Pragmatic Profiling for Disinformation Detection: An Exploratory Analysis of Stylistic Features in Spanish News

Alba Pérez-Montero[1], **María Miró Maestre**[2], **Elena Lloret Pastor**[2] **and Paloma Moreda Pozo**[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 maria.miro@ua.es {elloret, moreda}@dlsi.ua.es

Abstract

The spread of misleading and fabricated information has made automatic disinformation detection a central challenge for Natural Language Processing. While most approaches have focused on lexical, syntactic, or semantic cues, deceptive discourse is also shaped by pragmatic choices that reflect how information is framed, qualified, and directed toward readers. This paper investigates whether pragmatic information can improve disinformation detection in Spanish by explicitly incorporating two annotation layers: communicative intentions and subjectivity markers. We compare the impact of these pragmatic features across three modeling paradigms: traditional classifiers, an encoder-based transformer (RoBERTa), and generative language models (GPT-oss and Mistral-small). Our results show that pragmatic augmentation consistently improves over text-only baselines, with subjectivity markers displaying stronger discriminative power than intention labels. Statistical testing further confirms that the observed gains are robust for the generative models evaluated. These findings support the view that authorial stance and communicative purpose provide useful complementary evidence for veracity classification.

1 Introduction

The rapid circulation of digital content has amplified the visibility and impact of misleading and fabricated information, making automatic disinformation detection a key challenge for Natural Language Processing (NLP) and Computational Linguistics (Zhou and Zafarani, 2020). Although manual fact-checking remains essential, it is costly, time-consuming, and difficult to scale (Hassan et al., 2017). Text-based detection methods therefore offer a valuable complementary approach by identifying linguistic patterns associated with deceptive content without relying entirely on external verification (Shu et al., 2017).

This study examines disinformation detection in Spanish from a pragmatic perspective, asking whether pragmatic information improves classification. It assumes that deception is reflected not only in propositional content, but also in how information is oriented toward readers. In particular, metadiscourse captures how writers organise discourse and manage their relationship with readers (Hyland, 2005; Dynel, 2023), while epistemic modality reflects the degree of commitment authors express toward their claims (Zorraquino, 1999).

The main hypothesis is that pragmatic information provides relevant evidence for disinformation detection, complementing lexical, semantic, and syntactic representations that may not fully capture stance or communicative positioning. In this study, pragmatic information is operationalised through two dimensions: (i) communicative intentions and (ii) subjectivity markers. The task is formulated as a binary classification problem in which each text is labeled as TRUE or DISINFORMATION. An ablation-based design is used to evaluate the individual and combined contribution of linguistic and pragmatic information, making it possible to isolate the effect of intention-related and subjectivity-related features on classification performance.

The study also compares three modeling paradigms: traditional machine learning classifiers, an encoder-based transformer model (RoBERTa), and generative language models. This comparison makes it possible to assess not only whether pragmatic features improve detection, but also whether their contribution varies across model families. This comparative analysis makes it possible to assess whether pragmatic cues are more effectively exploited by feature-based approaches, contextual encoders, or generative models.

The main contributions of this paper are the following:

- A pragmatically informed approach to disinformation detection in Spanish, centered

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

on communicative intentions and subjectivity markers.

- A comparative evaluation across three modeling paradigms: traditional classifiers, an encoder-based transformer, and generative language models.
- An ablation-based framework to measure the contribution of pragmatic features relative to more conventional linguistic information.

The remainder of this paper is organised as follows. Section 2 reviews prior work on disinformation and on the evolution of automatic classification systems. Section 3 presents the dataset and the pragmatic annotation schemes adopted. Section 4 describes the experimental design, including feature settings, models, and the evaluation framework. Section 5 reports and discusses the descriptive and classification results. Finally, Section 6 concludes the paper and outlines future work.

2 Related Work

2.1 Disinformation and Information Disorder

The digital landscape has transformed from a space of open communication into a complex "information disorder", where disinformation is strategically deployed to deceive or cause harm (Wardle and Derakhshan, 2017). Disinformation is commonly understood as fabricated content that imitates the format of traditional news while lacking its ethical standards and organisational intent (Lazer et al., 2018). Its spread is often driven by political or economic incentives that exploit the high engagement potential of sensationalist and viral content (Saquete et al., 2020), as well as by individuals' tendency to accept and share information that reinforces partisan identities (Lazer et al., 2018).

Because the current volume of digital information exceeds human verification capacities, NLP has become essential for the automatic detection of unreliable claims (Saquete et al., 2020). As noted by (Yuan et al., 2024), detection technologies have progressed through three major stages: Machine Learning, based on manually engineered linguistic and user features; Deep Learning, which enabled the automatic extraction of complex semantic patterns; and pre-trained models, including Transformer-based architectures such as BERT and GPT, which capture broader contextual information and improve detection accuracy. At the same time, the scale of online content has increased dependence on algorithmic curation, fostering "filter bubbles" that isolate users from corrective information and intensify polarisation. In response, recent research has moved beyond isolated fact-checking toward triangular communication systems that combine human expertise, computational methods, and linguistic theory (Yuan et al., 2024). The framework proposed in this paper follows this explainable and multi-layered perspective to support transparent and sustainable disinformation detection.

2.2 Evolution of Computational Strategies for Disinformation Detection

Disinformation detection has been primarily approached in NLP as a text classification task, enabling models to distinguish between verified news and deceptive content based on linguistic patterns rather than manual fact-checking alone. Whereas early methods relied on metadata and simple lexical features, current approaches increasingly use deep learning architectures that capture more complex semantic and pragmatic signals in deceptive text (Pérez-Rosas et al., 2018).

Transformer-based models have significantly advanced this task. In Spanish, BETO, a BERT model trained on Spanish corpora, has shown strong performance in capturing language-specific cues, and BERT-based systems have achieved solid results in related binary classification settings (González-López et al., 2021). Other studies have highlighted the value of pre-trained BERT encoders for multitask classification, as well as the importance of implementation choices such as label encoding, stratified data handling, [CLS]-based classification, and hyperparameter optimization (Gallardo-Hernández et al., 2025; Martín-Hoz et al., 2025). These findings are consistent with results reported for RoBERTa, which reached 92.12% accuracy and an 89.47% F1-score with a batch size of 16 and a learning rate of 3×10^{-5} (Raza et al., 2025).

More recently, Large Language Models (LLMs) have introduced a different classification paradigm by using generative capabilities to support pragmatic reasoning. Although BERT-like models often perform better in controlled classification settings, LLMs such as Llama and Mistral 7B appear more robust under adversarial conditions and perturbations (Raza et al., 2025). They are also increasingly integrated into Generative AI workflows that use synthetic data and majority voting to improve label

stability, while requiring fewer training iterations than encoder-based models before overfitting becomes a concern.

Despite these advances, most computational approaches still prioritise lexical and semantic information, with limited attention to pragmatic dimensions. The present study addresses this gap by incorporating communicative intentions and subjectivity markers as explicit annotation layers, complementing the representational strategies reviewed above.

3 Data Framework and Pragmatic Annotation

3.1 News Dataset

This study builds on two corpora of Spanish news texts: *RUN corpus* (Bonet-Jover et al., 2024) and *FakeNewsSpanishCorpus* (Posadas-Durán et al., 2019). These datasets have been previously annotated with binary veracity labels (FAKE/TRUE). The corpus serves as the base layer onto which intentions and subjectivity markers are added in order to examine whether pragmatic enrichment improves automatic disinformation detection.

Both source datasets originate from prominent shared tasks in the Hispanic NLP community focused on news reliability: the FLARES task at IberLEF 2024 (Sepúlveda-Torres et al., 2024) and the FakeDeS task at IberLEF 2021 (Gómez-Adorno et al., 2021). For the present study, a subset of 33 articles was selected and processed. While this represents a limited volume of full texts, adopting the sentence as the primary unit of analysis, following established practices (Antici et al., 2021), provided the granular resolution necessary for capturing localised pragmatic markers. This approach yielded a balanced dataset of 634 sentences (317 per class), where rigorous manual revision ensured a high-quality human baseline.

Although this sample size is more limited than the original source corpora, this selection prioritizes two critical objectives: ensuring a perfect 50/50 class distribution to eliminate evaluation skewness, and allowing for an exhaustive manual review of the high-quality pragmatic annotations. To validate the statistical significance of the results despite the sample scale, a p-value analysis is provided in the classification results (Section 5.6).

3.2 Pragmatic Annotation Schema

Beyond propositional content, disinformation contains stylistic traces that function as markers of deception (Yuan et al., 2024). This section utilises metadiscourse and epistemic modality to examine how authors manage their stance and signal intentions (Hyland, 2005). Such pragmatic layers provide essential evidence of the speaker's certainty and underlying perspective (Zorraquino, 1999).

Among the approaches proposed to automate the classification of communicative intentions, Speech Act Theory (SAT) has remained one of the most influential linguistic frameworks in NLP (Austin, 1962; Searle, 1969). Under this perspective, utterances do not merely convey information; they also perform actions whose interpretation depends on both linguistic form and context, such as requesting, or promising. Accordingly, these 'speech acts', namely, the actions performed by speakers through their utterances (Yule, 2022), can be understood as intention-driven communicative acts that are often recoverable through observable linguistic cues, whether these intentions are expressed directly or indirectly. To identify the intentions conveyed in the news excerpts analysed in this study, we adopt the taxonomy proposed by Maestre et al. (2025), which distinguishes 13 global intention categories: informative, personal opinion, suggestion, command, request, question, threat, promise, praise, criticism, emotional, desire, and sarcasm/joke. Although this taxonomy was originally designed for social media messages, we consider it suitable for the present study because the news excerpts in our corpus are comparable in length and communicative density to prototypical social media texts. This makes the framework a useful basis for examining how intention-related cues contribute to disinformation classification.

The intention labels were assigned automatically with GPT-oss. Each news excerpt was processed in a zero-shot setting through a prompt specifically designed to identify its predominant intention according to the taxonomy adopted in this work. To guarantee annotation consistency, the model was instructed to produce a structured output restricted to a single label from the predefined inventory of intention categories. This automatically predicted label was then added to the corpus as an additional pragmatic annotation layer and subsequently incorporated into the classification settings of the study. More information on the intention annotation pro-

cess can be consulted in Appendix A.

The resulting dataset is therefore not only a binary veracity corpus, but also a pragmatically enriched resource that enables both descriptive analysis and supervised classification experiments. This design makes it possible to compare the contribution of pragmatic features against text-only baselines under the same classification framework.

3.3 Subjectivity Annotation

Determining an author's underlying implication requires a deep analysis of pragmatic signatures, following the principle that all discourse is an extension of the speaker's internal state (Tuchman and Aladro Vico, 1998). Utilising the concepts of metadiscourse (Hyland, 2005) and epistemic modality (Zorraquino, 1999), we propose an annotation system based on six dimensions of subjectivity mapped onto an axis of commitment. This model distinguishes between markers that denote active authorial engagement and those indicating a reduction in truth-claim stability, allowing for a granular assessment of how information is framed regarding its source and perceived reliability.

The annotation scheme is divided into two macro-groups, starting with **high author commitment**, which encompasses *Certainty* through factual affirmation or future tense, *Specificity* via precise references and named entities, and *Participation* through explicit first- or second-person presence. Conversely, **low author commitment** is characterised by *Doubt* through hedging and subjunctive forms, *Non-specificity* via generic expressions and indefinite pronouns, and *Distancing*, where the author separates themselves from the propositional content through passive voice or impersonal third-person references. It is important to note that these dimensions are not mutually exclusive, as a single sentence can be annotated with several of them simultaneously. Consequently, the specific combination and prominence of these overlapping dimensions are what ultimately reveal the overall degree of subjectivity.

These dimensions were annotated via a hybrid semi-automatic pipeline that integrates spaCy-based morphosyntactic tagging and custom lexicons with the metalinguistic reasoning of Gemini 2.0. The resulting labels functioned as preliminary hypotheses for expert human revision. This synergy of computational efficiency and manual rigor facilitates granular analysis and ensures data quality.

4 Experimental Design

4.1 Pragmatic Profiling Methodology

Before classification, we descriptively analyse the pragmatic profiles of DISINFORMATION vs. TRUE news. This stage establishes interpretable baseline evidence and guides feature selection for the experimental models. This analysis is organised into three layers:

- **Type-of-news profiling**: We characterise the distinct 'linguistic signatures' of both DISINFORMATION and TRUE news groups. By comparing their label distributions, we identify which pragmatic intentions and subjectivity markers are over-represented in deceptive content versus factual reporting.
- **Inter-label correlations**: Using heatmap visualizations, we analyse the co-occurrence and correlation between intentions and subjectivity categories. This allows us to observe how specific communicative goals align with varying levels of author commitment.
- **Information-theoretic importance**: We employ Mutual Information (MI) to quantify the discriminatory power of each feature. This non-linear analysis identifies the most predictive 'red flags' within the pragmatic domain.

4.2 Classification Settings

To systematically assess the contribution of pragmatic information, we define eight experimental settings that progressively incorporate annotation layers. This ablation-based design allows us to isolate the effect of each pragmatic dimension both individually and in combination.

The evaluated settings are the following:

1. **Plain Text**: original content without pragmatic augmentation.
2. **Intentions**: text enriched with communicative intention labels.
3. **Full Subjectivity**: text enriched with the full subjectivity feature set.
4. **Low Author Involvement**: text enriched only with low-commitment markers.
5. **High Author Involvement**: text enriched only with high-commitment markers.

6. **Low Author Involvement + Intentions**: combination of low-commitment subjectivity markers and intentions.

7. **High Author Involvement + Intentions**: combination of high-commitment subjectivity markers and intentions.

8. **Full Subjectivity + Intentions**: the most comprehensive pragmatic augmentation setting.

4.3 Classification Models

We evaluate the eight settings across three modeling paradigms with fundamentally different inductive biases: traditional machine learning classifiers, the encoder-only transformer RoBERTa[1], and two generative large language models: GPT-oss (proprietary, 120B) (Achiam et al., 2023) and Mistral-small (efficient and language-optimized, 24B) (Mistral AI, 2024). This comparative design allows us to assess whether pragmatic features are more effectively exploited by feature-based, contextual, or generative architectures.

For the encoder-based experiments, annotated texts are converted into structured inputs in which pragmatic labels are appended using explicit delimiters. This strategy preserves linguistic continuity and respects the model's context window while enabling the transformer to process pragmatic information together with the source text.

Following Liu (2021), the generative models are evaluated in a zero-shot setting. Prompt design enforces a constrained binary output format so that the models return only the required TRUE/DISINFORMATION labels. This decision avoids bias from in-context examples and tests the models' inherent ability to leverage the pragmatically enriched inputs.

Traditional classifiers are included as a strong baseline. Their presence is important because they make it possible to assess whether more complex neural architectures actually provide an advantage for this particular dataset and task.

4.4 Evaluation Framework

To ensure that performance differences reflect genuine improvements rather than dataset artifacts, we adopt a rigorous evaluation framework. Results are reported using stratified 5-fold cross-validation, preserving class proportions and reducing partition bias.

For traditional and encoder-based models, we report Precision, Recall, and F1-score. For generative models, robustness and reproducibility are further assessed by repeating evaluations across four random seeds (10, 20, 30, and 40). These seeds initialise the model's stochastic processes, ensuring that the results are consistent and not a product of specific random initialisations. Finally, statistical significance is verified with paired-sample *t*-tests comparing text-only baselines against pragmatically enriched settings (Student, 1908).

This combination of cross-validation and significance testing ensures that the final interpretation rests not only on raw performance values but also on statistically grounded evidence.

5 Results and Discussion

5.1 Pragmatic Profile of News

We begin by examining the distributional properties of pragmatic annotations across disinformation and true news. This exploratory analysis provides interpretive grounding for the automatic results reported in the following subsections.

Table 1 presents the distribution of subjectivity and intention labels across both veracity groups. At the level of subjectivity, disinformation shows higher proportions of *Doubt*, *Non-specificity*, *Certainty*, and *Participation*, whereas true news is more strongly associated with *Specificity* and, to a lesser extent, *Distancing*. In the intention layer, the *Informative* label is almost evenly distributed across both classes, but categories such as *Criticism*, *Emotive*, *Praise*, and especially *Threat* appear more often in deceptive content.

To distinguish meaningful contrasts from random variation, we apply a Mann–Whitney U test. As shown in Table 2, the strongest effects are found in the subjectivity layer: *Specificity*, *Certainty*, and *Non-specificity* reach statistical significance. By contrast, intention categories do not cross the conventional significance threshold, although several of them reveal interesting distributional tendencies.

These results reveal a clear divide in authorial commitment. Within the low-commitment macrogroup, both *Doubt* and *Non-specificity* are more prevalent in disinformation, while *Distancing* is slightly more frequent in true news. Within the high-commitment group, deceptive texts rely more heavily on *Certainty* and *Participation*, whereas

[1]The experiments were built on bertin-project/bertin-roberta-base-spanish, a RoBERTa-based model for Spanish, available at: https://huggingface.co/bertin-project/bertin-roberta-base-spanish

Subjectivity	**Absolute (N)**		**Percentage (%)**	
Label	*Disin.*	*True*	*Disin.*	*True*
Doubt	165	124	57.1	42.9
Nonspecificity	294	221	57.1	42.9
Distancing	186	197	48.6	51.4
Certainty	360	246	59.4	40.6
Specificity	632	844	42.8	57.2
Participation	88	69	56.1	43.9
Intentions	**Absolute (N)**		**Percentage (%)**	
Informative	246	255	49.1	50.9
Pers. Opinion	19	15	55.9	44.1
Criticism	17	10	63.0	37.0
Praise	3	2	60.0	40.0
Emotive	8	5	61.5	38.5
Question	13	10	56.5	43.5
Suggestion	2	3	40.0	60.0
Request	7	3	70.0	30.0
Threat	2	0	100.0	0.0
Desire	2	3	40.0	60.0
Obligation	1	3	25.0	75.0

Table 1: Distribution of Pragmatic Labels by Group. Disin. = Disinformation.

Note: Labels with zero frequency (Command, Promise, Sarcasm) are excluded from the analysis.

true news is characterized by a higher degree of *Specificity*. This pattern suggests that deceptive discourse may combine assertive framing with lower factual precision.

To estimate feature importance independently of significance testing, we compute Mutual Information (MI) scores. Table 3 ranks the most informative pragmatic features. Among subjectivity markers, *Certainty* and *Specificity* stand out as especially relevant. In the intention layer, *Question*, *Obligation*, *Request*, and *Informative* contribute useful class-discriminative information, even when their raw distributional differences are less pronounced.

Finally, the correlation analysis of pragmatic features shown in Figure 1 provides a global view of how subjectivity and intentions interact. In particular, *Criticism* aligns positively with *Certainty* and shows a meaningful relationship with *Specificity*, suggesting that critical discourse in the corpus tends to be assertive and factually framed. By contrast, *Personal Opinion* and other subjective functions tend to cluster with markers of *Doubt* and *Non-specificity*. These patterns help explain why combined feature settings often improve downstream classification: they make it easier for models to distinguish between assertive, factually grounded language and ambiguous or strategically subjective discourse.

Cat.	**Label**	**p-value**	**Sig.**	**Stars**
SUB	Specificity	< 0.001	Yes	***
SUB	Certainty	0.0019	Yes	**
SUB	Non-specificity	0.0179	Yes	*
SUB	Participation	0.1284	No	
SUB	Doubt	0.1758	No	
INT	Informative	0.0716	No	
INT	Threat	0.1656	No	
INT	Criticism	0.2024	No	
INT	Request	0.2252	No	
INT	Obligation	0.2988	No	
INT	Pers. Opinion	0.5547	No	

*Note: Only top labels by p-value shown for Intentions.

Table 2: Statistical Significance of Linguistic Labels (Mann-Whitney U test). Abbreviations: INT: Intentions, SUB: Subjectivity.

Cat.	**Feature**	**MI**	**Median (T/D)**	**Higher in**
INT	Quest.	0.0468	0.000 / 0.000	—
SUB	Cert.	0.0403	0.024 / 0.043	Disin.
SUB	Part.	0.0340	0.000 / 0.000	—
INT	Oblig.	0.0249	0.000 / 0.000	—
SUB	Spec.	0.0192	0.099 / 0.075	True
INT	Req.	0.0185	0.000 / 0.000	—
INT	Pers. Opin.	0.0174	0.000 / 0.000	—
INT	Info.	0.0156	0.037 / 0.042	Disin.
INT	Threat	0.0154	0.000 / 0.000	—
INT	Emot.	0.0116	0.000 / 0.000	—

Table 3: Top 10 linguistic features ranked by Mutual Information (MI) in descending order. Median values for True (T) and Disinformation (D) texts are shown as Median (T/D), and the column "Higher in" indicates the class with the higher median. Feature names are abbreviated to fit.

5.2 Traditional Classifier Results

For automatic classification, Table 4 presents the results of traditional classifiers across the eight settings. We employed the Optuna framework (Akiba et al., 2019) for hyperparameter tuning and model selection; the best configuration utilized a Logistic Regression model. Metrics reflect the top-performing fold under these optimized conditions.

The results show that traditional classifiers remain highly competitive in this task. The strongest performance is achieved by the pragmatically enriched settings *Full Subjectivity* and *Full Subjectivity + Intentions*, both reaching an F1-macro of 0.80. The *High Author Involvement + Intentions* condition also performs strongly, suggesting that pragmatic augmentation provides a selective advantage over relying exclusively on raw text.

At the same time, the impact of each pragmatic layer is not uniform. Intention labels

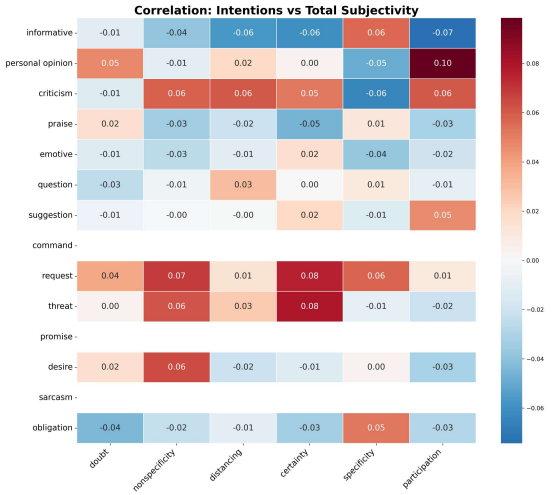

Figure 1: Correlation Heatmap of Intentions vs Full Subjectivity.

Setting	Prec.	Rec.	F1-macro
Baseline			
1. Plain Text	0.80	0.80	0.80
Single-Level			
2. + Intentions	0.77	0.76	0.76
3. + Full Subjectivity	0.81	0.80	0.80
4. + Low Author Involvement	0.79	0.79	0.79
5. + High Author Involvement	0.76	0.76	0.76
Hybrid			
6. + Low Author + Intentions	0.76	0.76	0.76
7. + High Author + Intentions	0.81	0.80	0.80
8. + Full Subjectivity + Intentions	**0.81**	**0.80**	**0.80**

Table 4: Logistic Regression performance metrics across experimental settings. Bold indicates the best results.

alone do not improve over the plain-text baseline, whereas subjectivity-based enrichments lead to clearer gains. This pattern is consistent with the descriptive analyses reported above, where subjectivity markers emerged as the most discriminative pragmatic features.

5.3 RoBERTa Results

The encoder-based experiments reveal that pragmatic augmentation provides consistent gains over the text-only baseline. Table 5 summarizes the average accuracy, F1-score, and standard deviation across the RoBERTa settings.

The best-performing RoBERTa configuration is *High Author Involvement + Intentions*, which achieves an average accuracy of 0.72 and an F1-score of 0.71. This result suggests that the combination of assertive subjectivity markers with intent is particularly informative for the encoder architecture. In contrast, the *Intentions*-only and *Full Subjectivity*-only settings perform worse than more selective pragmatic combinations, indicating that not all annotation layers are equally useful when introduced into contextual encoders.

Setting	Avg. Acc.	F1	SD Acc.
Baseline			
1. Plain Text	0.66	0.62	0.0802
Single-Level			
2. + Intentions	0.64	0.60	0.0935
3. + Full Subjectivity	0.59	0.53	0.1063
4. + Low Author Involvement	0.69	0.68	0.0593
5. + High Author Involvement	0.66	0.64	0.0479
Hybrid			
6. + Low Author + Intentions	0.67	0.65	0.0554
7. + High Author + Intentions	**0.72**	**0.71**	0.0186
8. + Full Subjectivity + Intentions	0.61	0.59	0.0697

Table 5: RoBERTa disinformation classification. Bold indicates the best results; higher SD in some settings indicates signal instability, partly undercutting a "consistent gains" framing.

Another relevant observation is the low standard deviation of the best RoBERTa setting. The *High Author Involvement + Intentions* condition shows substantially lower variability than the baseline (0.0186 vs. 0.0802), which suggests not only higher performance but also more stable behavior across folds. This supports the idea that selected pragmatic cues help the encoder capture more robust signals of veracity.

5.4 LLM Results

The generative models confirm the general trend observed with RoBERTa: pragmatic annotations improve over text-only baselines across both architectures. Table 6 compares GPT-oss and Mistral-small under the eight settings.

	GPT-oss			Mistral		
Setting	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline						
1. Plain Text	0.55	0.82	0.54	0.60	0.81	0.63
Single-Level						
2. + Intentions	0.62	0.74	0.63	0.72	0.75	0.73
3. + Full Subjectivity	0.69	0.68	0.68	0.80	0.62	0.73
4. + Low Author Involvement	0.70	0.64	**0.68**	0.77	0.67	**0.73**
5. + High Author Involvement	**0.71**	0.51	0.64	**0.85**	0.50	0.70
Hybrid						
6. + Low Author + Intentions	**0.71**	0.62	**0.68**	0.66	**0.90**	0.69
7. + High Author + Intentions	0.61	**0.84**	0.63	0.72	0.73	0.72
8. + Full Subjectivity + Intentions	0.65	0.72	0.66	0.68	0.81	0.71

Table 6: GPT-oss and Mistral comparative metrics. Bold indicates the best result per model in each metric.

Mistral-small achieves the strongest absolute F1-scores, with peak performance around 0.73 in the *Intentions*, *Subjectivity*, and *Low Author Involvement* settings. GPT-oss starts from a weaker base-

line but shows a clear improvement once pragmatic annotations are incorporated, reaching an F1-macro of 0.68 in several enriched conditions.

The two models also differ in how they exploit the pragmatic information. Mistral-small tends to benefit from single-layer pragmatic augmentation, whereas GPT-oss appears more reactive to hybrid configurations in terms of recall. In particular, the *Intentions + High Author Involvement* setting yields the highest disinformation recall for GPT-oss (0.84), suggesting that generative models may leverage pragmatic cues differently depending on whether the task favors balanced performance or stronger sensitivity to deceptive items.

Overall, LLM results confirm that augmenting raw text with structured pragmatic indicators provides a more nuanced discourse representation, consistently outperforming text-only baselines across architectures.

5.5 Cross-Architecture Comparison

A cross-architecture synthesis reveals that pragmatic features impact classifiers differently based on their underlying paradigms. While the traditional Logistic Regression model benefits from dense, global profiles (*Full Subjectivity*), deep learning models require more targeted cues. Specifically, RoBERTa and the generative LLMs show a clear affinity for structured interaction, where combining *Intentions* with specific dimensions of author commitment (*High* or *Low*) yields the highest stability and performance gains. Ultimately, while traditional models offer competitive absolute scores, neural and generative architectures prove more sensitive to the fine-grained interplay of hybrid pragmatic markers.

5.6 Statistical Significance

To move beyond anecdotal metric comparisons, we apply paired *t*-tests across four random seeds. Table 7 compares mean accuracy under the text-only baseline and under the pragmatically enriched setting used for significance testing.

The results confirm that the inclusion of pragmatic annotations produces statistically significant gains for both generative models. GPT-oss improves from a mean accuracy of 0.5435 to 0.6222, while Mistral improves from 0.6772 to 0.7088. Since both p-values are below 0.05, the null hypothesis can be rejected in both cases.

These findings are important because they show that the effect of pragmatic augmentation is not merely accidental or dependent on a particular random configuration. Rather, the gains appear robust across repeated runs, which strengthens the main claim of the paper: pragmatic information provides meaningful additional evidence for disinformation detection in Spanish.

Model	**Mean Text**	**Mean Combined**	p**-value**	**Signif.**
GPT-oss	0.5435	0.6222	0.00044	True
Mistral	0.6772	0.7088	0.00611	True

Table 7: Statistical Significance Tests (T-Test)

6 Conclusions and Future Work

This study demonstrates that pragmatic enrichment is a valuable asset for disinformation detection in Spanish. By integrating communicative intentions and subjectivity markers, we have shown that modeling authorial stance and intent significantly improves performance across multiple paradigms. Our findings highlight that subjectivity markers (specifically *Certainty* and *Specificity*) possess the strongest discriminative power, providing interpretable evidence that deceptive content often differs from factual reporting in its management of precision and commitment. Statistical testing confirms that these pragmatic gains are robust, particularly for generative models, suggesting that detection systems should move beyond surface-level lexical analysis.

Despite these contributions, this work faces certain limitations. The study is restricted to a binary veracity setup in Spanish, which simplifies the realistic mixture of true and false elements within actual news texts, making deception detection inherently complex. Additionally, the dataset scale remains focused, and the intention taxonomy, originally designed for social media, may not fully encapsulate all journalistic nuances. Finally, the semi-automatic pipeline limits large-scale, real-time deployment.

Future work will scale the framework to multi-class, multilingual corpora and develop automated tagging via fine-tuned LLMs. Crucially, we will investigate sentence-level correlations to determine how localized subjectivity shifts propagate to affect the text's profile. Ultimately, we seek to integrate these features into explainable architectures for a transparent understanding of structural deception signatures.

Ethics Statement

The study addresses disinformation detection in order to support more transparent and interpretable NLP systems. Care should be taken to avoid overclaiming automatic detection capabilities in high-stakes settings without human oversight.

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 del 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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico 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.

John Langshaw Austin. 1962. *How to Do Things with Words*. Oxford at the Clarendon Press.

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, Patricio Martínez-Barco, and Mario Nieto-Pérez. 2024. Run-as: a novel approach to annotate news reliability for disinformation detection. *Language Resources and Evaluation*, 58(2):609–639.

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

Alvaro Zaid Gallardo-Hernández, Ramón Aranda, and Angel Diaz-Pacheco. 2025. A multi-task beto-based framework with synthetic data augmentation for sentiment and contextual classification of spanish tourist reviews. *IberLEF@ SEPLN*.

Helena Gómez-Adorno, Juan Pablo Posadas-Durán, Gemma Bel Enguix, and Claudia Porto Capetillo. 2021. Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task. *Procesamiento del lenguaje natural*, 67:223–231.

Samuel González-López, Steven Bethard, Francisca Cecilia Encinas Orozco, and Adrián Pastor López-Monroy. 2021. Consumer cynicism identification for spanish reviews using a spanish transformer model. *Procesamiento del Lenguaje Natural*, 66:111–120.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Ken Hyland. 2005. Metadiscourse: Exploring interaction in writing. *Continuum*.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Yi Liu. 2021. A corpus-based lexical semantic study of Mandarin verbs of zhidao and liaojie. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 646–651, Shanghai, China. Association for Computational Linguistics.

María Miró Maestre, Ernesto L Estevanell-Valladares, Robiert Sepúlveda-Torres, and Armando Suárez Cueto. 2025. Enhancing pragmatic processing: A two-dimension approach to detecting intentions in spanish. *Procesamiento del lenguaje natural*, 74:263–276.

Laura Martín-Hoz, Samuel Yanes-Luis, Jerónimo Huerta Cejudo, Daniel Gutiérrez-Reina, and Evelia Franco Álvarez. 2025. A comparative study of bert-based models for teacher classification in physical education. *Electronics*, 14(19):3849.

Ivan Martinez Murillo, María Miró Maestre, Aitana Martínez, Snorre Ralund, Elena Lloret, Paloma Moreda, and Armando Suárez Cueto. 2025. GPLSI-CORTEX at SemEval-2025 task 10: Leveraging intentions for generating narrative extractions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 575–583, Vienna, Austria. Association for Computational Linguistics.

Maria Miro Maestre, Robiert Sepulveda-Torres, Ernesto Luis Estevanell-Valladares, Armando Suarez Cueto, and Elena Lloret. 2025. Towards intention-aligned reviews summarization: Enhancing LLM outputs with pragmatic cues. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 737–747, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mistral AI. 2024. Mistral large: Our most advanced model. `https://mistral.ai/news/mistral-large/`. Accessed: 2026-03-20.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3391–3401.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data. *Knowledge and Information Systems*, 67(4):3267–3292.

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press.

Robiert Sepúlveda-Torres, Alba Bonet-Jover, Isam Diab, Ibai Guillén-Pacho, Isabel Cabrera-de Castro, Carlos Badenes-Olmedo, Estela Saquete, M Teresa Martín-Valdivia, Patricio Martínez-Barco, and L Alfonso Ureña-López. 2024. Overview of flares at iberlef 2024: Fine-grained language-based reliability detection in spanish news. *Procesamiento del lenguaje natural*, 73:369–379.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Gaye Tuchman and Eva Aladro Vico. 1998. La objetividad como ritual estratégico: un análisis de las nociones de objetividad de los periodistas. *CIC: Cuadernos de información y comunicación*, 4:199–218.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policymaking. Technical report, Council of Europe, Strasbourg.

L Yuan, H Jiang, H Shen, L Shi, and N Cheng. 2024. Sustainable development of information dissemination: A review of current fake news detection research and practice. systems 2023, 11, 458.

George Yule. 2022. *The study of language*. Cambridge university press.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

María Antonia Martín Zorraquino. 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 Intention annotation with GPT-oss

Communicative intentions were annotated through a two-stage procedure. First, GPT-oss was used as an automatic pre-annotation model to assign one predominant communicative intention to each sentence in the corpus. The annotation followed the 13-category taxonomy proposed by Maestre et al. (2025). This choice was motivated by previous work showing that GPT-family language models can support pragmatic intention classification in Spanish with strong empirical results (Miro Maestre et al., 2025; Martinez Murillo et al., 2025). In the present study, GPT-oss was selected because it could be used within the access and computational constraints of our experimental setting,

while still allowing a controlled, reproducible, and constrained-output annotation protocol.

For each sentence, the model received only the text to be annotated and the predefined inventory of intention labels. It was not provided with the TRUE/DISINFORMATION label, the source corpus identifier, the cross-validation fold, or any downstream classification output. The annotation was performed in a zero-shot setting using the following prompt:

> *From now on you're going to classify the global communicative intention of the text shown below. The text intention must be one of the following 13 categories: "informative", "personal opinion", "praise", "criticism", "desire", "request", "question", "command", "suggestion", "sarcasm / joke", "promise", "threat" or "emotional". I want your answer to be: The global intention of the text is: Text:*

The raw outputs were normalised to the predefined label inventory. Cases in which the model returned an invalid label, multiple labels, or additional explanatory text were manually checked during the validation stage.

Once the intentions had been automatically annotated, a linguistics expert reviewed a subset of 100 automatically assigned intention labels and assessed whether they matched the predominant communicative function conveyed by each sentence. Only five labels were manually modified by the expert, meaning that 95 out of the 100 reviewed annotations were accepted without correction. On this basis, the remaining annotations were retained as automatically assigned labels. We nevertheless acknowledge that this procedure does not eliminate all possible annotation errors, as some degree of disagreement or uncertainty is expected in automatic intention classification tasks (Maestre et al., 2025).

Because the validated intention labels were subsequently used as pragmatic features in the classification experiments, it is necessary to clarify how this annotation step relates to the later use of GPT-oss as one of the evaluated classifiers. In this study, GPT-oss played two distinct roles: first, it was used as an automatic pre-annotation tool for communicative intentions; later, it was evaluated as a zero-shot generative model for the separate task of veracity classification. These two stages differed in input assumptions, output space, and experimental function. The annotation step produced pragmatic intention labels from the sentence text and the predefined taxonomy, whereas the classification step predicted binary TRUE/DISINFORMATION labels. This setup does not constitute direct data leakage, because the intention annotation stage did not use the TRUE/DISINFORMATION labels, the train-test splits, downstream predictions, or information from the classification experiments. Moreover, the same validated intention annotations were used as input features across all model families, not only in the GPT-oss classification condition. Therefore, GPT-oss was not given privileged access to information unavailable to the other classifiers. Any remaining concern should therefore be understood as a possible model-specific coupling effect rather than as leakage from the veracity-classification task.