgos* [risks]).

Table 4: Summary of linguistic phenomena leading to automated annotation errors: description of linguistic phenomena and representative examples.

hybrid framework integrates the adjusted prompt ensembling and the revised ML analysis processes.

Full Dataset. To validate this final architecture, we scaled the pilot dataset to create the full dataset, which statistics are included in Table 5. This expansion represents an increase of 74.18% in tokens and almost doubles the pilot dataset sentence count. This final dataset was also subjected to a dynamic annotation methodology to ensure the optimization of the final guidelines.

Metric	Value
Sentences	626
Tokens	17,437
Avg. Tokens / Sentence	27.85

Table 5: Full dataset statistics presenting sentences, tokens and average tokens per sentence.

Consolidated Hybrid Framework. The final framework integrates a traditional ML pipeline with Gemini 2.5 Flash (Gemini Team, 2025) through a robust prompt-ensembling strategy. To optimize annotation quality for complex cases, we implemented a single meta-prompt that synthesizes three techniques: Decomposition for stage segmentation, Composition for chain-of-thought reasoning, and Augmentation for few-shot contextualization (see Appendix A.2). This approach aligns with recent findings on the benefits of detailed guidance and diverse prompting strategies in LLMs (Kasner et al., 2025). The system was executed on Google Colab[2] using free API tier; due to rate limits, annotation was completed over several days.

Methodological enhancements were implemented to address the limitations identified during the pilot study. Specifically, the initial human vs. model comparison was replaced by a semi-automatic workflow that integrates automated tagging with expert human revision. We utilize three distinct annotation setups for this comparative analysis:

- **A1: Raw Manual Annotation:** Annotation produced by a human expert without subsequent revision.
- **A2: Unrevised Automatic Annotation:** Annotation produced automatically by the consolidated hybrid framework without any post-processing or quality filtering.
- **A3: Consolidated Semi-Automatic Annotation (Expert-Validated Reference)**: This configuration represents the final annotation derived from the hybrid framework's output, systematically refined through expert manual revision. By integrating automated tagging with human oversight, we establish a robust reference standard that mitigates models bias and ensures high annotation reliability. Consequently, this consolidated output serves as the definitive ref-

[2] `https://colab.google/`

erence benchmark for all subsequent performance evaluations within this study.

Each annotation setup (A1, A2, A3) functions as a distinct "annotator" in this evaluation. While the hybrid setups (A2 and A3) are not intended to replace the necessary quality of human annotators, we present a hybrid option to achieve comparable quality given the task's complexity and limited expert annotators availability. This landscape justifies the implementation of a hybrid approach that relies on ML, LLMs, and human expertise. This approach allows for a comprehensive assessment of quality, speed, and agreement across all setups, enabling a direct comparison between purely manual and semi-automatic annotation methods.

5 Validation of the Consolidated Hybrid Framework: Results and Analysis

This section presents the findings obtained from applying the consolidated hybrid framework to the full dataset. The analysis offers a evaluation of the framework across three dimensions: quantitative quality metrics, inter-annotator reliability against human expert, and computational efficiency and resource consumption. The measurement of diverse evaluation aspects allows us to construct a final conclusion that validates the best methodology in terms of quality and efficiency.

5.1 Performance and Quality Metrics

Our primary objective is to evaluate system performance against the expert-validated reference established by the consolidated hybrid annotation (A3). The designation of A3 as the reference benchmark is justified by two factors: (1) its status as the most comprehensive annotation outcome, resulting from expert manual revision of the hybrid output, and (2) the lack of pre-existing benchmarks for Spanish subjectivity detection. Notably, the achievement of perfect performance metrics for the A3 setup is an inherent result of the experimental design, as this configuration serves as the study's reference point. These results, therefore, do not imply an absence of error or model overfitting; instead, they reflect the framework's alignment with the expert-validated consensus used to establish the gold standard.

Framework performance is evaluated using three standard NLP metrics: Precision, Recall, and F1-score.

Model	Precision	Recall	F1
Pilot Study			
Human	1.00	1.00	1.00
GPT	0.72	0.48	0.57
Gemini	0.73	0.90	0.81
Llama	0.57	0.70	0.63
spaCy	0.87	0.80	**0.83**
Main Experiment ***(using Pilot Dataset)***			
Manual Raw (A1)	1.00	0.89	0.94
Hybrid Unrevised (A2)	1.00	0.77	0.87
Consolidated Revised (A3)	1.00	1.00	1.00
Main Experiment ***(using Full Dataset)***			
Manual Raw (A1)	0.91	0.91	0.91
Hybrid Unrevised (A2)	0.80	0.90	0.84
Consolidated Revised (A3)	1.00	1.00	1.00

Table 6: Consolidated performance metrics comparing pilot study, main experiment (using pilot dataset) and main experiment (using full dataset).

Table 6 presents the performance metrics of different annotation setups. The pilot study results presented on the first block of the table serve only as a visual summary, primarily emphasizing the relationship between the two methodological phases, given that the raw scores are not directly comparable due to the differences in the pipeline structure. Moreover, to preserve result comparability, performance metrics for the main experiment were calculated using the pilot dataset across both evaluated setups (A1, A2, A3), reported in the second block of Table 6. We also report the final results using our consolidated hybrid framework on the full dataset in the last block of the same table.

F1-Score and Recall directly quantify the loss of efficacy compared to the expert-validated reference (A3). The A1 (raw manual annotation) F1 of 0.91 implies a 9% efficacy loss, which establishes the maximum expected human quality before exhaustive revision. In contrast, the A2 (unrevised hybrid annotation) F1 of 0.84 sets the system's baseline performance, where the 16% efficacy

loss (compared to A3) is observed. Recall also measures the linguistic coverage versus missed instances (false negatives), precisely quantifying the required human review effort. For example, the A2 (hybrid unrevised) Recall of 0.90 shows that 10% of relevant annotations were missed, which is the direct measure of required human correction.

5.2 Inter-Annotator Reliability

To evaluate our hybrid approach, we will also calculate inter-annotator agreement (IAA) to measure not only the correctness of the different system's output but also its degree of similarity with both manual and revised semi-automatic annotation methods.

We selected Krippendorff's α as the metric for measuring inter-annotator reliability due to its statistical power and versatility(Klie, Castilho, and Gurevych, 2024). However, applying it to complex subjective analysis requires caution, as high agreement does not guarantee high quality, needing complementary manual validation (Braylan, Alonso, and Lease, 2022). For interpretation, we apply the standard thresholds established by (Krippendorff, 2004): α above 0.8 is considered highly reliable, 0.7 to 0.8 represents moderate agreement, and below 0.7 is reserved for tentative conclusions.

Annotators Pair	**Krippendorf's α**	**Reliability**
A1 vs. A2	0.72	Moderate
A2 vs. A3	0.60	Low
A1 vs. A3	0.89	High

Table 7: Inter-Annotator Agreement results using Krippendorf's α (Krippendorff, 2004).

The results presented in Table 7 demonstrate a moderate agreement (α above 0.70) between the raw manual annotation (A1) and the unrevised hybrid annotation (A2). This level of reliability partially validates the system's ability to act as a synthetic annotator, producing output comparable to that of a non-expert human. However, the most relevant interpretation in this landscape is that the implication of human revision on the hybrid framework output (A3) increases the agreement to a strong level (α above 0.80), confirming the consolidated hybrid approach as the most robust and consistent methodology.

5.3 Efficiency and Time Consumption

To directly evaluate the efficiency of our proposed annotation framework, we analyze the time consumption of all setups to assess the practical utility of the hybrid approach. This evaluation identifies the optimal configuration by balancing performance and reliability against operational costs.

Metric	**A1**	**A2**	**A3**
Avg. Time per Sentence (sec)	50.95	2.00	9.65
Avg. Tokens per Sentence	27.81	27.81	27.81
Speed (sec/token)	1.83	0.07	0.35
Relative Efficiency (vs. A1)	1.00	25.48	5.28

Table 8: Efficiency metrics comparison across annotation setups. Metrics include average time per sentence, processing speed (sec/token), and efficiency gains relative to manual annotation (A1).

Table 8 presents the efficiency results, measured as the average time per sentence in seconds against the raw manual annotation (A1) baseline. While the automated setup (A2) is the fastest, the hybrid A3 method offers a significant efficiency gain over traditional manual annotation (A1), reducing total processing time by 81%.

Despite the advantages of LLMs, their deployment implies significant drawbacks, most notably a relevant environmental impact. The high energy demands of model inference contribute to a considerable carbon footprint, presenting a critical ethical and sustainability challenge that must be considered against the framework's efficiency gains.

5.4 Discussion

This section synthesizes the quantitative results and performs a qualitative analysis to assess the overall efficiency and quality of the consolidated hybrid framework.

As illustrated in the diverging bar chart (Figure 2), the performance balance validates A3 as the most cost-effective and high-quality solution. A3 achieves the optimal balance by delivering high quality while maintaining near-maximum efficiency (minimal loss compared to the fastest setup, which is A2).

This final evaluation confirms the framework's reliability in terms of quality and efficiency gains. The F1-Score quantifies the loss

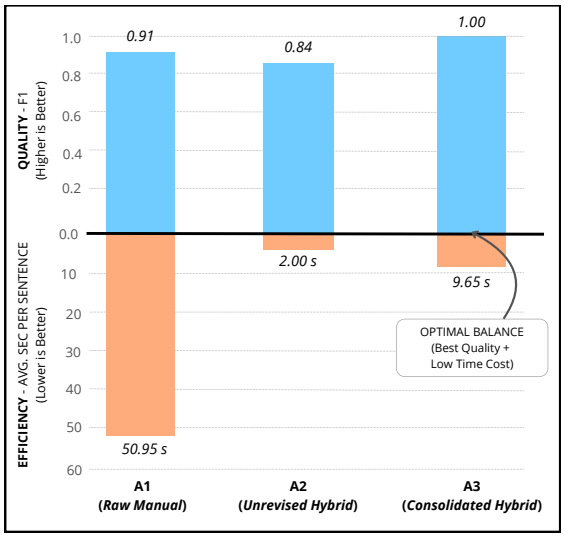

Figure 2: Diverging Bar Chart illustrating the trade-off between annotation quality measured using F1-score (*top*) and efficiency measured using average seconds per sentence (*bottom*).

of efficacy relative to the expert-validated reference (A3). For instance, the A2 (unrevised hybrid annotation) F1 of 0.87 implies a 13% efficacy loss relative to the A3 expert-Validated reference (F1=1.00). This loss requires human intervention, which is quantified by the A2 Recall of 0.77. This score indicates that 23% of relevant instances were missed by the automated system, directly measuring the required subsequent human correction effort. Ultimately, comparing the time consumption of A1 and A2 with their respective Recall scores allows us to determine the optimal balance between maximizing coverage and minimizing the overall time cost. While the automated A2 setup is 25.48 times faster than manual annotation (A1), the consolidated hybrid annotation (A3) proves to be the most effective option in terms of quality and efficiency, achieving perfect quality at a time cost that is only 19% of the pure manual effort.

6 Conclusions and Future Work

This paper developed and validated a novel hybrid annotation framework for subjectivity annotation of linguistic markers in Spanish. This methodology notably integrates traditional ML efficiency with LLM reasoning, achieving annotations of human-comparable quality in a significantly shorter period of time. The pipeline was optimized through a preliminary pilot study and subsequent exhaustive error analysis, eventually confirming the framework's reliability in establishing the primary objective: an optimal balance between efficiency and quality, while also providing a transparent and easily traceable framework.

The consolidated hybrid annotation (A3) method is presented as the most practicable option for high-quality annotation generation. It achieves high quality at a time cost that is only 19% of the pure manual effort (A1). The F1-Score helps us quantify the required manual effort: for instance, the 23% efficacy loss observed in the unrevised setup (A2) sets the minimum required manual review necessary to reach A3 quality. This findings lead to three key contributions: the consolidated hybrid framework, a comprehensive set of refined annotation guidelines, and the creation of a refined and validated annotated dataset for Spanish SD.

The demonstrated efficacy of this framework also reveals several limitations that establish interesting directions for future research. For instance, our methodology should be applied to new domains where subjectivity information is relevant, such as hate speech or specialized discourse, including political or advertising language. In this cases, the analysis of linguistic subjectivity is implicated in public opinion or client persuasion. Furthermore, future work needs to investigate the LLMs' ability to efficiently capture technical terms (technicisms) as specific linguistic markers within these diverse domains. Addressing this unresolved challenge is essential for advancing the framework's optimization and generalization potential across different domains.

Finally, to strengthen the statistical justification for LLM deployment, we plan to increase the number of human annotators to implement the Alt-Test (Calderon, Reichart, and Dror, 2025), introduced in Section 2. This metric is relevant because it provides robust, actionable statistical evidence regarding LLM viability, as it intrinsically accounts for inter-human variability.

Acknowledgements

This research is funded by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social Europeo Plus of Generalitat Valenciana -

European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects; "SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP" (AIA2025-163322-C63); "Mecánica cuántica para la comprensión y generación del lenguaje" (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE; Proyecto Desarrollo de Modelos ALIA within the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 of the Ministerio para la Transformación Digital y de la Función Pública y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; "Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)", developed in the II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, with the topic "Humanidades Digitales" (Referencia: FRAHUMANIDADES25-01), funded by Fundación Ramón Areces.

References

Akkurt, F., O. Güngör, B. Marşan, T. Güngör, B. Ö. Başaran, A. Özgür, and S. Üsküdarlı. 2024. Evaluating the quality of a corpus annotation scheme using pretrained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6504–6514.

Antici, F., L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, and F. Ruggeri. 2021. Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 40–52. Springer.

Bajtin, M. 1982. *Estética de la creación verbal*. Siglo XXI Editores.

Bonet-Jover, A., R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, A. Piad-Morffis, and S. Estevez-Velarde. 2023. Applying human-in-the-loop to construct a dataset for determining content reliability to combat fake news. *Engineering Applications of Artificial Intelligence*, 126:107152.

Braylan, A., O. Alonso, and M. Lease. 2022. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, pages 1720–1730.

Bybee, J. and S. Fleischman. 1995. Modality in grammar and discourse: An introductory essay. *Modality in grammar and discourse*, 14:503–517.

Calderon, N., R. Reichart, and R. Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*.

Da San Martino, G., S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov. 2019. Fine-grained analysis of propaganda in news articles. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China, November. Association for Computational Linguistics.

Dynel, M. 2023. Lessons in linguistics with chatgpt: Metapragmatics, metacommunication, metadiscourse and metalanguage in human-ai interactions. *Language & Communication*, 93:107–124.

Gemini Team, G. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*. Release date derived from search results: July 2025.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Hyland, K. 2005. Metadiscourse: Exploring interaction in writing. *Continuum*.

Jahan, M., H. Wang, T. Thebaud, Y. Sun, G. H. Le, Z. Fagyal, O. Scharenborg, M. Hasegawa-Johnson, L. M. Velazquez, and N. Dehak. 2024. Finding spoken identifications: Using gpt-4 annotation for an efficient and fast dataset creation pipeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Re-*

sources and Evaluation (LREC-COLING 2024), pages 7296–7306.

Kasmuri, E. and H. Basiron. 2017. Subjectivity analysis in opinion mining—a systematic literature review. *Int J Adv Soft Comput Appl*, 9(3):132–159.

Kasner, Z., V. Zouhar, P. Schmidtová, I. Kartáč, K. Onderková, O. Plátek, D. Gkatzia, S. Mahamood, O. Dušek, and S. Balloccu. 2025. Large language models as span annotators. *arXiv preprint arXiv:2504.08697*.

Klie, J.-C., R. E. d. Castilho, and I. Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Koo, R., M. Lee, V. Raheja, J. I. Park, Z. M. Kim, and D. Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand, August. Association for Computational Linguistics.

Krippendorff, K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Krüger, K. R., A. Lukowiak, J. Sonntag, S. Warzecha, and M. Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), January.

Nasution, A. H. and A. Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *Ieee Access*, 12:71876–71900.

Ott, M., Y. Choi, C. Cardie, and J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Pang, B., L. Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Posadas-Durán, J.-P., H. Gómez-Adorno, G. Sidorov, and J. J. M. Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943.

Schroeder, H., D. Roy, and J. Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. *arXiv preprint arXiv:2507.15821*.

Shokri, M., V. Sharma, E. Filatova, S. Jain, and S. Levitan. 2024. Subjectivity detection in english news using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 215–226.

Stureborg, R., D. Alikaniotis, and Y. Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Vaamonde Dos Santos, G., F. González, J. M. García-Miguel, et al. 2010. Adesse. a database with syntactic and semantic annotation of a corpus of spanish.

Vieira, L. L., C. L. M. Jeronimo, C. E. C. Campelo, and L. B. Marinho. 2020. Analysis of the subjectivity level in fake news fragments. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 233–240, New York, NY, USA. Association for Computing Machinery.

Voutsa, M. C., N. Tsapatsoulis, and C. Djouvas. 2025. Biased by design? evaluating bias and behavioral diversity in llm annotation of real-world and synthetic hotel reviews. *AI*, 6(8):178.

Wang, S., X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, and C. Guo. 2025. GPT-NER: Named entity recognition via large language models. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albu-

querque, New Mexico, April. Association for Computational Linguistics.

Xia, M., S. Maharjan, T. Le, W. Taylor, and M. Song. 2025. Syncode: Synergistic human–llm collaboration for enhanced data annotation in stack overflow. *Information*, 16(5):392.

Zorraquino, M. A. M. 1999. Aspectos de la gramática y de la pragmática de las partículas de modalidad en español actual. In *Español como lengua extranjera, enfoque comunicativo y gramática: actas del IX congreso internacional de ASELE. Santiago de Compostela, 23-26 de septiembre de 1998*, pages 25–56. Servicio de Publicaciones= Servizo de Publicacións.

A Appendix 1: Prompting Techniques

A.1 Prompting Techniques used in Pilot Study

Following the approach by Liu et al. (2023), we implement three distinct prompting techniques: Decomposition, Composition, and Augmentation. Table 9 presents the specific prompts utilized for these techniques in their original Spanish form, alongside English translations provided solely for clarity in this article. 'Doubt' category is used as an example and this structural framework was replicated consistently across all other subjectivity categories.

A.2 Prompt Ensembling used in Consolidated Hybrid Annotation Framework

As detailed in Section 4.3, the consolidated framework utilizes a meta-prompting strategy that ensembles Decomposition, Composition, and Augmentation techniques. This approach facilitates complex annotation by leveraging segmented reasoning and contextual few-shot learning to improve model performance. The specific configuration of this prompt ensembling is provided in Table 10.

Prompt Technique	Spanish Prompt (Original)	English Translation
Prompt Decomposition	*Tienes que actuar como un analizador lingüístico. Tu objetivo es extraer tanto palabras como conjuntos de palabras o signos de puntuación que indiquen DUDA, expresados verbalmente a través de Interrogaciones retóricas (concretamente representada con el signo de interrogación '?'), Verbos en condicional, Verbos en modo subjuntivo, Verbos de creencia según su significado, y Pronombres interrogativos. Realiza el análisis del texto y devuélveme el texto anotado señalando entre corchetes angulares <> el fragmento del texto que se anota y entre corchetes normales [ ] la categoría DUDA. El texto a analizar es el siguiente:*	You must act as a linguistic analyzer. Your goal is to extract words, sets of words, or punctuation marks that indicate DOUBT, expressed through rhetorical questions (specifically the '?' mark), conditional verbs, subjunctive verbs, verbs of belief, and interrogative pronouns. Analyze the text and return the annotated text using angle brackets <> for the annotated fragment and square brackets [ ] for the DOUBT category. The text to analyze is:
Prompt Composition	*Tienes que actuar como un analizador lingüístico. Tu objetivo es extraer tanto palabras como conjuntos de palabras o signos de puntuación que indiquen DUDA. Paso 1: Buscar signos de interrogación ('?'). Paso 2: Localizar verbos en condicional. Paso 3: Detectar verbos en subjuntivo. Paso 4: Marcar verbos de creencia (creer, parecer, dudar, etc.). Paso 5: Extraer pronombres interrogativos. Paso 6: Realizar el análisis y devolver el texto anotado con <> y [DUDA]. El texto a analizar es el siguiente:*	You must act as a linguistic analyzer. Your goal is to extract words, phrases, or punctuation indicating DOUBT. Step 1: Search for '?' marks. Step 2: Locate conditional verbs. Step 3: Detect subjunctive verbs. Step 4: Mark verbs of belief (creer, parecer, dudar, etc.). Step 5: Extract interrogative pronouns. Step 6: Perform the analysis and return the annotated text using <> and [DOUBT]. The text to analyze is:
Prompt Augmentation	*Tienes que actuar como un analizador lingüístico. Tu objetivo es anotar palabras o signos que indiquen DUDA. Ejemplos: 1. 'Si fuera... habría...' (subjuntivo y condicional). 2. '¿Por qué...?' (pronombre interrogativo). 3. '¿...?'. (pregunta retórica). 4. '...parece...' (verbo de creencia). Realiza el análisis del texto y devuélveme el texto anotado señalando entre corchetes angulares <> y entre corchetes normales [ ] la categoría DUDA. El texto a analizar es el siguiente:*	You must act as a linguistic analyzer. Your goal is to annotate words or signs indicating DOUBT. Examples: 1. 'Si fuera... habría...' (subjunctive/conditional). 2. '¿Por qué...?' (interrogative pronoun). 3. '¿...?'. (rhetorical question). 4. '...parece...' (verb of belief). Analyze the text and return it annotated with angle brackets <> and square brackets [ ] for the DOUBT category. The text to analyze is:

Table 9: Prompt techniques for subjectivity annotation used in the pilot study, using 'Doubt' category as an example.

Prompt Technique	Spanish Prompt (Original)	English Translation
Prompt Ensembling	*Actúa como experto lingüista. Detecta verbos en modo subjuntivo en el texto.* *PROCESO: 1. Análisis Morfológico: Identifica verbos en subjuntivo (ej. 'fuera' vs adverbio 'fuera'). 2. Análisis Semántico: Busca verbos que expresen deseo, duda, posibilidad o hipocresía. 3. Consolidación de resultados: Cruza ambos análisis para eliminar falsos positivos.* *REGLAS: Extrae la lista con etiqueta <* $SUBJUNTIVOS$ *>. Formato: verbo <* DUD *>. Si no hay: <* $SUBJUNTIVOS$ *> N/A <* $/SUBJUNTIVOS$ *>. Sin explicaciones.* *TEXTO: [texto_a_analizar]*	Act as an expert linguist. Detect verbs in the subjunctive mood within the text. PROCESS: 1. Morphological Analysis: Identify subjunctive verbs (e.g., 'were' vs other forms). 2. Semantic Analysis: Look for verbs expressing desire, doubt, possibility, or hypocrisy. 3. Result Consolidation: Cross-reference both analyses to eliminate false positives. RULES: List with <* $SUBJUNCTIVES$ *> tag. Format: verb <* $DOUBT$ *>. If none: <* $SUBJUNCTIVES$ *> N/A <* $/SUBJUNCTIVES$ *>. No explanations.* TEXT: [text_to_analyze]

Table 10: Prompt used for subjectivity annotation applying prompt ensembling in the consolidated hybrid framework, using 'Doubt' category as an example.