Attitude Analysis in Systemic Functional Linguistics: A New Corpus and Benchmark Task

Jorge Ortiz Fuentes[1,2,3] **Felipe Bravo-Marquez**[1,2,3] **Beatriz Quiroz**[4]
[1]Department of Computer Sciences, Universidad de Chile, Chile
[2]National Center for Artificial Intelligence, Chile
[3]Millennium Institute for Foundational Research on Data, Chile
[4]Facultad de Letras, Pontificia Universidad Católica de Chile, Santiago, Chile
jorge@ortizfuentes.com fbravo@dcc.uchile.cl bquiroz@uc.cl

Abstract

We introduce a corpus for Systemic Functional Linguistics (SFL) Appraisal analysis in Chilean Spanish, together with TULIO, a BERT model adapted to Chilean Spanish through continual pre-training. The corpus contains 2,546 texts annotated for the Attitude subsystem, with a focus on negative Judgment, and supports two span-level sequence labeling tasks: Attitude Types, which distinguishes Affect, Judgment, and Appreciation, and Judgment Subtypes, which identifies five finer-grained forms of social evaluation. To reduce the boundary ambiguity, we use a two-stage protocol that first marks structural spans and then assigns Attitude labels. Three expert annotators reach almost perfect agreement at the general level (Krippendorff's $\alpha = 0.834$) and substantial agreement for Judgment Subtypes ($\alpha = 0.647$). We benchmark supervised sequence labelers and structured LLM agents. The highest observed F1 scores are 0.554 on Attitude Types from a Flair BiLSTM-CRF with TULIO embeddings and 0.410 on Judgment Subtypes from Qwen 3.6 Plus. Under Twitter-only training, supervised performance is genre-sensitive, with higher F1 on Twitter than on other genres. All model scores remain below expert agreement.

1 Introduction

How language encodes evaluations and interpersonal positions has been studied across linguistics, psychology, and social science (Hunston and Thompson, 2000; Martin and White, 2005; Gray and Biber, 2012). Currently, there is a limited dialogue between two primary fronts: computational approaches, which prioritize scalability and materialize in tasks such as Sentiment Analysis and Stance Detection, and qualitative linguistic approaches, which develop rich theoretical frameworks but lack automation. A widely used qualitative framework in discourse studies is Appraisal Theory within Systemic Functional Linguistics (SFL) (Martin and White, 2005; Martin and Rose, 2003). We model two Attitude analysis problems from this theory as Natural Language Processing (NLP) sequence-labeling tasks. To this end, we use negative judgments in Chilean Spanish as a case study, presenting a new annotated corpus and a set of modeling experiments to automate these tasks and establish a foundational benchmark using both Machine Learning and generative LLMs.[1]

Our work intersects two fields: NLP and linguistic discourse studies. In NLP, it is common to employ Sentiment Analysis or Stance Detection tools to understand positions. However, the analytical focus of these methods remains centered on identifying the valence (or polarity) and the target of an opinion. In contrast, in our linguistically grounded approach, Attitude analysis aims to classify the very nature of the evaluation itself: distinguishing, for example, between a moral judgment, an aesthetic appreciation, and an emotional reaction (Martin and Rose, 2003; Martin and White, 2005).

In SFL, language is analyzed through distinct aspects known as systems. This study focuses on the *Attitude* system, which categorizes evaluations into *Affect* (emotions), *Judgment* (evaluations of people's behavior), and *Appreciation* (evaluations of things). Each of these options is divided into multiple categories that can be analyzed in greater or lesser depth depending on the analysts' objectives.

As a first step towards automating this complex analysis, we narrowed our focus to studying Affect, Judgment, and Appreciation in general, and in particular focusing only on negative *Judgment*. This decision was driven by three factors. First, negative Judgment is socially relevant: toxic, uncivil, and

[1]The corpus is available at https://huggingface.co/datasets/jorgeortizfuentes/chilean-spanish-attitude-corpus and the experiment code at https://github.com/jorgeortizfuentes/chilean-spanish-attitude-experiments. Appendix J lists all released resources.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 28000–28023
October 24-29, 2026 ©2026 Association for Computational Linguistics

offensive language is prevalent in online discourse, eroding public debate and harming individuals (Sobieraj and Berry, 2011; Anderson et al., 2018; Kim et al., 2021). Second, this scope matches the intended analytical setting: in online news comments, Cavasso and Taboada (2021) report that over 83% of Judgment realizations are negative, and hostile public discourse is precisely where negative Judgment is central for critical discourse analysis and forensic linguistics. Third, restricting the scope to negative Judgment is methodologically advantageous, as it lets us reuse offensive-language detection resources to build a corpus dense with the phenomenon of interest.

To conduct this work, we consider the entire pipeline involved in defining the problem. This encompasses data annotation, problem formulation as an NLP task, automation using distinct techniques, and performance evaluation. To this end, we formulate three research questions: **RQ1**: Can three expert linguists achieve at least substantial agreement when annotating SFL Attitude and Judgment categories in Chilean Spanish? **RQ2**: Can supervised machine learning models effectively learn to automate Attitude analysis? **RQ3**: Can Generative Large Language Models (LLMs) perform Attitude analysis through structured prompt-based classification without task-specific fine-tuning?

2 Background and Related Work

We introduce the key concepts of Appraisal Theory and then review related computational tasks and previous attempts to annotate and automate Appraisal analysis.

2.1 Appraisal Theory

Systemic Functional Linguistics (SFL) is a theory that views language as a social semiotic system, a resource for making meaning in context (Matthiessen, 2023; Matthiessen and Teruya, 2023). Within SFL, Appraisal Theory offers a systematic framework for analyzing evaluative language, or how speakers and writers express their feelings and negotiate interpersonal positions (Martin and White, 2005). The theory conceptualizes three interrelated subsystems: Attitude, Engagement, and Graduation. Our work focuses on *Attitude*, which is concerned with analyzing feelings, including emotional reactions, judgments of behavior, and the evaluation of things.

The Attitude subsystem is divided into three semantic regions: *Affect* (the expression of emotion, e.g., *I am happy*), *Judgment* (the evaluation of human behavior and character, e.g., *corrupt politician*), and *Appreciation* (the evaluation of objects, processes, and phenomena, e.g., *beautiful sunset*).

Our work focuses on the more specific categories of Judgment. As illustrated in Figure 1, Judgments are divided into two main types: *Social Esteem* and *Social Sanction*. *Social Esteem* is subdivided according to evaluations based on norms of Normality (how unusual?), Capacity (how capable?), and Tenacity (how resolute?). In contrast, *Social Sanction* is subdivided into Veracity (how truthful?) and Propriety (how ethical?). For example, calling someone a *dishonest* person is a negative judgment of Veracity, questioning their integrity, and calling someone an *incompetent* manager is a negative judgment of Capacity, questioning their skill.

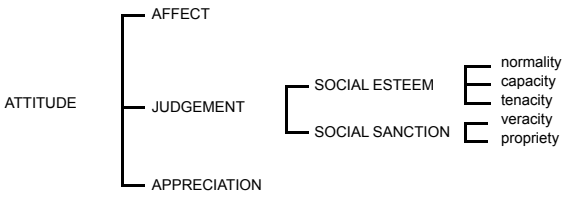


Figure 1: Attitude subsystem and Judgment subcategories.

2.2 Computational Analysis of Stance and Sentiment

Several NLP tasks already address subjective language, though each captures only part of what Appraisal describes. Sentiment Analysis (SA), the most common of these, classifies text polarity (Jurafsky and Martin, 2026). More advanced variations, such as Aspect-Based Sentiment Analysis (ABSA), identify the polarity towards specific aspects or targets within a text (e.g., positive sentiment towards the *screen* but negative towards the *battery* of a phone) (Do et al., 2019; Jiang et al., 2011). Another related task is Stance Detection, which determines an author's position (e.g., favor, against, neutral) towards a specific target or claim (Mohammad et al., 2016). Although useful, these tasks lack the descriptive richness of qualitative discourse analysis. They do not explain *how* evaluations are constructed.

There are existing offensive-language datasets that contain annotations similar to those we aim to capture in this research. One notable dataset

labeled at the span level is SemEval-2021 Task 5, which identifies the offsets responsible for a toxic label (Pavlopoulos et al., 2021). Similarly, HateXplain provides document labels for hate, offensive, and normal content, along with token-level annotator rationales for hateful and offensive posts (Mathew et al., 2021). While these types of annotations resemble those in our research, they solely locate harmful content rather than classifying evaluative meaning within the SFL (Systemic Functional Linguistics) Attitude hierarchy, which is our primary objective.

2.3 Computational Approaches to Appraisal Analysis

The application of computational methods to Appraisal Theory is a nascent field. Most research has focused on manual annotation, and the use of technology has usually been limited to corpus linguistics tools like AntConc (Anthony, 2005) or annotation software such as the UAM Corpus Tool (O'Donnell, 2009). However, some studies have worked on creating annotated resources and exploring automatic annotation.

First, Read and Carroll (2012) studied inter-annotator agreement for full Appraisal labeling in English. They document difficulties in span delimitation and fine-grained classification. They report F1 agreement per level of their Appraisal hierarchy, with 0.528 at Level 2 and 0.396 at Level 4, the level at which the five Judgment Subtypes are classified. Each level aggregates categories from Attitude, Engagement, and Graduation, so these values are not restricted to the Attitude subsystem. Second, Mora and Lavid-López (2018) labeled a corpus of reviews with Appraisal categories. In their work, they report moderate agreement with a Kappa score of 0.49 for Attitude subtype annotations. Third, Barriere et al. (2018) used Hidden Conditional Random Fields (HCRF) to classify the valence (positive or negative) of user attitude in human-agent dialogues. Their approach consisted of classifying interaction pairs (one agent turn and one user turn) and they obtained an F1 score of 0.80 on this task.

Our corpus operationalizes the five Judgment Subtypes in Chilean Spanish. Its two-stage procedure first delimits nominal and adjectival groups and then assigns semantic labels. This design responds to the boundary problems reported by Read and Carroll (2012). The agreement comparison in Section 3.3 is indicative rather than direct because the studies differ in language, label inventory, span granularity, and annotation interface. Targeted sampling increases the density of negative Judgment. The corpus also provides a benchmark for supervised Transformers, Flair models, and structured LLM agents.

3 Corpus: A New Resource for Attitude Analysis

A primary contribution of this research is the creation and validation of a new corpus for Attitude analysis in Chilean Spanish. This section details the corpus design, annotation methodology, and final statistics.

3.1 Corpus Design and Selection

The first challenge was to create a corpus enriched with instances of Attitude and particularly with negative Judgments. For SFL, the text is the minimum semantic unit (Matthiessen, 2023; Matthiessen and Teruya, 2023), so we needed to obtain complete texts. To this end, we collected a large body of 4.7 million texts in Chilean Spanish from the 2010-2022 period, sourced from Twitter, news articles, opinion columns, and a consumer complaint website. Random sampling from this collection would be methodologically inadequate for our research objectives. Expressions of Attitude, particularly fine-grained evaluative language, are relatively rare in most discourse genres (Martin and White, 2005). Therefore, a simple random sample would produce a corpus too sparse to capture our phenomenon of interest (Xiao, 2010).

To address this problem, we employed a classifier-based targeted sampling strategy, inspired by confidence-based instance selection in fBERT Sarkar et al. (2021). The approach involves using pre-trained classifiers to identify and oversample texts with a high probability of containing the target phenomenon. In our work, we used a BERT model fine-tuned on a Chilean Spanish corpus for offensive language detection (Arango et al., 2022). The rationale for this selection strategy is based on the overlap between the definitions of offensive language and the system of Judgment. Because Judgment is defined as the evaluation of human behavior and character against social and ethical norms (Martin and White, 2005) and offensive language is inherently a devaluing or discriminatory attack based on the transgression of such norms (Waseem et al., 2017; Zampieri et al., 2019), we

infer that texts containing offensive language have a high probability of also containing negative Judgments. This link is further supported by forensic research that has applied Appraisal Theory to identify evaluative patterns in malicious and threatening discourse (Reczek, 2021).

To assess the density of the selected stratum, we manually reviewed 100 randomly sampled texts from it. The review found Judgment expressions in 86% of the sample. This check supports the use of the classifier as an enrichment proxy, but it does not measure coverage of negative Judgment across registers. Because the proxy targets offensive language, it likely over-represents overt hostility and under-represents understated criticism. Within the Appraisal framework, the devaluation of human behavior is a central linguistic resource for constructing hostility and malicious threats (Reczek, 2021). This theoretical premise aligns with the consensus in NLP abusive language detection, where abuse is defined as intentionally harmful attacks and derogatory remarks directed at specific individuals or groups (Waseem et al., 2017). Our classifier-based sampling therefore yields a corpus dense with negative Judgment. We discuss the scope implications of excluding positive polarity in the Limitations section.

3.2 Annotation Methodology

The annotation was conducted by three linguists with expertise in SFL, using the open-source platform Argilla (Argilla, 2023). Following best practices for complex annotation tasks (Fort, 2016; Hovy and Lavid, 2010), the process was iterative and structured into two main campaigns.

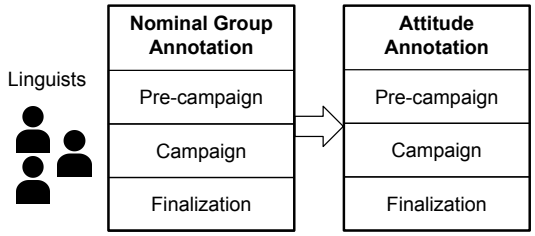


Figure 2: The two-campaign annotation workflow, designed to first establish structural boundaries (nominal groups and adjectival groups) and then apply the more subjective semantic labels (Attitude).

Campaign 1: Group Annotation. To delimit the scope of analysis and improve consistency, in the first campaign annotators annotated the boundaries of two types of syntactic groups: nominal groups (noun phrases) and adjectival groups that could potentially carry attitudinal meaning. This structural layer is comparatively more objective and provided a stable foundation for the subsequent semantic labeling stage.

Campaign 2: Attitude Labeling. In the second campaign, annotators labeled the previously identified nominal and adjectival groups with the most specific Attitude category possible, following the hierarchy in Figure 1. Annotators could choose a more general category (e.g., 'Judgment') if a more specific one (e.g., 'Propriety') was not clearly distinguishable. Annotators discussed difficult cases regularly, and these discussions produced explicit decision rules that were folded back into the guidelines.

Both campaigns followed a three-phase process: (1) a *pre-campaign* phase for training, guideline development, and initial agreement checks on a shared set of texts; (2) a *campaign* phase where annotators worked on distinct text sets but with a shared subset to continuously monitor agreement; and (3) a *finalization* phase to consolidate the guidelines and document the process.

For texts with multiple annotations, we retained the full annotation layer of the annotator identified as most consistent in that phase. This preserves coherent span boundaries and hierarchical label choices, at the cost of releasing a selected expert annotation rather than an adjudicated consensus.

3.3 Inter-Annotator Agreement (IAA)

To investigate whether expert annotators can achieve at least substantial inter-annotator agreement on SFL Attitude categories in Chilean Spanish and how agreement varies across different levels of semantic granularity (**RQ1**), we measured IAA using Krippendorff's Alpha (α) (Krippendorff, 2018; Artstein and Poesio, 2008). To investigate the impact of semantic granularity on annotation consistency, we computed α at four distinct hierarchical levels: (1) the general level using original BIO-tagged labels; (2) Level 1 - Attitude Types (Affect, Judgment, Appreciation); (3) Level 2 - Judgment Types (Social Esteem vs. Social Sanction); and (4) Level 3 - Judgment Subtypes (Normality, Capacity, Tenacity, Propriety, Veracity). For each level, labels outside the specific taxonomy were systematically converted to 'O' tags to maintain token-level comparability across all three annotators.

We interpret α qualitatively following the widely used benchmarks of Landis and Koch (1977): values below 0.00 indicate *poor* agreement, 0.00–0.20 *slight*, 0.21–0.40 *fair*, 0.41–0.60 *moderate*, 0.61–0.80 *substantial*, and 0.81–1.00 *almost perfect* agreement.

Table 1 presents the average Krippendorff's Alpha and the corresponding span-level micro F1-score between annotators across all annotation phases. We report results at the four hierarchical levels described above. These figures provide both a reliability estimate that corrects for chance agreement (α) and a task-aligned span-level F1 reference for comparison with the scores obtained by automatic models in Section 4.

Table 1: Average Krippendorff's Alpha (α) and span-level micro F1 between annotators across levels of granularity.

Level	**Avg α**	**Avg F1**
General	0.834	0.830
Level 1 (Attitude Types)	0.725	0.719
Level 2 (Judgment Types)	0.684	0.678
Level 3 (Judgment Subtypes)	0.647	0.600

The agreement levels achieved validate the reliability of our annotation methodology. Overall agreement at the general level reaches $\alpha = 0.834$ and F1 ≈ 0.83, which corresponds to *almost perfect* agreement in the Landis and Koch scale (Landis and Koch, 1977) and represents strong reliability according to established benchmarks (Artstein and Poesio, 2008; Krippendorff, 2018). Span-level F1 gives no credit for agreement on `O`. We also recomputed α after excluding tokens labeled `O` by all three annotators. This retains 54.3% of tokens on average and yields mean $\alpha = 0.716$ across the nine subcampaigns, still within the substantial range. However, when decomposing the distinct categories, agreement decreases following a systematic and theoretically expected pattern for finer-grained distinctions: Level 1 achieves $\alpha = 0.725$ and F1 ≈ 0.72, Level 2 reaches $\alpha = 0.684$ and F1 ≈ 0.68, and Level 3 (the most detailed) obtains $\alpha = 0.647$ and F1 ≈ 0.60.

The observed hierarchical decline aligns with established findings in annotation theory (Artstein and Poesio, 2008), which predict that increasing the number of categories and the subtlety of their semantic boundaries makes discrimination more cognitively demanding and inherently more subjective. Even at the most challenging granularity level, agreement remains in the *substantial* range ($\alpha = 0.647$). This pattern also reflects an important aspect of our annotation instructions: when annotators could not confidently assign a fine-grained category, they were instructed to assign the broader category. While methodologically sound for capturing annotator certainty, this instruction systematically concentrates fine-grained labels in the most complex cases, naturally increasing nominal disagreement at finer levels.

From the pre-campaign phases to the production campaigns, α increased from 0.819 to 0.842 at the general level, from 0.679 to 0.687 at Level 2, and from 0.628 to 0.657 at Level 3; it decreased from 0.735 to 0.720 at Level 1. These descriptive changes do not isolate the effect of individual workflow components.

Read and Carroll (2012) report F1 agreement per level of their full Appraisal hierarchy, with 0.528 at Level 2 and 0.396 at Level 4, the level at which the five Judgment Subtypes are classified. Our span-level F1 is 0.830 at the general level and 0.600 for Judgment Subtypes. This comparison is indicative only. The studies differ in language, label inventory, span granularity, and annotation interface. Their metric also gives credit to overlapping spans, unlike our exact-boundary matching. Our protocol separates group delimitation from semantic labeling, addressing one source of difficulty identified by Read and Carroll. The present comparison does not isolate the effect of that procedural difference.

Our IAA analysis provides an affirmative answer to **RQ1**: expert annotators can achieve agreement ranging from *substantial* (Levels 1–3, $\alpha = 0.647$–0.725) to *almost perfect* at the general level ($\alpha = 0.834$, F1 ≈ 0.83). Therefore, our corpus constitutes an expert-annotated benchmark for computational discourse analysis. The span-level F1 scores reported here also serve as human performance baselines for the two automation tasks considered in Section 4: Attitude Types and Judgment Subtypes.

3.4 Final Corpus Statistics and Availability

After consolidating the annotations, our corpus consists of 2,546 texts. The distribution of labels across all three annotation layers is shown in Table 2. The corpus is dominated by *Judgment* and *Appreciation*, which is consistent with the nature of the source texts (online discussions and reviews) and our sampling strategy.

Beyond benchmarking, the corpus supports SFL-

based discourse analysis of online incivility and forensic linguistics, supplies finer-grained signal than polarity for moderation and stance detection, and enables studies of dialectal variation in evaluative language.

Table 2: Span distribution across the three annotation layers. Layer totals are 6,340 for Attitude Types, 3,293 for Judgment Types, and 3,288 for Judgment Subtypes.

Layer	Label	Spans[a]
Attitude Types	Affect	483
	Appreciation	2,574
	Judgment	3,283
Judgment Types	Social Esteem[b]	1,707
	Social Sanction[b]	1,586
Judgment Subtypes	Capacity	902
	Normality	428
	Tenacity	374
	Propriety	1,423
	Veracity	161

[a] A span is a maximal contiguous run of the same non-0 label. Each layer is counted independently, so adjacent lower-level spans can merge at broader levels and totals across layers are not additive.
[b] Social Esteem subsumes Capacity, Normality, and Tenacity; Social Sanction subsumes Propriety and Veracity.

4 Automation: Benchmarking Models for Attitude Analysis

This section details our second main contribution: a benchmark of state-of-the-art NLP models on the new *Attitude* analysis tasks.

4.1 TULIO: A Chilean Spanish BERT

Given the dialectal specificity of our study, we introduce TULIO, a BERT-based model tailored for Chilean Spanish. TULIO inherits BETO's tokenizer, architecture, and Spanish lexical coverage (Cañete et al., 2020), and is further pre-trained with Masked Language Modeling (Devlin et al., 2019) on a Chilean Spanish corpus. Since the only difference with BETO is this additional pre-training, the BETO-vs-TULIO comparison isolates the contribution of dialect adaptation.

The 43.7 GB training corpus contains 37,213,992 documents: 73.4% Chilean Twitter, 23.4% mC4 `.cl` (Xue et al., 2021), and 3.2% news, books, and complaints. Pre-training consists of one epoch of continual whole-word MLM over BETO, with a learning rate of 2×10^{-5} and a batch size of 64. Full corpus composition, pre-training hyperparameters, and the loss trajectory are reported in Appendix G.

4.2 Experimental Setup

Tasks and Data We frame Attitude annotation as two sequence labeling tasks: **Attitude Types** (3 classes: *Affect*, *Judgment*, *Appreciation*) and **Judgment Subtypes** (5 classes: *Normality*, *Capacity*, *Tenacity*, *Veracity*, *Propriety*). These correspond to Levels 1 and 3 of the IAA analysis in Section 3.3; we omit Level 2 (Social Esteem vs. Social Sanction) as it has limited standalone interpretive value. The 2,546 annotated texts are split 70/15/15 into train, validation, and test.

Models We compare three paradigms. Hyperparameters and training details are in Appendix B. Each supervised configuration selected for Table 3 was retrained over three additional random seeds, reported as mean and standard deviation alongside the observed run.

Transformers. We fine-tune seven BERT-based encoders: three general Spanish (BETO (Cañete et al., 2020), BERTIN-RoBERTa (de la Rosa et al., 2022), PlanTL-RoBERTa-bne (Gutiérrez-Fandiño et al., 2022)), three multilingual (XLM-RoBERTa (Conneau et al., 2020), mDeBERTa-v3 (He et al., 2023), Twitter-XLM-RoBERTa (Barbieri et al., 2022)), and our dialect-adapted TULIO (§4.1; Appendix G). Table 3 reports the best encoder; the full sweep and a multi-seed BETO-vs-TULIO ablation are in Appendix E.

LSTM (Flair). A BiLSTM-CRF on top of Flair contextual string embeddings (Akbik et al., 2019) stacked with BETO and TULIO embeddings, used as frozen feature extractors.

LLMs. We evaluate six LLMs through LangGraph agents: proprietary (GPT-5, Gemini 2.5 Flash, Gemini 2.5 Pro, Qwen 3.6 Plus) and open-weight (Gemma 4 31B, Qwen 3.5-397B-A17B). Direct token-level prompting proved unstable on fine-grained boundaries, a known limitation of LLMs on structured prediction (Adibhatla et al., 2023; Li et al., 2025). We therefore use a *feedback* agent that produces JSON token labels and re-prompts on validation failures, with a single theory-informed prompt per task (Appendices D.1 and D.2). The full agent leaderboard is in Appendix D.

Evaluation We report entity-level micro *Precision*, *Recall*, and *F1*: a span counts as correct only if both its label and its token boundaries match the gold annotation. This is the same exact span metric used for IAA in Section 3.3. The IAA score averages annotator-pair F1 across annotation phases, whereas model scores compare predictions with the retained corpus layer on the test set; the comparison is therefore descriptive rather than an estimate of a shared sample-level ceiling. Comparisons between two systems on the same test set use a paired sentence-level bootstrap: the 382 test sentences are resampled with replacement (2,000 iterations, seed 42) and both systems are scored on each resample, so we report the 95% percentile interval of the F1 gap and treat a gap as significant when that interval excludes zero. The encoder and recipe ablations in Appendix E use this same procedure; the cross-genre intervals in Appendix F use a hierarchical variant described there. To probe cross-genre transfer, we also train each supervised recipe on Twitter only. We evaluate it on Twitter and non-Twitter texts over three seeds (Table 4; Appendix F). Span-level error decompositions are reported in Appendix C.

4.3 Results

Table 3 reports the highest observed single-run test F1 within each architecture family and task. Each supervised row reports the saved seed-42 score and the mean and sample standard deviation over seeds 1–3. Each LLM row reports one saved evaluation run. Table A1 reports precision and recall for the best model per task; per-class scores are reported in Appendix A. The error decomposition and LLM leaderboard appear in Appendices C and D. The full encoder sweep and cross-genre results appear in Appendices E and F.

Flair BiLSTM-CRF + TULIO has the highest observed score on *Attitude Types*, whereas Qwen 3.6 Plus has the highest observed score on *Judgment Subtypes*. Thus, no architecture has the highest observed score on both tasks.

Across seeds 1–3, the supervised means preserve the observed Flair–Transformer ordering on *Attitude Types* but reverse it on *Judgment Subtypes* (Table 3). The LLM scores come from single saved runs. Conditional on the observed runs, paired bootstrap comparisons support Flair over the Transformer (95% CI $[+0.035, +0.093]$) and Gemma (95% CI $[+0.029, +0.100]$) on *Attitude Types*. On *Judgment Subtypes*, the Qwen–Flair interval excludes zero only narrowly ($[+0.002, +0.106]$), whereas the Qwen–Transformer interval includes zero ($[-0.002, +0.093]$). We therefore do not identify a single best architecture for *Judgment Subtypes*.

4.4 Analysis and Discussion

The results reveal three findings. First, the task of automating *Attitude* analysis is challenging. The best *F1-scores*, ranging from 0.36 to 0.55 across tasks and architectures, are considerably lower than state-of-the-art results on more established *sequence labeling* tasks like *Named Entity Recognition*, where scores often exceed 0.90. Framed against the expert span-level baselines of Section 3.3, the highest observed scores reach 77% on *Attitude Types* ($0.554/0.719$) and 68% on *Judgment Subtypes* ($0.410/0.600$). Both remain below expert agreement.

Second, no architecture has the highest observed score on both tasks. On the 3-class *Attitude Types* task, the Flair BiLSTM-CRF with combined embeddings (Flair contextual + TULIO) leads, with the dialect-adapted Transformer (TULIO) and the best LLM agent (the open-weight Gemma 4 (31B); Table D1) about 6.5 F1 points behind. On the finer-grained *Judgment Subtypes* task, four LLM agents have observed point estimates above the highest observed supervised score (0.3635; Table A2). Qwen 3.6 Plus is close to 5 F1 points above that score, Gemini 2.5 Flash is about 4 points above it, and the two open-weight models are about 2 to 3 points above it. The F1 difference between Qwen 3.6 Plus and the TULIO Transformer is inconclusive (95% CI $[-0.002, +0.093]$), so these results do not establish a single winner for the task. Averaged within each model across the two tasks, micro F1 drops about 32% for the Transformer encoders (Table E1) and 36% for the BiLSTM-CRF, versus about 16% for the LLM agents (Table D1). The higher observed LLM scores concentrate in two of the three rarest subtypes, *Veracity* and *Tenacity* (Appendix A), which we read as pretraining priors compensating for the scarce supervision available for those categories.

Third, the relative difficulty for models mirrors that of human annotators at the task level: both score higher on *Attitude Types* than on *Judgment Subtypes*. At the class level, the best *Attitude Types* model scores highest on *Judgment*, followed by *Appreciation* and *Affect* (F1 ≈ 0.59 vs. 0.51 and 0.49; Table A3), in line with the dominance of *Judgment*

Table 3: Observed test F1 and multi-seed estimates for the selected configurations. Appendix D describes the available run metadata.

Task	Model	Observed F1	3-seed mean ± SD
Attitude Types (3-class)	LSTM (Flair + TULIO)	**0.554**	0.537 ± 0.008
	Transformer (TULIO)	0.489	0.497 ± 0.011
	LLM (Gemma 4 (31B))	0.489	n/a
Judgment Subtypes (5-class)	LSTM (Flair + TULIO)	0.356	0.356 ± 0.016
	Transformer (TULIO)	0.363	0.346 ± 0.012
	LLM (Qwen 3.6 Plus)	**0.410**	n/a

in our corpus (Table 2). On *Judgment Subtypes*, the best LLM model also exhibits substantial variation across categories (Table A4): *Propriety* (F1 ≈ 0.48), *Veracity* (F1 ≈ 0.44), and *Capacity* (F1 ≈ 0.40) are learned relatively well, while *Tenacity* (F1 ≈ 0.33) and particularly *Normality* (F1 ≈ 0.19) remain harder.

Encoder Ablation TULIO outperforms the other six encoders on both tasks (Appendix E). A three-seed ablation under the Transformer fine-tuning recipe confirms that its advantage over BETO is statistically significant on *Attitude Types* ($\Delta = +0.033$, 95% CI $[+0.014, +0.052]$), while the *Judgment Subtypes* gain is suggestive but inconclusive under the same criterion ($\Delta = +0.028$, 95% CI $[-0.002, +0.055]$; Table E2). Under the Flair recipe, a three-seed *Attitude Types* ablation also favors TULIO over BETO (0.537 ± 0.008 vs. 0.507 ± 0.009, $\Delta = +0.029$, 95% CI $[+0.009, +0.050]$).

Cross-Genre Transfer F1 is lower outside Twitter for both recipes and tasks, with every task-level interval for the ID–OOD difference above zero (Table 4). The decrease is larger on *Judgment Subtypes* for both recipes. The gold Judgment span-onset rate per token is 4.2 times higher on Twitter (Table F3), whereas Flair's pooled Appreciation F1 is similar across splits (0.470 ID vs. 0.445 OOD; Table F2); a separate hierarchical bootstrap for the ID–OOD difference includes zero (95% CI $[-0.018, +0.071]$). Across all four task–recipe pairs, precision and recall both decrease outside Twitter (Table F1). These patterns are consistent with a label-distribution shift, but they do not isolate its causal effect. Appendix F reports split-level precision, recall, confidence intervals, and per-class F1 results.

The main benchmark instead uses the full training split, including both genres. Three of its four supervised configurations have higher point estimates on the 18 non-Twitter test texts than on Twitter. This small subset does not establish stable transfer; it shows that the decreases in Table 4 concern Twitter-only training.

Error Analysis We matched gold and predicted *spans* and decomposed non-exact outcomes into False Positives, False Negatives, Classification Errors, Boundary Errors, and Boundary + Classification Errors (Appendix C, Table C1). For the best *Attitude Types* model (Flair BiLSTM-CRF + TULIO), boundary mistakes account for 13.2% of the combined gold and predicted span total (plus 3.4% combining boundary and classification errors), and false negatives for 10.9%. For the best *Judgment Subtypes* model (Qwen 3.6 Plus), false negatives account for 17.0%, boundary errors for 14.3%, and classification errors for 10.4%. Boundary errors are therefore a dominant source of error in both tasks, which is consistent with difficulties reported in previous *Appraisal* annotation studies. Our human workflow separates the identification of nominal and adjectival groups from semantic labeling, whereas models make both decisions in a single BIO sequence. The present analysis does not isolate whether this procedural difference contributes to boundary errors. Appendix I presents the boundary rules, a worked two-stage annotation, and selected gold–prediction comparisons.

The three paradigms have different precision–recall profiles (Table A2). The Transformer changes from $P = 0.46$, $R = 0.53$ on *Attitude Types* to $P = 0.34$, $R = 0.39$ on *Judgment Subtypes*; the LLM agent changes from $P = 0.48$, $R = 0.49$ to $P = 0.43$, $R = 0.39$; and the Flair BiLSTM-CRF changes from $P = 0.56$, $R = 0.55$ to $P = 0.43$, $R = 0.30$. For Flair, the false-negative share rises from 11.0% to 26.9% of spans while false positives change little (Table C2). Because Flair always uses a CRF in our experiments, these results do not isolate the decoder from the BiLSTM and embeddings. The best observed Transformer CRF variants score about 1.2

Table 4: Cross-genre F1 after Twitter-only training, averaged over three seeds. ID is held-out Twitter ($n = 364$); OOD is non-Twitter ($n = 126$). The 95% CIs for the ID–OOD difference use 10,000 hierarchical bootstrap iterations with paired seeds and independently resampled texts.

Task	**Recipe**	**ID F1**	**OOD F1**	ΔF_1 **95% CI**
Attitude Types	Flair + TULIO	0.523	0.406	$[+0.084, +0.151]$
Attitude Types	Transformer TULIO	0.473	0.315	$[+0.125, +0.193]$
Judgment Subtypes	Flair + TULIO	0.351	0.184	$[+0.118, +0.217]$
Judgment Subtypes	Transformer TULIO	0.339	0.138	$[+0.152, +0.247]$

F1 points below the best observed softmax variants on *Judgment Subtypes* for both TULIO and BETO (Appendix E), but this separate sweep comparison does not establish the mechanism behind Flair's precision–recall shift. The observed Qwen–Transformer difference is concentrated in precision, whereas the observed Qwen–Flair difference is concentrated in recall.

Inspection of the *Judgment Subtypes* confusions reveals a pattern shared by the three paradigms: the minority subtypes (*Tenacity*, *Normality*, *Veracity*) are predominantly mislabeled as one of the two majority subtypes, *Capacity* or *Propriety*, and *Capacity* and *Propriety* are themselves frequently confused with each other across the Social Esteem and Social Sanction divide. For the supervised models this is consistent with the skewed label distribution of Table 2; the LLM agents, which are not trained on the corpus, reproduce the same bias, which suggests that *Capacity* and *Propriety* also dominate the priors that LLMs bring from pretraining on general Spanish discourse.

The results answer **RQ2** and **RQ3**: supervised models and structured LLM agents recover exact labeled spans on both tasks, but all remain below expert agreement. The evidence does not establish a single best architecture for *Judgment Subtypes*.

5 Conclusion

We formulate two Attitude analysis problems as NLP tasks and present the first Chilean Spanish corpus annotated with the *Attitude* and *Judgment* subsystems, reaching $\alpha = 0.834$ at the general level and $\alpha = 0.647$ at the finest level. The highest observed F1 scores are 0.554 on *Attitude Types* from a Flair BiLSTM-CRF with TULIO embeddings and 0.410 on *Judgment Subtypes* from a structured Qwen 3.6 Plus agent. These scores are 77% and 68% of the respective expert span-level baselines and remain below expert agreement. Under Twitter-only training, both supervised architectures score lower outside Twitter, particularly on *Judgment Subtypes*; the results do not establish stable cross-genre transfer.

Future Work We see three follow-ups. First, an extended version of the corpus with polarity-balanced sampling that combines classifier-based seeding with sentiment-lexicon seeding for positive evaluations, lifting the negative-only restriction without re-introducing sparsity. Second, broader genre and dialect coverage: additional Chilean Spanish genres (formal news, parliamentary debate, literary text) and other Spanish varieties (Mexican, Argentine, Peninsular) to study cross-dialectal transfer. Third, span-based decoders can model annotation spans directly. Hierarchy-aware models can use the unmodeled Level 2 labels and enforce consistency across Attitude levels.

Limitations

Although our work establishes a baseline for *Attitude* analysis, we acknowledge several limitations that frame the interpretation of our results.

Scope: Negative Judgment Our benchmark is restricted to negative Judgment. This is a deliberate scope decision, not a sampling artifact. Negative Judgment is the evaluative phenomenon that critical discourse analysis and forensic linguistics are concerned with when they study hostile and malicious language, and it is the use case for which this corpus was built. Positive Judgment lies on an orthogonal polarity axis and constitutes a distinct phenomenon; because it is comparatively rare in the genres we study, capturing it reliably calls for dedicated, polarity-aware sampling rather than the offensive-language classifier used here. We therefore treat a polarity-balanced corpus as a separate resource, outlined in Future Work, rather than a gap in the present one. Consequently, the corpus and the models trained on it are most appropriate for genres in which negative Judgment is the dominant phenomenon, and are not intended for general-purpose evaluative analysis requiring balanced po-

larity detection. The classifier-based proxy may over-represent overt hostility and under-represent more subtle forms of negative Judgment.

Structural Scope The annotation is restricted to Attitude realized in nominal and adjectival groups. Realizations outside these groups, including verbal, adverbial, and clause-level cases, are not annotated. We do not estimate their prevalence because that estimate requires a separate annotation pass over unrestricted spans.

Dialectal Specificity The corpus is composed exclusively of Chilean Spanish texts. This variety is characterized by distinct lexical and morphosyntactic features. Therefore, the performance of our models could degrade when applied to other varieties of Spanish without further domain adaptation.

Cross-Genre Generalization The Twitter-only transfer experiment in Table 4 shows lower performance on the non-Twitter genres. The drop is larger for Judgment Subtypes, and the gold label distribution also differs across genres. Models trained without non-Twitter examples should therefore be used cautiously across genres. Broader genre coverage and domain adaptation remain future work.

Fine-grained Judgment Distinctions Agreement decreases at finer levels of the annotation hierarchy, from $\alpha = 0.725$ for *Attitude Types* to $\alpha = 0.684$ for *Judgment Types* and $\alpha = 0.647$ for *Judgment Subtypes* (Table 1). Annotators could retain a broader category when a subtype could not be assigned with confidence. These results indicate that the finer-grained distinctions are more difficult to apply consistently; they do not establish an upper bound on model performance. Our annotation scheme uses non-overlapping spans by design; nested attitudinal expressions are mapped to the longest containing nominal/adjectival group at annotation time, which simplifies the labeling task at the cost of discarding some structural information.

Joint and Hierarchical Modeling We model Attitude Types and Judgment Subtypes as independent BIO sequence-labeling tasks. The corpus spans support evaluation of span-based decoders. The intermediate Judgment Types labels support hierarchy-aware evaluation even though we do not treat them as a standalone task; their agreement is $\alpha = 0.684$. We do not enforce consistency across the hierarchy.

LLM Reproducibility Each LLM score comes from one saved evaluation run, so run-to-run variation is not estimated. The retained artifacts include the agent configurations, prompts, predictions, and LLM-to-BIO alignment procedure.

Ethics Statement

Data Collection and Privacy The data collection process adhered to the terms of service of the source platforms. Twitter data was collected using the Academic Research product track of the Twitter API, strictly for research purposes. The remaining texts were obtained from public web sources (news outlets and consumer complaint websites). Although the source texts were publicly accessible, we treated their redistribution as a potential privacy risk. Before release, platform user handles were replaced with stable generic identifiers (e.g., `@user1`, `@user2`), and detected phone numbers and URLs were redacted. These measures reduce identification risk while preserving the referential consistency required for discourse analysis.

Annotator Recruitment and Compensation We recruited three annotators who hold Master's degrees in Linguistics and have specialized training in Systemic Functional Linguistics. They were compensated at an hourly rate commensurate with their professional qualifications and consistent with the market rates for specialized linguistic work in Chile.

Offensive Content and Research Intent We acknowledge that a significant portion of the corpus contains offensive, toxic, and discriminatory language. We do not support or condone the views expressed in these texts. However, as stated in Section 1, analyzing this type of discourse is necessary to understand how negative judgments function in public debate and to develop tools capable of detecting uncivil language.

Acknowledgments

This work was funded by the National Center for Artificial Intelligence (CENIA), FB210017, Basal ANID; the ANID Chile Millennium Science Initiative Program, IMFD ICN17_002; and FONDEF IDeA I+D 2025, grant ID25I10330.

References

Hiranmai Sri Adibhatla, Pavan Baswani, and Manish Shrivastava. 2023. Fine-grained contract NER us-

ing instruction based model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 889–902, Hong Kong, China. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2018. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, 30(1):156–168.

Laurence Anthony. 2005. Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 729–737. IEEE.

Aymé Arango, Jorge Pérez, Bárbara Poblete, Valentina Proust, and Magdalena Saldaña. 2022. Multilingual resources for offensive language detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130.

Argilla. 2023. Argilla: Open-source data platform. Software repository; accessed August 25, 2026.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Valentin Barriere, Chloé Clavel, and Slim Essid. 2018. Attitude classification in adjacency pairs of a human-agent interaction with hidden conditional random fields. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4949–4953. IEEE.

Luca Cavasso and Maite Taboada. 2021. A corpus analysis of online news comments using the appraisal framework. *Journal of Corpora and Discourse Studies*, 4:1–38.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero, and María Grandury. 2022. BERTIN: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Bethany Gray and Douglas Biber. 2012. Current conceptions of stance. In Ken Hyland and Carmen Sancho Guinda, editors, *Stance and Voice in Written Academic Genres*, pages 15–33. Palgrave Macmillan, London.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

Susan Hunston and Geoff Thompson, editors. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, Oxford.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160.

Daniel Jurafsky and James H. Martin. 2026. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. Online manuscript released August 19, 2026.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*, fourth edition. SAGE Publications, Thousand Oaks, California.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Xinxin Li, Huiyao Chen, Chengjun Liu, Jing Li, Meishan Zhang, Jun Yu, and Min Zhang. 2025. LLMs can also do well! breaking barriers in semantic role labeling via large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23162–23180, Vienna, Austria. Association for Computational Linguistics.

James R. Martin and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, London.

James Robert Martin and David Rose. 2003. *Working with discourse: Meaning beyond the clause*. Continuum, London.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Christian M. I. M. Matthiessen. 2023. *System in Systemic Functional Linguistics: A System-based Theory of Language*. Key Concepts in Systemic Functional Linguistics. Equinox, Sheffield.

Christian M. I. M. Matthiessen and Kazuhiro Teruya. 2023. *Systemic Functional Linguistics: A Complete Guide*. Routledge, London.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Natalia Mora and Julia Lavid-López. 2018. Building an annotated dataset of app store reviews with appraisal features in english and spanish. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 16–24.

Mick O'Donnell. 2009. The uam corpustool: Software for corpus annotation and exploration. In Carmen María Bretones Callejas, José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, María Elena García Sánchez, María Enriqueta Cortés de los Ríos, María Sagrario Salaberri Ramiro, María Soledad Cruz Martínez, Nobel Augusto Perdú Honeyman, and Blasina Cantizano Márquez, editors, *Applied Linguistics Now: Understanding Language and Mind*, pages 1433–1448. Universidad de Almería, Almería, Spain.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in English. *Language Resources and Evaluation*, 46:421–447.

Andrew Patrick Reczek. 2021. Forensic applications of appraisal theory & genre to threatening & malicious language. Mphil thesis, Aston University, Birmingham, UK.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Sobieraj and Jeffrey M Berry. 2011. From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, 28(1):19–41.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Richard Xiao. 2010. Corpus creation. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, second edition, pages 147–165. CRC Press, London.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Detailed Model Results

This appendix reports detailed span-level results for the best models used in Section 4. We first summarize the best model per task, then the best model per framework and task, and finally provide per-class F1-scores for these configurations. The supervised values in this appendix correspond to the observed seed-42 runs, each LLM value corresponds to one saved evaluation run.

B Model Hyperparameters

Table B1 summarizes the key hyperparameters used for the best Flair (LSTM-CRF) configurations on each task. Table B2 reports the corresponding settings for the best Transformer (TULIO) configurations.

C Span-level Error Analysis

This appendix summarizes span-level error distributions for the best models, decomposing non-exact outcomes into False Positives (FP), False Negatives (FN), Classification Errors, Boundary Errors, and Boundary+Classification Errors. The supervised values in this appendix correspond to the observed seed-42 runs, each LLM value corresponds to one saved evaluation run.

Table A1: Best models by task on the test set.

Task	Framework	Model	F1	Prec.	Rec.
Attitude Types	Flair	Flair BiLSTM-CRF + TULIO	0.5537	0.5608	0.5468
Judgment Subtypes	LLM	Qwen 3.6 Plus (feedback)	0.4099	0.4323	0.3898

Table A2: Best models by framework and task on the test set.

Task	Framework	Model	F1	Prec.	Rec.
Attitude Types	Flair	Flair BiLSTM-CRF + TULIO	0.5537	0.5608	0.5468
Attitude Types	Transformers	TULIO Transformer	0.4889	0.4567	0.5260
Attitude Types	LLM	Gemma 4 (31B) (feedback)	0.4887	0.4847	0.4927
Judgment Subtypes	Flair	Flair BiLSTM-CRF + TULIO	0.3558	0.4347	0.3012
Judgment Subtypes	Transformers	TULIO Transformer	0.3635	0.3420	0.3878
Judgment Subtypes	LLM	Qwen 3.6 Plus (feedback)	0.4099	0.4323	0.3898

Table A3: Per-class F1-scores for the best model from each framework on Attitude Types.

Framework	Model	Affect	Appreciation	Judgment
Flair	Flair BiLSTM-CRF + TULIO	0.4853	0.5145	0.5924
Transformers	TULIO Transformer	0.3816	0.4249	0.5553
LLM	Gemma 4 (31B) (feedback)	0.3221	0.4663	0.5264

Table A4: Per-class F1-scores for the best model from each framework on Judgment Subtypes.

Framework	Model	Capacity	Normality	Propriety	Tenacity	Veracity
Flair	Flair BiLSTM-CRF + TULIO	0.3863	0.1758	0.4424	0.0645	0.0000
Transformers	TULIO Transformer	0.4101	0.2778	0.4128	0.2051	0.2174
LLM	Qwen 3.6 Plus (feedback)	0.4044	0.1913	0.4828	0.3333	0.4444

Table B1: Key hyperparameters for best Flair models (LSTM-CRF with combined embeddings) by task.

Hyperparameter	Attitude Types	Judgment Subtypes
Label type	`attitude_tags`	`subtypes_jugdment_tags`
Embedding type	combined	combined
BERT model	TULIO (Chilean Spanish BERT)	
Max epochs	100	100
Learning rate (Flair layers)	0.1	0.05
Learning rate (BERT)	2×10^{-5}	3×10^{-5}
Train mini-batch size	16	8
Eval batch size	16	8
Hidden size (LSTM)	128	256
Patience (early stopping)	3	3
Anneal factor	0.5	0.5
Seeds (observed / multi-seed)	42 / 1, 2, 3	42 / 1, 2, 3
Use CRF	True	

Table B2: Key hyperparameters for best Transformer models (TULIO) by task.

Hyperparameter	**Attitude Types**	**Judgment Subtypes**
Label type	`attitude_tags`	`subtypes_jugdment_tags`
Model name	TULIO (Chilean Spanish BERT)	
Max epochs	8	10
Learning rate	3×10^{-5}	2×10^{-5}
Per-device train batch size	16	16
Per-device eval batch size	32	32
Effective batch size	16	16
Weight decay	0.01	0.01
Warmup steps	500	500
Warmup ratio	0.06	0.10
Evaluation strategy	epoch	epoch
Early stopping patience	2	3
Max sequence length	512	512
Gradient accumulation steps	1	1
Max grad norm	1.0	1.0
LR scheduler	linear	linear
Label smoothing factor	0.05	0.1
Adam ϵ	10^{-8}	10^{-8}
Adam (β_1, β_2)	(0.9, 0.999)	
Seeds (observed / multi-seed)	42 / 1, 2, 3	42 / 1, 2, 3
Data loader workers	0	0
Save total limit	2	2
Logging steps	100	100

Table C1: Span-level error breakdown for the best model by task on the test set. Error percentages use the sum of gold and predicted spans as denominator; together, the error columns sum to $100 \times (1 - \mathrm{F1})$.

Task	**Model**	**F1**	**FP**	**FN**	**Classification**	**Boundary**	**Bnd.+Clas.**
Attitude Types	Flair BiLSTM-CRF + TULIO	0.5537	9.68	10.95	7.47	13.16	3.37
Judgment Subtypes	Qwen 3.6 Plus (feedback)	0.4099	11.80	16.98	10.35	14.29	5.59

Table C2: Span-level error breakdown by framework and task on the test set. Error percentages use the sum of gold and predicted spans as denominator; together, the error columns sum to $100 \times (1 - \mathrm{F1})$.

Task	**Framework**	**F1**	**FP**	**FN**	**Classification**	**Boundary**	**Bnd.+Clas.**
Attitude Types	Flair BiLSTM-CRF + TULIO	0.5537	9.68	10.95	7.47	13.16	3.37
Attitude Types	TULIO Transformer	0.4889	14.88	7.83	6.96	17.10	4.35
Attitude Types	LLM (Gemma 4 (31B), feedback)	0.4887	11.19	10.36	8.35	15.77	5.46
Judgment Subtypes	Flair BiLSTM-CRF + TULIO	0.3558	8.72	26.86	14.19	9.30	5.35
Judgment Subtypes	TULIO Transformer	0.3635	19.46	13.19	11.99	12.55	6.46
Judgment Subtypes	LLM (Qwen 3.6 Plus, feedback)	0.4099	11.80	16.98	10.35	14.29	5.59

D LLM Setup and Prompts

Model Configuration We used LangGraph-based agents with six models: four proprietary (GPT-5, Gemini 2.5 Flash, Gemini 2.5 Pro, Qwen 3.6 Plus) and two open-weight (Gemma 4 (31B), Qwen 3.5-397B-A17B). Each score comes from one saved evaluation run over the 382 test texts per model and task. Table D1 reports entity-level micro F1 for the six models.

Table D1: LLM agent results across all evaluated models on the test set (*feedback* variant; entity-level micro F1). Best LLM per task in **bold**.

Model	Open?	Att. F1	Sub. F1
GPT-5	No	0.462	0.363
Gemini 2.5 Flash	No	0.475	0.404
Gemini 2.5 Pro	No	0.448	0.363
Qwen 3.6 Plus	No	0.450	**0.410**
Gemma 4 (31B)	Yes	**0.489**	0.392
Qwen 3.5 (397B-A17B)	Yes	0.447	0.383

Agent Configuration For each task, the *feedback* agent runs a single token classification chain with structured JSON output and retries when the output fails validation.

Prompting Strategy All agents relied on a single prompt with detailed instructions and inline examples embedded in the label definitions, rather than explicit k-shot lists of annotated instances. For full reproducibility, we include the complete system prompts used in production below. Each system prompt is shown together with the shared human-message template that provides the input text, the enumerated token list, and formatting constraints. All prompts are in Spanish, as the task targets Chilean Spanish; we retain the original language to avoid introducing translation artifacts into the evaluation.

D.1 Attitude Types System Prompt

The following is the full system prompt used by all LLMs for the *Attitude Types* task (3-class: `affect`, `judgment`, `appreciation`, plus `O`).

```
<ROLE>
Eres un lingüista computacional experto en Lingüística
↪  Sistémica Funcional (SFL) y en el marco de
↪  Appraisal/Valoración. Tu especialidad es la anotación
↪  del subsistema de ACTITUD (Attitude) con alto grado de
↪  precisión para el español de Chile.
</ROLE>

<TASK>
Tu tarea es analizar un texto en español (variedad de
↪  Chile) y anotar una secuencia de tokens con etiquetas
↪  de Actitud a nivel de token, restringiendo el foco a
↪  Grupos Nominales (GN) y Grupos Adjetivales (GA). Las
↪  únicas etiquetas posibles son: `appreciation`,
↪  `judgment`, `affect`, `O`.
</TASK>

<ANNOTATION_SCOPE>
El análisis debe centrarse en GN y GA que expresen valores
↪  actitudinales (Attitude):

* **Grupo Nominal (GN):** Unidad cuyo núcleo es un
↪  sustantivo o pronombre; puede incluir determinantes,
↪  modificadores (adjetivos), y cláusulas relativas.
  * *Ej.:* “una persona honesta”, “ese gobierno inoperante”,
  ↪  “la pésima idea”, “los valientes que enfrentaron el
  ↪  peligro”.
* **Grupo Adjetival (GA):** Unidad cuyo núcleo es un
↪  adjetivo (con posibles cuantificadores/complementos).
  * *Ej.:* “muy enojado”, “bastante mediocre”,
  ↪  “extremadamente bello”.
</ANNOTATION_SCOPE>

<LABEL_DEFINITIONS>
Usa únicamente estas etiquetas; los listados léxicos son
↪  ilustrativos y siempre debes considerar el contexto.

* `affect` (Afecto): Recursos para expresar
↪  emociones/sentimientos, estados afectivos o reacciones
↪  emocionales (del emisor o atribuidas a alguien).
  * *Núcleos típicos (es/CL):* contento, triste, feliz,
  ↪  furioso, asustado, enamorado, “tengo rabia”, “me
  ↪  carga”, “me encanta”, “estoy chato”.
  * *GN/GA prototípicos:* “profundamente **triste**”, “su
  ↪  **rabia**”, “gente **asustada**”.
  * *Nota:* No confundir con evaluación ética o estética;
  ↪  si el núcleo expresa emoción → `affect`.
  * Marca affect solo si el estado emocional se atribuye
  ↪  explícitamente a una persona/grupo (1ª persona,
  ↪  ‘su/mi’, sujeto humano, o predicativo sobre humanos).
  ↪  Si el adjetivo afectivo modifica
  ↪  cosas/artefactos/eventos (p. ej., ‘música depresiva’,
  ↪  ‘una triste noticia’), anota appreciation.”

* `judgment` (Juicio): Evaluaciones de personas (o grupos)
↪  respecto de su conducta, carácter o capacidad
↪  socialmente normada (ética, confiabilidad,
↪  cordura/competencia, carácter).
  * *Núcleos típicos (es/CL):* honesto, mentiroso, corrupto,
  ↪  cruel, generoso, irresponsable, incompetente,
  ↪  valiente, cobarde, “facho”, “aweonao” (tonto),
  ↪  “pesao” (carácter), “desubicado”.
  * *GN/GA prototípicos:* “una **persona justa**”, “esos
  ↪  **corruptos**”, “un **cobarde**”, “la **gente
  ↪  decente**”.
  * *Transferencia / metonimia:* Si el GN evalúa
  ↪  “palabras/argumentos” con rasgos morales
  ↪  (“**deshonestas** palabras”), transfiere el juicio a
  ↪  la persona → `judgment`.

* `appreciation` (Apreciación): Evaluaciones de cosas,
↪  entidades, procesos, eventos, productos e instituciones
↪  por su calidad, valor social, impacto, composición o
↪  estética.
  * *Núcleos típicos (es/CL):* hermoso, feo, excelente,
  ↪  mediocre, inútil (cosa), brillante (idea/plan),
  ↪  nefasto (evento/medida), “una **mierda**
  ↪  (institución/idea)”, “penca” (cuando califica
  ↪  cosa/situación).
  * *GN/GA prototípicos:* “**pésima** idea”, “un **buen**
  ↪  plan”, “la **hermosa** obra”, “ese **desgobierno**
  ↪  inoperante” (si personifica al colectivo, ver regla
  ↪  más abajo).
  * *Personificación:* Si una institución se evalúa como
  ↪  actor humano (p. ej., “gobierno **inoperante**”),
  ↪  clasifica según el objetivo real: si juzga la
  ↪  capacidad/ética de quienes la componen → `judgment`;
  ↪  si se enfoca en la calidad del “artefacto/gestión”
  ↪  como cosa/proceso → `appreciation`.

* `O` (Otro): Todo token que NO sea parte del núcleo
↪  evaluativo de un GN/GA de Attitude.
  * Incluye determinantes, preposiciones, sustantivos no
  ↪  evaluativos cuando no proyectan la evaluación, verbos
  ↪  fuera de GA, adverbios no evaluativos, y lenguaje
  ↪  fuera del alcance (p. ej., compromiso/gradación).
</LABEL_DEFINITIONS>
```

```
<RULES>

### Reglas Generales de Anotación
1. **Atender solo Attitude inscrita (explícita):** Etiqueta
↪  únicamente evaluaciones explícitas lexicalizadas.
↪  Ignora actitudes evocadas/implícitas.
2. **Conjunto cerrado de etiquetas:** Usa solo `affect`,
↪  `judgment`, `appreciation`, `O`.
3. **Variedad chilena:** Mapea chilenismos según contexto
↪  (p. ej., “aweonao”→ `judgment` [capacidad], “penca”→
↪  `appreciation` si califica cosa/situación; `judgment`
↪  si califica persona).
4. **Metáforas léxicas inscritas:** Sustantivos/adjetivos
↪  metafóricos que lexicalizan evaluación (p. ej., “rata”,
↪  “parásito”, “lacra”) → `judgment` cuando refieren a
↪  personas.
5. **Análisis contextual:** Decide la etiqueta por el
↪  blanco semántico: ¿emoción? (`affect`);
↪  ¿persona/conducta/carácter? (`judgment`);
↪  ¿cosa/proceso/evento/obra/gestión? (`appreciation`).

### Reglas de Alcance (Scope)
1. **Foco sintáctico:** Etiqueta únicamente GN y GA.
2. **Extensión de la etiqueta (proyección de grupo):** Una
↪  vez identificado el núcleo evaluativo en el GN/GA,
↪  aplica la misma etiqueta a TODOS los tokens del grupo
↪  desde su inicio (determinantes) hasta su final,
↪  incluyendo modificadores y cláusulas relativas que
↪  especifican el rasgo evaluado.
   * *Ej.:* “**un montón de cobardes**” → todos los tokens
   ↪  `judgment`; “**esas lacras**” → todos `judgment`;
   ↪  “**la pésima idea**” → todos `appreciation`; “**el
   ↪  valiente que enfrentó el peligro**” → todos
   ↪  `judgment`.
3. **Cláusulas relativas:** Si la relativa aporta el rasgo
↪  evaluativo o lo especifica, se incluye en el alcance
↪  (misma etiqueta).
4. **Hashtags:** Si el Juicio/Afecto/Apreciación está en un
↪  hashtag, etiqueta `#` y el contenido, y extiende al
↪  GN/GA si forma parte de él. Ej.: “el plan **#pésimo**”
↪  → todo el grupo `appreciation`.

### Reglas de Exclusión
1. **Exclusión de otras unidades:** Ignora verbos y
↪  cláusulas completas que no constituyen GN/GA
↪  evaluativos.
2. **Diferenciar valoración de estado/rol:** “niño enfermo”
↪  puede ser estado físico (→ `O`) salvo uso metafórico
↪  evaluativo. “gente enferma” (moralmente depravada) →
↪  `judgment`.
3. **Instituciones y cosas:** A diferencia del esquema
↪  centrado solo en Juicio, aquí **sí** se anotan
↪  evaluaciones de cosas/procesos/instituciones como
↪  `appreciation`, salvo que se personifique a sus
↪  miembros (→ `judgment`).
4. **Juicio transferido:** Evaluaciones dirigidas a
↪  “palabras/argumentos/acciones” que implican cualidad
↪  ética del emisor (“palabras **deshonestas**”) →
↪  `judgment`.
5. **Nominalizaciones afectivas** (p. ej., ‘la tristeza’)
↪  se marcan affect solo si se atribuyen a humanos (‘su
↪  tristeza’, ‘están con tristeza’). Si aparecen como
↪  título/tema o referencia genérica (‘Encuentro de la
↪  Soledad’), etiqueta O (o appreciation si se evalúa el
↪  evento).”
6. **Sustantivos de rol/identidad colectiva** (‘ricos’,
↪  ‘izquierda’, ‘elite’, ‘inmigrantes’, ‘manifestantes’)
↪  no se etiquetan salvo que estén modificados por léxico
↪  evaluativo (‘esos corruptos’, ‘lacras’).”

</RULES>

<HEURISTICS>
1) **¿Hay experimentante humano explícito o 1ª persona?**
↪  (“estoy furioso”, “su rabia”) → `affect`.
2) **Si no, el evaluado es cosa/proceso/artefacto/evento?**
↪  → `appreciation`.
3) **Si no, trata de conducta/carácter/ética/capacidad de
↪  personas/grupos?** → `judgment`.
4) **Personificación / transferencia:** Si el blanco
↪  “cosa/discurso” sirve para juzgar a personas, prioriza
↪  `judgment`; si es calidad/impacto del
↪  artefacto/proceso, `appreciation`.
5) **Casos limítrofes (rápidos):**
   - “**triste** persona” → `affect`; “**triste** noticia”
   ↪  → `appreciation`.
   - “cumbias **depresivas**” → `appreciation`.
   - “**penca** situación” → `appreciation`; “ese
   ↪  **penca**” (persona) → `judgment`.
   - “**weón/weona**” (insulto) → `judgment`.
   - Adjetivos clasificatorios/denotativos (“servicios
   ↪  **básicos**”) → `O`.
</HEURISTICS>

<INSTRUCTIONS>
1. **Lectura global:** Lee el texto completo, considerando
↪  chilenismos y contexto.
2. **Proceso (token por token):**
   a) **Escaneo previo:** Identifica posibles núcleos
   ↪  evaluativos (adjetivos, sustantivos valorativos,
   ↪  nominalizaciones afectivas).
   b) **Detecta GN/GA evaluativos** y determina el blanco
   ↪  (persona vs cosa/proceso vs emoción).
   c) **Clasifica** con `<HEURISTICS>` y
   ↪  `<LABEL_DEFINITIONS>`.
   d) **Extiende la etiqueta** a todo el GN/GA evaluativo
   ↪  (determinantes, modificadores, relativas).
   e) **Resuelve conflictos** con la regla del primer
   ↪  núcleo evaluativo.
   f) **Etiqueta `O`** a todo lo demás.
3. **Polaridad y gradación:** No anotes polaridad ni
↪  gradación; solo el tipo de actitud.
4. **Formato de Salida:** Produce JSON válido según
↪  `<OUTPUT>`.

</INSTRUCTIONS>

<INPUTS>
* Texto completo.
* Lista de tokens con índices (0, 1, 2, . . . )
↪  correspondiente al texto.
</INPUTS>

<OUTPUT>
Responde en formato JSON válido con la siguiente estructura:
  {{
    "annotations": [
      {{"index": 0, "token": "primer", "label":
      ↪  "etiqueta"}},
      ...
    ]
  }}
</OUTPUT>
```

The following human-message template was paired with this system prompt at runtime.

```
Texto: {text}
Tokens enumerados (indice -> token):
{enumerated_tokens}
Número de tokens: {token_count}
{format_instructions}

CRÍTICO: Debes producir EXACTAMENTE {token_count}
↪  anotaciones. Cada token en la lista debe tener una
↪  anotación correspondiente.
- Si un token parece raro (URLs fragmentadas, símbolos,
↪  hashtags), de todos modos anótalo con su etiqueta
↪  correspondiente o 'O'.
- NO combines ni omitas tokens.
- NO agregues tokens extra.
```

D.2 Judgment Subtypes System Prompt

The following is the full system prompt used by all LLMs for the *Judgment Subtypes* task (5-class: `propriety`, `normality`, `capacity`, `tenacity`, `veracity`, plus 0).

```
<ROLE>
Eres un lingüista computacional experto en Lingüística
↪  Sistémica Funcional (SFL) y análisis de la Actitud. Tu
↪  especialidad es la anotación del subsistema de Juicio
↪  (Judgment) con un alto grado de precisión para el
↪  español de Chile.
</ROLE>
```

```
<TASK>
Tu tarea es analizar un texto en español (proveniente de Chile) y anotar una secuencia de tokens con etiquetas de subtipos de Juicio. Debes seguir un esquema de anotación a nivel de token.
</TASK>

<ANNOTATION_SCOPE>
Tu análisis debe centrarse en grupos nominales (GN) y grupos adjetivales (GA) que expresen una evaluación de Juicio.

* **Grupo Nominal (GN):** Unidad sintáctica cuyo núcleo es un sustantivo o pronombre. Puede incluir modificadores como adjetivos, artículos, etc., y en ocasiones cláusulas subordinadas (relativas) que modifican al núcleo.
    * *Ejemplos de GN:* "una persona honesta", "esos criminales", "los valientes que son extremadamente justos", "los ladrones que asaltaron el banco".
* **Grupo Adjetival (GA):** Unidad sintáctica cuyo núcleo es un adjetivo. Puede incluir modificadores como cuantificadores (ej. "muy") o complementos.
    * *Ejemplos de Juicio en GA:* "**muy valiente**", "**extremadamente honesto**".
</ANNOTATION_SCOPE>

<LABEL_DEFINITIONS>
Debes usar únicamente las siguientes etiquetas. Ten en cuenta que las listas de palabras proporcionadas son solo ejemplos léxicos ilustrativos y el significado puede variar según el contexto. Por lo tanto, que un token aparezca en una lista no garantiza su etiquetado automático, siempre considera el contexto.

* `propriety` (Propiedad): Evalúa la conducta de una persona en términos de ética, moralidad y cumplimiento de normas sociales.
    * *Ejemplos:* bueno, malo, moral, ético, perverso, corrupto, justo, injusto, sensible, cruel, amable, generoso, egoísta, educado, grosero, respetuoso
    * *Ejemplos chilenos:* 'penca', 'mala onda', 'fresco' (caradura).
* `normality` (Normalidad): Evalúa cuán especial, usual o inusual es una persona o su comportamiento. Esta etiqueta también incluye evaluaciones del carácter general (ej. 'simpático', 'pesao') y apariencia física (ej. 'lindo', 'feo').
    * *Ejemplos:* afortunado, normal, natural, raro, peculiar, excéntrico, anticuado, vanguardista, predecible, estable.
    * *Ejemplos chilenos:* 'simpático', 'pesao', 'lindo', 'feo''.
* `capacity` (Capacidad): Evalúa la habilidad, aptitud o capacidad de una persona (física, mental o psicológica). Incluye insultos intelectuales (ahueonao, gil, estúpido, cabeza perturbada). Se enfoca en la efectividad y habilidad, no en la voluntad o determinación.
    * *Ejemplos:* poderoso, vigoroso, sano, inteligente, estúpido, dotado, cuerdo, sensato, experto, competente, exitoso, débil, enfermo, lento, tonto, ignorante, incompetente.
    * *Ejemplos chilenos:* 'aweonao', 'leso', 'malo'.
* `tenacity` (Tenacidad): Evalúa la determinación, confiabilidad, perseverancia e independencia de una persona. Se enfoca en la voluntad, la determinación y la consistencia del carácter, no en la habilidad innata.
    * *Ejemplos:* valeroso, valiente, heroico, paciente, cuidadoso, perseverante, resolutivo, confiable, leal, constante, tímido, cobarde, impaciente, temerario, débil, infiel, terco, obstinado, pecho frío.
* `veracity` (Veracidad): Evalúa la honestidad, credibilidad y franqueza de una persona.
    * *Ejemplos:* veraz, honesto, creíble, franco, cándido, directo, deshonesto, engañoso, mentiroso, manipulador, que distorsiona la verdad, manipulador, que miente.
* `O` (Otro): Asigna esta etiqueta a cualquier token que **no** sea la palabra clave que expresa el Juicio dentro de un GN o GA. Esto incluye artículos, preposiciones, cuantificadores (como "muy"), sustantivos no evaluativos (como "persona" en "persona honesta"), tokens dentro de cláusulas relativas no evaluativas, y todos los tokens fuera del alcance del análisis (verbos, adverbios no ligados a GA, etc.).
</LABEL_DEFINITIONS>

<RULES>

### Reglas Generales de Anotación
1. **Enfoque en Juicio Inscrito:** Etiqueta únicamente Juicios **inscritos** (explícitos) con las palabras evaluativas presentes. No intentes interpretar Juicios **evocados** (implícitos).
2. **Conjunto de etiquetas:** Debes usar **únicamente** las etiquetas definidas en `<LABEL_DEFINITIONS>`.
3. **Contexto (Chilenismos)**: Presta especial atención al contexto chileno. Palabras con alta carga contextual, ideológica o de uso local deben ser mapeadas a las etiquetas según los ejemplos provistos en `<LABEL_DEFINITIONS>` (ej. 'facho', 'aweonao', 'desubicado').
4. **Metáforas Léxicas Inscritas:** GNs que funcionan como metáforas léxicas que **inscriben** un juicio (ej. 'parásito', 'buitre', 'rata', 'mentalidad de cloaca', ) deben ser etiquetados, ya que la evaluación es explícita y se ha lexicalizado.
5. **Análisis del significado contextual:** el significado de una palabra evaluativa puede cambiar según las palabras que la acompañan. Analiza siempre el grupo sintáctico completo según la función que cumple en el texto. Ejemplo: 'weon' generalmente puede ser capacity (tonto) o normality (amigo). Pero en "las weonas nefastas", el adjetivo 'nefastas' fuerza la interpretación de todo el grupo hacia propriety (mala gente).
6. **Prioridad de las definiciones:** las definiciones de cada categoría son la principal fuente de verdad, por sobre los ejemplos. Si encuentras un ejemplo que parece contradecir el significado o las reglas, sigue las reglas y definiciones. Se deben analizar los significados de las palabras en su contexto específico.

### Reglas de Alcance (Scope)
1. **Foco Sintáctico:** Tu análisis debe centrarse **exclusivamente** en grupos nominales (GN) y grupos adjetivales (GA).
2. **Foco Semántico:** Identifica los GN y GA que evalúan a **personas** o grupos de personas.

### Reglas de Exclusión
1. **Exclusión Sintáctica:** Ignora verbos que no funcionen como adjetivos (participios), cláusulas completas y cualquier otra unidad sintáctica que no sea un GN o GA que exprese Juicio.
2. **Exclusión de Cosas, Conceptos e Instituciones:** Si un GN o GA evalúa una **cosa**, un **concepto abstracto**, una **acción**, una **situación** o una **institución**, etiquétalo como `O`.
    * *Ejemplo (Concepto):* En "principios morales", 'morales' describe un concepto. -> `O`.
    * *Ejemplo (Persona):* En "son personas morales", 'morales' evalúa a personas. -> `propriety`.
    * *Ejemplo (Acción):* En "Lo encuentro malo exponerlo", 'malo' evalúa la acción "exponerlo". -> `O`.
    * *Ejemplo (Situación):* En "la situación la encuentro penca", 'penca' evalúa "la situación". -> `O`.
    * *Ejemplo (Institución):* En "la facultad es una mierda", 'una mierda' evalúa una institución (Apreciación) -> `O`.
    * *EXCEPCIÓN (Personificación):* Si la institución está **personificada** (juzgada como un actor humano), SÍ es Juicio. (ej. 'un gobierno inoperante' -> `capacity`, ya que 'inoperante' juzga a las personas que lo componen).
    *  EXCEPCIÓN (Juicio Transferido / Metonimia): SÍ debe etiquetar un GN si el "concepto" o "cosa" evaluado no es neutro y está relacionado con un juicio dirigido hacia una persona o grupo.
        * *Ejemplo:* En "sus palabras deshonestas", 'deshonestas' evalúa las palabras, pero el juicio se transfiere a la persona que las pronuncia. Etiqueta `veracity`.
        * *Ejemplo:* En "un argumento que niega la realidad", 'niega la realidad' evalúa el argumento, pero el juicio se transfiere a la persona que lo presenta. Etiqueta `veracity`.
3. **Exclusión de Estado/Rol vs. Juicio:** Debes distinguir entre una evaluación de **Juicio** y una mera **descripción de estado físico, mental o rol social**.
    * *Ejemplo de Estado (O):* En "el niño enfermo", 'enfermo' describe un estado físico, no un Juicio de `capacity`. Etiqueta `O` a menos que el contexto indique que se está evaluando su capacidad.
```

```
    * EXCEPCIÓN POR METÁFORA: Si la palabra (ej. 'enfermo')
    ↪  se usa metafóricamente para juzgar el carácter o la
    ↪  mente (ej. 'Esos enfermos' refiriéndose a
    ↪  depravados), SÍ es Juicio.
4. **Exclusión de Juicio Evocado (Figurado):** Ignora el
↪  lenguaje figurado complejo o modismos que **evocan**
↪  una actitud en lugar de inscribirla (ej. 'una sociedad
↪  de avestruz', 'le importa un moco').
5.  **Exclusión de Vocativos**: Ignora los vocativos
↪  (formas de dirigirse a alguien) aunque contenga
↪  adjetivos evaluativos, a menos que sean el foco central
↪  del Juicio. Ejemplos: 'Estimado presidente' -> O.
6.  **Regla de Etiqueta 'O':** Asigna `O` a todos los demás
↪  tokens que no cumplan los criterios para una etiqueta
↪  de Juicio según las reglas anteriores.

### Reglas de Aplicación y Extensión
1.  **Regla de extensión de etiqueta:** al identificar un
↪  Juicio en un GN o GA, la etiqueta debe aplicarse a
↪  TODOS los tokens que componen el grupo sintáctico
↪  evaluativo, desde su inicio (artículo) hasta su final.
    * Si el Juicio está en un sustantivo o adjetivo, la
    ↪  etiqueta se extiende a sus artículos, determinantes,
    ↪  cuantificadores directos y cláusulas relativas que
    ↪  modifiquen al núcleo.
        * En "un montón de cobardes", los cuatro tokens
        ↪  (`un`, `monton`, `de`, `cobardes`) se
        ↪  etiquetan.
        * En "esas lacras", ambos tokens (`esas`, `lacras`)
        ↪  se etiquetan..
        * En "al corrupto", ambos tokens (`al`, `corrupto`)
        ↪  deben etiquetarse.
        * En "este desgobierno de incapaces", los cuatro
        ↪  tokens (`este`, `desgobierno`, `de`,
        ↪  `incapaces`) se etiquetan.
        * En "los asesinos de un senador", el núcleo
        ↪  asesinos (propriety) proyecta su evaluación a
        ↪  todo el grupo que lo describe. Por lo tanto,
        ↪  `los`, `asesinos`, `de`, `un`, `senador` deben
        ↪  ser etiquetados.
        * En "el valiente que enfrentó al peligro", los
        ↪  tokens `el`, `valiente`, `que`, `enfrentó`,
        ↪  `al`, `peligro` deben etiquetados porque la
        ↪  cláusula relativa es parte del GN que evalúa a
        ↪  la persona.
    * Cláusulas relativas: si una cláusula relativa
    ↪  (generalmente introducida por 'que') está
    ↪  subordinada a un grupo nominal y describe una
    ↪  cualidad o acción que juzga a la persona, la
    ↪  etiqueta de Juicio debe extenderse a todo el grupo
    ↪  nominal, incluyendo la cláusula relativa.
        * En "esas mujeres que resuelven", la cláusula "que
        ↪  resuelven" evalúa su tenacidad/capacidad. Por
        ↪  lo tanto, "esas mujeres que resuelven" deben
        ↪  etiquetarse juntos.
        * En "un weon pecho frío que no moja la camiseta",
        ↪  la segunda parte es una cláusula relativa que
        ↪  evalúa al sujeto. Por lo tanto, todos los
        ↪  tokens deben ser etiquetados.
    * Los hashtags sí deben ser etiquetados.
        * Si el Juicio está en un hashtag, entonces todos
        ↪  los tokens que componen el hashtag deben ser
        ↪  etiquetados, incluyendo el símbolo `#`.
            * En "#valiente", ambos tokens (`#`, `valiente`)
            ↪  se etiquetan.
        * Un hashtag puede ser parte de un GN o GA más
        ↪  grande. En ese caso, la etiqueta se extiende a
        ↪  todo el GN o GA.
            * En "el presidente #valiente", todos los tokens
            ↪  (`el`, `presidente`, `#`, `valiente`) se
            ↪  etiquetan.
2.  **Regla de Conflicto (Múltiples Juicios):** Si un solo
↪  GN contiene múltiples adjetivos con diferentes tipos de
↪  Juicio (ej. 'La señora loca [capacity] y mala onda
↪  [propriety]'), aplica la etiqueta del **primer**
↪  adjetivo evaluativo (`capacity` en este caso) a todo el
↪  GN. Esta es una regla operativa para resolver
↪  ambigüedades.

</RULES>

<HEURISTICS>
Para decidir la etiqueta correcta, sigue este orden de
↪  preguntas:
1.  **¿Evalúa la veracidad/honestidad?** ¿Se juzga a la
↪  persona como honesta, mentirosa, manipuladora, directa
↪  o falsa? -> Si es así, usa `veracity`.
2.  **¿Evalúa la capacidad/habilidad?** ¿Se juzga a la
↪  persona por su inteligencia, competencia, habilidad,
↪  fuerza o cordura? (ej. 'aweonao') -> Si es así, usa
↪  `capacity`.
3.  **¿Evalúa la determinación/perseverancia?** ¿Se juzga a
↪  la persona por ser valiente, cobarde, tenaz, perezosa,
↪  confiable o débil de voluntad? -> Si es así, usa
↪  `tenacity`.
4.  **¿Evalúa cuán (in)usual es?** ¿Se juzga a la persona
↪  por ser normal, rara, especial, excéntrica, o por su
↪  carácter/apariencia? (ej. 'lindo', 'pesao',
↪  'desubicado') -> Si es así, usa `normality`.
5.  **Si no es ninguna de las anteriores, ¿evalúa la
↪  ética/moralidad?** ¿Se juzga a la persona por ser buena,
↪  mala, corrupta, justa, cruel o irrespetuosa, o por su
↪  ideología? (ej. 'facho', 'culiado') -> Usa `propriety`.
</HEURISTICS>

<INSTRUCTIONS>
1.s **Contexto General:** Lee el texto completo para
↪  entender el contexto global y local (chilenismos).
2.  **Proceso de Anotación (Token por Token):**
    a. **Escaneo Previo:** Antes de analizar
    ↪  sintácticamente, realiza una lectura para
    ↪  identificar posibles "palabras clave" o grupos de
    ↪  palabras que expresen una evaluación (ej. 'inútil',
    ↪  'valiente', 'mentiroso', 'facho', 'aweonao',
    ↪  'mierda', 'nefasto'). Estas palabras serán tus
    ↪  anclas para el análisis
    b. Para cada token o grupo de tokens, identifica si son
    ↪  parte de un Grupo Nominal (GN) o Grupo Adjetival
    ↪  (GA) que evalúan a una **persona** o **grupo de
    ↪  personas**.
    c. **Identifica el núcleo evaluativo:** Dentro del
    ↪  GN/GA, encuentra las palabras clave que carga el
    ↪  Juicio (ej. 'honesta' en "una persona honesta",
    ↪  'facho' en "ese facho").
    d. **Clasifica el Juicio:** Usa las `<HEURISTICS>` y
    ↪  las listas de ejemplos en `<LABEL_DEFINITIONS>`
    ↪  para asignar la etiqueta correcta.
    e. **Aplica la Regla de Extensión:** Extiende la
    ↪  etiqueta asignada a **TODOS** los tokens que forman
    ↪  parte de ese GN/GA evaluativo.
    f. **Aplica Reglas de Conflicto y Exclusión:** Verifica
    ↪  si se aplican reglas de exclusión (ej. es una
    ↪  institución no personificada) o de conflicto
    ↪  (múltiples adjetivos).
    g. **Etiqueta 'O':** Si un token no cumple con estos
    ↪  criterios, etiquétalo como `O`.
3.  **Formato de Salida:** Genera un JSON válido con la
↪  estructura especificada en `<OUTPUT>`.
</INSTRUCTIONS>

<INPUTS>
* Texto completo.
* Lista de tokens con índices (0, 1, 2, . . . )
↪  correspondiente al texto.
</INPUTS>

<OUTPUT>
Responde en formato JSON válido con la siguiente estructura:
  {{
    "annotations": [
      {{"index": 0, "token": "primer token", "label":
      ↪  "etiqueta"}},
      ...
    ]
  }}
</OUTPUT>
```

The following human-message template was paired with this system prompt at runtime.

```
Texto: {text}
Tokens enumerados (indice -> token):
{enumerated_tokens}
Número de tokens: {token_count}
{format_instructions}

CRÍTICO: Debes producir EXACTAMENTE {token_count}
↪  anotaciones. Cada token en la lista debe tener una
↪  anotación correspondiente.
- Si un token parece raro (URLs fragmentadas, símbolos,
↪  hashtags), de todos modos anótalo con su etiqueta
↪  correspondiente o 'O'.
```

```
- NO combines ni omitas tokens.
- NO agregues tokens extra.
```

LLM-to-BIO Alignment Protocol For seqeval-based evaluation, all LLM outputs are normalized to the same BIO scheme used by supervised models. The agent emits, for every token in the input, either `O` or a class name (e.g., `judgment`, `propriety`). When converting to BIO, we treat each maximal run of consecutive tokens sharing the same non-`O` class as a single span: the first token receives a `B-`*class* prefix and subsequent tokens an `I-`*class* prefix; tokens labeled `O` are left unchanged. This yields BIO sequences of identical length to the gold annotations, so supervised and LLM-based systems are evaluated under the same exact span-matching rule used for IAA.

E Additional Encoder Baselines

This appendix reports test-set performance for the seven Transformer encoders. All models were fine-tuned under the same four-configuration hyperparameter sweep (*base*, *aggressive*, *stable*, *small_batch*), and we report the best F1 per model. TULIO and BETO results in this table are without the optional CRF decoder, to ensure all numbers in the same row are directly comparable.

Multi-seed BETO-vs-TULIO ablation Table E1 shows TULIO ahead of BETO on both tasks, but each entry there is a single-seed point estimate. We therefore ran a multi-seed ablation of the two encoders to check whether this advantage is stable across random seeds and statistically reliable. The comparison isolates the effect of dialect adaptation, since TULIO differs from BETO only in the additional pre-training on Chilean Spanish (§4.1). We fine-tuned both encoders three times under fixed per-task configurations selected from the single-seed TULIO sweep by validation F1 (*base* for *Attitude Types* and *small_batch* for *Judgment Subtypes*). The same configuration was applied to TULIO and BETO within each task, varying only the random seed. These configurations differ from the test-selected TULIO configurations in Table E1, so the multi-seed means are not replicas of those point estimates. We assess the gap between the two encoders with the paired sentence-level bootstrap of §4.2, which scores both encoders on the same resample and so removes sentence-sampling variation from the estimated gap. On *Attitude Types*, TULIO outperforms BETO by 0.033 F1 on average, and the interval excludes zero (95% CI $[+0.014, +0.052]$). On *Judgment Subtypes*, TULIO leads by 0.028 F1 and is ahead on all three seeds, but the evidence is suggestive rather than significant under this criterion because the interval includes zero (95% CI $[-0.002, +0.055]$). This contrast is consistent with the role of dialect adaptation: the gain is larger on *Attitude Types*, where surface form carries more of the signal, and smaller on *Judgment Subtypes*, where the difficulty is primarily semantic and driven by class imbalance.

Flair Recipe Ablation Because the main *Attitude Types* result is obtained by a Flair BiLSTM-CRF model with TULIO embeddings, we also ran a targeted three-seed ablation under that full recipe for *Attitude Types*. TULIO+Flair obtains 0.537 ± 0.008 F1, compared with 0.507 ± 0.009 for BETO+Flair; the paired sentence-level bootstrap of §4.2 gives a gap of $+0.029$ F1, with a 95% CI that excludes zero ($[+0.009, +0.050]$). Across these seeds, the Flair recipe favors TULIO. We do not treat this as a pure pre-training ablation, however, because the comparison is embedded in the full Flair BiLSTM-CRF recipe rather than in an encoder-only setup.

Optional CRF Decoder We additionally evaluated a Conditional Random Field decoder on top of the best two Transformer encoders (TULIO and BETO), to test whether a structured decoder closes the gap to the Flair recipe (which already pairs the encoder with a BiLSTM-CRF decoder). Comparing the highest observed test F1 in the CRF and softmax sweeps, the CRF maximum is lower on both tasks: TULIO Attitude 0.4850 vs. 0.4889, BETO Attitude 0.4601 vs. 0.4615, TULIO Subtypes 0.3513 vs. 0.3635, and BETO Subtypes 0.3260 vs. 0.3389. These are comparisons between sweep maxima rather than paired configurations. All reported Transformer results use the softmax decoder.

F Out-of-Domain Evaluation

Protocol To evaluate the cross-genre robustness of our proposed model, we partitioned the corpus by source genre into *Twitter* (microblog) and *non-Twitter*. For each task we trained TULIO on the Twitter portion only and evaluated it on (a) the held-out Twitter test split (*in-domain*, $n = 364$ texts) and (b) the entire non-Twitter set (*ood*, $n = 126$

Table E1: Additional encoder baselines on the test set. Best F1 over 4 hyperparameter configurations per model. Decoder: softmax. Best per task in **bold**.

	Attitude Types			Judgment Subtypes		
Model	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**
TULIO (Chilean BERT)	**0.4889**	0.4567	0.5260	**0.3635**	0.3420	0.3878
BETO (Spanish BERT)	0.4615	0.4293	0.4990	0.3389	0.3072	0.3780
mDeBERTa-v3-base	0.4664	0.4280	0.5125	0.2864	0.2515	0.3327
XLM-RoBERTa-base	0.4731	0.4370	0.5156	0.3119	0.2891	0.3386
Twitter-XLM-RoBERTa-base	0.4699	0.4346	0.5114	0.2873	0.2654	0.3130
PlanTL-GOB-ES-RoBERTa-base-bne	0.4247	0.3794	0.4823	0.2886	0.2983	0.2795
BERTIN-RoBERTa-base-Spanish	0.4349	0.3946	0.4844	0.3019	0.3046	0.2992

Table E2: Multi-seed BETO-vs-TULIO ablation using validation-selected configurations, which differ from the test-selected configurations retrained in Table 3. Mean F1 and standard deviation over 3 random seeds, using a frozen per-task configuration and a softmax decoder. Δ is the difference between the mean F1 of TULIO and BETO. The 95% confidence interval comes from the paired sentence-level bootstrap of §4.2.

Task	**TULIO**	**BETO**	Δ	**95% CI**	**Verdict**
Attitude Types	0.481 ± 0.001	0.448 ± 0.009	$+0.033$	$[+0.014, +0.052]$	Significant
Judgment Subtypes	0.351 ± 0.006	0.323 ± 0.018	$+0.028$	$[-0.002, +0.055]$	Inconclusive

texts). To prevent cherry-picking on the OOD signal, the hyperparameters and decoder (softmax for both tasks) were frozen in advance based on **validation** F1 over the Twitter-only training split, *before any non-Twitter data was inspected*. We repeated each experiment with 3 random seeds and computed bootstrap 95% CIs by resampling at the sentence level (500 iterations) over the pooled across-seed prediction set. Reporting score distributions over multiple runs rather than a single value follows established methodological recommendations for sequence tagging (Reimers and Gurevych, 2017) and for fine-tuning pre-trained models (Dodge et al., 2020). Note that the in-domain F1 of this experiment is lower than that of the main results in Table 3 because here the training set is restricted to one genre, so model capacity and data both differ.

We additionally replicated the protocol for the Flair BiLSTM-CRF + TULIO configuration (the *Attitude Types* winner in Table 3) with the same 3 seeds. The Flair recipe configuration (hyperparameters and embedding stack) was frozen from the main in-domain configuration before any non-Twitter data was inspected, and the bootstrap convention matches the Transformer protocol.

These results concern Twitter-only training and are separate from the full-corpus supervised results in Table 3.

G TULIO Pre-training Details

This appendix reports TULIO's initialization, pre-training corpus, and training procedure.

Initialization and architecture TULIO is obtained by continual masked-language-model (MLM) pre-training of *BETO* (Cañete et al., 2020) (specifically, *bert-base-spanish-wwm-cased*), a cased Spanish BERT trained with whole-word masking. The architecture (12 layers, 768 hidden units, 12 attention heads, ∼110M parameters), WordPiece tokenizer, and vocabulary are inherited unchanged from BETO. Because TULIO differs from BETO only in this additional pre-training stage, the BETO-vs.-TULIO ablation in Appendix E isolates the contribution of the Chilean Spanish adaptation corpus.

Pre-training corpus We pre-train on the *Universal and Chilean Spanish Corpus*, which combines Chilean Spanish web text (tweets, news articles, and consumer complaints) with multi-dialectal Spanish from books and the `.cl` domain partition of mC4 (Xue et al., 2021). Table G1 reports the composition by source. The corpus is dominated by Twitter (73.4%), which mirrors the dominant genre of our *Attitude* corpus and helps explain the gains observed in the in-domain Twitter setting (Appendices E and F).

Pre-training procedure Table G2 reports the pre-training hyperparameters. The objective is

Table F1: Out-of-domain evaluation under both supervised recipes. F1 mean ± SD is computed across three seeds. Precision, recall, and F1 are pooled across the three seed-level prediction sets; the 95% F1 CIs use sentence-level resampling (500 iterations).

Task (Recipe)	**Split**	**F1 mean ± SD**	**P pooled**	**R pooled**	**F1 pooled [95% CI]**
Attitude Types (Flair)	in-domain	0.523 ± 0.004	0.539	0.508	0.523 [0.504, 0.542]
Attitude Types (Flair)	ood	0.406 ± 0.015	0.438	0.378	0.406 [0.380, 0.434]
Judgment Subtypes (Flair)	in-domain	0.351 ± 0.012	0.408	0.308	0.351 [0.325, 0.378]
Judgment Subtypes (Flair)	ood	0.184 ± 0.009	0.238	0.150	0.184 [0.144, 0.229]
Attitude Types (Transformer)	in-domain	0.473 ± 0.014	0.441	0.511	0.473 [0.454, 0.493]
Attitude Types (Transformer)	ood	0.315 ± 0.001	0.267	0.383	0.315 [0.286, 0.339]
Judgment Subtypes (Transformer)	in-domain	0.339 ± 0.016	0.332	0.348	0.340 [0.314, 0.364]
Judgment Subtypes (Transformer)	ood	0.138 ± 0.010	0.133	0.144	0.138 [0.107, 0.173]

Table F2: Per-class F1 in-domain vs. OOD under both supervised recipes. F1 is pooled across the three seed-level prediction sets; n is the number of unique gold entity spans in the evaluation split.

Task (Recipe)	**Class**	**In-domain F1 (n)**	**OOD F1 (n)**
Attitude Types (Flair)	affect	0.465 (68)	0.271 (55)
Attitude Types (Flair)	appreciation	0.470 (331)	0.445 (400)
Attitude Types (Flair)	judgment	0.567 (490)	0.357 (169)
Judgment Subtypes (Flair)	capacity	0.379 (122)	0.177 (46)
Judgment Subtypes (Flair)	normality	0.108 (64)	0.028 (39)
Judgment Subtypes (Flair)	propriety	0.430 (223)	0.306 (49)
Judgment Subtypes (Flair)	tenacity	0.095 (53)	0.039 (27)
Judgment Subtypes (Flair)	veracity	0.080 (29)	0.000 (8)
Attitude Types (Transformer)	affect	0.399 (68)	0.249 (55)
Attitude Types (Transformer)	appreciation	0.419 (331)	0.330 (400)
Attitude Types (Transformer)	judgment	0.521 (490)	0.292 (169)
Judgment Subtypes (Transformer)	capacity	0.363 (122)	0.185 (46)
Judgment Subtypes (Transformer)	normality	0.200 (64)	0.034 (39)
Judgment Subtypes (Transformer)	propriety	0.406 (223)	0.234 (49)
Judgment Subtypes (Transformer)	tenacity	0.176 (53)	0.043 (27)
Judgment Subtypes (Transformer)	veracity	0.179 (29)	0.000 (8)

Table F3: Prior shift across genres: per-token rate of B-* tags by domain. Twitter rates are systematically higher, especially for *Judgment* subtypes.

Task	**Class**	**Twitter %**	**Non-Twitter %**	**Ratio**
Attitude Types	affect	1.14	1.01	1.13×
Attitude Types	appreciation	8.98	7.85	1.14×
Attitude Types	judgment	13.28	3.14	4.23×
Judgment Subtypes	capacity	3.52	0.95	3.71×
Judgment Subtypes	normality	1.56	0.73	2.15×
Judgment Subtypes	propriety	5.94	0.96	6.17×
Judgment Subtypes	tenacity	1.43	0.42	3.43×
Judgment Subtypes	veracity	0.71	0.09	8.08×

Table G1: Composition of the TULIO pre-training corpus.

Source	Documents	Share
Twitter (Chilean)	27,306,583	73.4%
mC4 (.cl domain)	8,706,681	23.4%
News (Chilean)	1,081,542	2.9%
Books (Spanish)	87,967	0.24%
Consumer complaints	31,219	0.08%
Total	**37,213,992**	100%

whole-word MLM, matching BETO's pre-training scheme. We train for a single epoch with a constant learning rate. The logged training loss decreases from 1.56 at 10% of the epoch to 1.34 at 90%.

Table G2: TULIO pre-training hyperparameters.

Setting	Value
Initialization	BETO (*bert-base-spanish-wwm-cased*)
Objective	MLM (whole-word masking)
Epochs	1
Learning rate	2×10^{-5} (constant)
Train batch size	64
Eval batch size	64
Optimizer	Adam (β_1=0.9, β_2=0.999, ϵ=10^{-8})
Random seed	13
Frameworks	Transformers 4.30.2, PyTorch 2.0.1+cu117

H Compute Environment

The main training and evaluation suite was executed on a workstation with an Intel Core i7-11700K CPU, 128 GB DDR4 RAM, three NVMe SSDs, and two NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory each. This suite used approximately 900 GPU-hours.

I Annotation and Error Examples

Content warning: The corpus examples below contain offensive or hostile language. User handles are replaced with generic identifiers. English glosses are approximate.

Boundary Rules Campaign 1 used three tests to delimit nominal groups.

1. *Permutation*: the candidate group must move as one unit within its clause.
2. *Commutation*: the group must be replaceable by a pronoun or clitic.
3. *Relative clauses*: restrictive relatives belong to the group they specify; non-restrictive relatives are excluded.

Campaign 2 assigned the most specific Attitude label supported by the context. Annotators used a broader label when a subtype could not be assigned with confidence.

We use GRP for group, APPR for appreciation, and JUDG for Judgment. The subtype labels used below are CAP for capacity, PROP for propriety, and VER for veracity.

Two-Stage Annotation The following example shows the output of each campaign. Campaign 1 marks candidate groups. Campaign 2 assigns Attitude labels to the groups that express evaluation. For readability, the second line shows only groups assigned an Attitude label.

Stage 1: *[@user1]*$_{\text{GRP}}$ *Tampoco es [justo]*$_{\text{GRP}}$ *[Señor Bóric]*$_{\text{GRP}}$ *usar [el poder que el soberano le otorga a su representante]*$_{\text{GRP}}$ *para que [este]*$_{\text{GRP}}$ *lo utilice para [su beneficio personal de sus parientes y amigos]*$_{\text{GRP}}$*, [eso]*$_{\text{GRP}}$ *es abusar y ser [deshonesto]*$_{\text{GRP}}$.

Stage 2: *@user1 Tampoco es [justo]*$_{\text{APPR}}$ *Señor Bóric usar el poder que el soberano le otorga a su representante para que este lo utilice para [su beneficio personal de sus parientes y amigos]*$_{\text{PROP}}$*, eso es abusar y ser [deshonesto]*$_{\text{VER}}$.

Gloss: "@user1 Nor is it fair, Mr. Boric, to use the power granted to a representative by the sovereign for the personal benefit of relatives and friends; that is abuse and dishonesty."

Boundary Split at an Internal Comma This recurring boundary pattern appears when the selected Flair Attitude Types run splits one gold Judgment span at an internal comma. Strict evaluation counts one false negative and two false positives.

Gold: *Jovino Novoa es [el personaje más oscuro, nefasto y fáctico de la derecha, cómplice y engendro de la dictadura de Pinochet]*$_{\text{JUDG}}$*. @user1*

Predicted: *Jovino Novoa es [el personaje más oscuro, nefasto y fáctico de la derecha]*$_{\text{JUDG}}$*, [cómplice y engendro de la dictadura de Pinochet]*$_{\text{JUDG}}$*. @user1*

Gloss: "Jovino Novoa is [the darkest and most nefarious de facto figure of the right, an accomplice and product of Pinochet's dictatorship]$_{\text{JUDG}}$."

Capacity and Propriety The selected Qwen 3.6 Plus Judgment Subtypes run preserves both boundaries but assigns the second span to Propriety instead of Capacity.

Gold excerpt: *@user1 Jaja, [idiota]*$_{\text{CAP}}$ *y [s/sentimientos]*$_{\text{CAP}}$*. . . .*

Predicted excerpt: *@user1 Jaja, [idiota]*$_{\text{CAP}}$ *y [s/sentimientos]*$_{\text{PROP}}$*. . . .*

Gloss: "@user1 Haha, [an idiot]$_{\text{CAP}}$ and [without feelings]$_{\text{CAP}}$. . . ."

J Release

The experiment code is available at https://github.com/jorgeortizfuentes/chilean-spanish-attitude-experiments.

The *Attitude* corpus and its three annotation layers are released at https://huggingface.co/datasets/jorgeortizfuentes/chilean-spanish-attitude-corpus.
The selected supervised checkpoints are grouped at https://huggingface.co/collections/jorgeortizfuentes/chilean-spanish-attitude-emnlp-2026.
The TULIO model weights and tokenizer are available at https://huggingface.co/dccuchile/tulio-chilean-spanish-bert.