Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 263-276 recibido 05-12-2024 revisado 20-01-2025 aceptado 07-02-2025

Enhancing Pragmatic Processing: A Two-Dimension Approach to Detecting Intentions in Spanish

Avances en el Procesamiento Pragmático: un Enfoque Bidimensional para la Detección de Intenciones en Español

María Miró Maestre,[1] Ernesto L. Estevanell-Valladares,[1,2]
Robiert Sepúlveda-Torres,[1] Armando Suárez Cueto[1]
[1]Department of Software and Computing Systems, University of Alicante, Spain
[2]University of Havana
maria.miro@ua.es

Abstract: Recent advancements in Natural Language Processing (NLP), driven by the impressive performance of Large Language Models (LLMs), enable studies to address more complex linguistic levels such as semantics and pragmatics. However, available resources annotated with pragmatic information remain scarce for most languages. To address this gap, we present a Spanish annotation scheme for communicative intentions comprising two typologies: one for identifying the global intention of a message and another for the intentions of its textual segments. After validating the scheme, we introduce INTENT-ES, the first Spanish corpus of tweets annotated with their global and segment intentions. We leverage this corpus to evaluate the performance of traditional Machine Learning systems and current LLMs on intention classification. Considering the results, we believe these resources will benefit the NLP research community, facilitating the evaluation of LLMs in pragmatic tasks and integrating pragmatic information into NLP systems.

Keywords: intention classification, spanish corpus, annotation scheme, large language models.

Resumen: Los avances en el Procesamiento del Lenguaje Natural (PLN), derivados del increíble rendimiento de los Grandes Modelos del Lenguaje (LLMs), han motivado la investigación de niveles lingüísticos más complejos como la semántica y la pragmática. Sin embargo, el número de recursos disponibles anotados con información pragmática todavía es muy escaso para la mayoría de idiomas. Para abordar esta brecha de investigación presentamos un esquema de anotación en español para las intenciones comunicativas que consta de dos tipologías: una para identificar la intención global de un mensaje y otra para las intenciones de sus segmentos textuales. Al validar este esquema, presentamos INTENT-ES, el primer corpus en Español de tuits anotados con sus intenciones globales y segmentales. Aprovechamos este corpus para evaluar el desempeño de sistemas tradicionales de Machine Learning y los LLMs en la tarea de clasificación de intenciones. A la vista de los resultados, creemos que estos recursos serán de provecho para la comunidad investigadora de PLN al facilitar la evaluación de LLMs en tareas pragmáticas, además de permitir la integración de información pragmática en sistemas de PLN.

Palabras clave: clasificación de intenciones, corpus en español, esquema de anotación, grandes modelos del lenguaje.

1 Introduction

Pragmatics is steadily positioning itself as an issue to investigate in NLP research. Indeed, setting aside the impressive results Large Language Models (LLMs) have achieved in various tasks, these models still show limited complex reasoning and common sense (Bang et al., 2023; Ignat et al., 2024). These mental aspects, which are fundamental to delivering a message in a human-like manner, partially rely on the intention of the message to understand its actual meaning. Thus, detecting intentions and their discriminatory features could be of interest to adequately address pragmatic-related tasks. In a similar

ISSN 1135-5948 DOI 10.26342/2025-74-18

fashion, incorporating intentions into an NLP system could enable chatbots like ChatGPT to provide responses that align more closely with users' goals. This approach may also facilitate more empathetic conversations and allow for better monitoring of users' mental states, both of which are areas that require further research (Azaria, Azoulay, and Reches, 2024).

From the perspective of theoretical linguistics, intentions have been extensively explored within the philosophy of language tradition. Authors such as Austin (1962), Searle (1969), Searle and Vanderveken (1985), Grice (1975), Dijk (1980), Sperber and Wilson (1986) or Bach (2012), have focused on defining communicative intentions and their distinguishing features, given that intentions are considered one of the fundamentals that shape the pragmatic discipline of research.

The recognition of communicative intentions as a central element in pragmatic linguistic analysis has extended to the field of NLP, particularly with the increasing focus on intention classification in current research tasks (Schopf, Arabi, and Matthes, 2023). Well-established NLP tasks, such as Question Answering (QA) (Mirzaei, Meshgi, and Sekine, 2023; Srikanth et al., 2024), Text Summarization (Zhang and Liu, 2022; Mu et al., 2023), political discourse analysis (Subramanian, Cohn, and Baldwin, 2019; Reinig, Rehbein, and Ponzetto, 2024), and sarcasm detection (Scola and Segura-Bedmar, 2021; Alnajjar and Hämäläinen, 2021) have demonstrated the advantages of incorporating intentional information to enhance system performance. However, despite the growing body of research on intentions in NLP, pragmatically annotated linguistic resources remain limited to a small subset of languages.

Consequently, this paper introduces the first Spanish annotation scheme for detecting communicative intentions in short texts. The scheme operates on two textual dimensions: the overall intention of the message and the intention of each individual segment within the message. This approach allows us to analyze how global intentions are expressed through the different intentions reflected in the segments that make up the message, thereby contributing to advancing research on the understanding of language at the pragmatic level.

Additionally, derived from the design and evaluation of the annotation scheme, we also created INTENT-ES, the first Spanish corpus of tweets annotated with communicative intentions on two textual dimensions. With this corpus, we tested the performance of different NLP models when addressing the intention classification task, aiming to analyze how accurately current NLP systems can handle classification tasks involving complex inferential aspects of language, such as communicative intentions.

Overall, the main contributions of this paper are:

- The first annotation scheme for Spanish communicative intentions in short texts encompassing two dimensions: the global intention of the message and the intention of each segment within the message.
- The first Spanish dataset of tweets annotated with both global and segment-specific intentions, encompassing a total of 8,112 tweets.
- A statistical analysis of the annotated dataset, highlighting the correlation between segment and global intentions.
- A set of experiments using traditional machine learning (ML) algorithms, deep learning (DL) methods, and LLMs to evaluate the automatic classification of communicative intentions in Spanish messages.

In the remainder of this paper, we first contextualize the notion of intentions within linguistics and NLP research. Then, we present a detailed description of the annotation scheme and the corpus we developed for intention classification in Spanish. In addition to introducing our new linguistic resources, we conduct an analysis of automatic intention classification using ML and DL methods. Finally, we discuss the results obtained and outline potential directions for future research.

2 Related Work

One of the key linguistic and philosophical theories that helped establish pragmatics as an independent field of research is the Speech Act Theory (SAT). Originally introduced by Austin (1962) and later expanded by Searle (1969), this theory defends that, depending on the words used and the context in which they are said, messages can perform actions such as ordering someone to do something or promising a future action. Consequently, these 'speech acts', i.e., the actions performed by speakers through their utterances (Yule, 2022), are governed by particular intentions, which can often be identified through linguistic features, depending on whether they are expressed directly or indirectly.

Similar ideas to the SAT can be found in the work of Grice (1957), who introduced the concept of 'implicature'. This refers to the underlying intention behind a message, which is deduced through inferential processes not always explicitly stated. Later, Dijk (1980) studied speech acts at the discourse level with the micro and macropragmatic structures. In his works, these structures are referred to as 'sequences of speech acts'. The goal of identifying these speech acts within a message is to detect their internal relationships to give a global meaning to that message, thus helping the analysis of communicative context through intentions.

These linguistic theories laid the foundation for developing pragmatic research, which has since expanded into various disciplines. One area where the study of intentions plays a crucial role is in NLP, particularly in the context of Computer-mediated Communication (CMC). CMC includes many textual genres common in Web 2.0, such as blogs, emails, social media posts, and reviews. A unique feature of CMC is that it exists at the intersection of oral and written communication, combining aspects of both channels and incorporating paratextual elements like emojis, memes, and GIFs (Herring, Stein, and Virtanen, 2013; Herring, 2019).

Given the hybrid nature of CMC, intentions in these textual genres can be expressed with different direct or inferential processes and linguistic features. As a result, there has been growing research interest in analyzing the intentions of online users across different CMC textual genres, particularly on the social media platform Twitter[1]. For instance, Zhang, Gao, and Li (2011), Vosoughi and Roy (2016), and Saha, Saha, and Bhattacharyya (2019) focused on the automatic classification of English tweets based on their intentions. These studies adapted the SAT taxonomy to fit their specific research topics. They employed different ML approaches, ranging from traditional algorithms to advanced DL methods, including transformers, to improve the state of the art in this task.

Considering the growing research on the automatic classification of English speech acts in CMC textual genres, researchers have extended this classification to languages such as Arabic (Algotiml, Elmadany, and Magdy, 2019), Portuguese (Resende de Mendonça et al., 2020), Mexican Spanish (Díaz-Torres et al., 2020), German (Plakidis and Rehm, 2022), and French (Laurenti et al., 2022). However, studies on European Spanish remain limited, with most relying on statistical analyses (Sampietro, 2017; Rodrigo, 2020; Rodrigo, 2021; Pascual, 2021) rather than utilizing NLP techniques for automatic classification.

3 Communicative Intentions Annotation Scheme

Our annotation scheme[2] establishes two sets of intention tags: one for individual message segments and another for the global intention of the message. Given the significance of the SAT in NLP research, as discussed in Section 2, we adapted the intention categories from Searle (1969) for segment-level intentions and developed new categories for global intentions, drawing on Dijk (1980)'s concept of micro and macropragmatic structures.

This dual annotation scheme enables an analysis of the relationship between segmental and global intentions, offering insights into the formation of speech acts and the role of segmental intentions in shaping a message's global intention. Within this framework, the message represents the global discourse, while its segments provide the linguistic context for interpreting meaning across the discourse, facilitating a holistic understanding of how intentions are both semantically and pragmatically conveyed.

The segment intention categories are based on the SAT taxonomy (Searle, 1969), with some modifications. In contrast, the global intention categories were designed from scratch to represent the diverse intentions that short messages can convey, particularly within CMC textual genres.

[1] Also referred to as 'X'.

[2] The annotation scheme is available online at: https://github.com/Maria3mmm/INTENT_ES_corpus

3.1 Segment intention

The first set of intentional categories is based on the taxonomy presented in the SAT (Searle, 1969), except for the "declarative" class. This intention depends on more contextual factors to perform the intended action, such as the speaker's social status or the situational context. In the context of online communication, the number of declarative segments that fulfill this action are relatively scarce compared to other categories. Thus, we set the "declarative" intention aside for this research.

Segment intentions are identified through grammatical patterns that, according to the SAT, help determine a textual segment's purpose. Thus, we adapted linguistic elements that Searle (1969) considered indicative of intentions, such as 'speech act verbs', which identify specific speech acts in particular contexts and grammatical forms. We selected a lexicon of English speech act verbs (Wierzbicka, 1987), translated them into Spanish while preserving semantic nuances, and classified them with their intention according to Searle (1969)'s taxonomy for our annotation scheme. Beyond lexical aspects, we outlined grammatical conditions under which those verbs convey the intention they are associated with. These conditions include verb tense, mood, and person in which the verbal form is conjugated, as such aspects can determine whether a segment reflects a promise, an order, or a description of an external situation (Real Academia Española, 1999; Real Academia Española, 2009; Real Academia Española, 2016). Finally, we enriched the scheme with a list of phraseological units and interjections as in Corpas Pastor (1996), who classified these pragmatic units depending on their communicative intention, also following Searle's taxonomy.

The categories to annotate segment intentions[3] are the following:

- **Representative**. The user aims at asserting something he/she believes to be true.

  [*Al hacerte amigx de #Kifkif, estás contribuyendo a la #acogida digna de las personas refugiadas LGTBI.*] [*En España se estima que una cuarta parte de las personas refugiadas son #LGTBI.*] ([When joining #Kifkif, you contribute to the dignified hosting of LGTBI refugees.] [It is estimated that a quarter the refugees in Spain are LGTBI.])[4]

- **Directive**. They aim to get the interlocutor to do something.

  [*@laraanam @mjferreiro les recomiendo un documental por hbo max que se llama 'el arma perfecta', acerca de cómo ya no solo hackean computadoras, sino también mentes y sociedades.*] ([I recommend you a documentary in hbo max called 'the perfect weapon', focused not only on how computers are hacked, but also our minds and society.])

- **Commissive**. The user commits to an action that will take place in the future.

  después de la jugada que le hicieron a cavill, na. [*de mi parte ya no recibirán un centavo más.*] (after the dirty trick they did to cavill, nope. [they won't get a single penny from my side.])

- **Expressive**. This intention reflects the mental state of the user through feelings and emotions.

  [*En nombre del colectivo #LGTBI de #Chivilcoy agradecemos siempre la predisposición a la escucha y a la acción para hacer un municipio cada día más justo.*] ([In name of Chivilcoy's LGTBI collective, we always thank the predisposition to listen and to act to make a city increasingly fair.])

3.2 Global intention

The global intention taxonomy was designed to encompass some of the most common intentions found in CMC texts. Given that many CMC genres are focused on Web 2.0 and social networks, we determined that the intention categories should be related to sharing objective and subjective information, prompting other users to perform an action, and expressing users' perceptions through emotional statements. The taxonomy comprises 13 mutually exclusive global intention categories, highlighting the fine-grained nature of this annotation scheme.

[3] The textual segments annotated with the intention category described in each section are delimited within square brackets.

[4] Examples translated into English for clarity purposes.

- **Informative**. It aims at providing impartial information about the topic mentioned in the message.

  @betoc72070977 hola. te comentamos que en la suscripción prime cuentas con amazon prime, no amazon music unlimited.

  (@betoc72070977 hi. Note that your prime subscription includes amazon prime, not amazon music unlimited.)

- **Personal opinion**. The user includes his/her personal point of view about the topic mentioned in the text.

  @joseluisrubin yo las descargué en hbo max para este viaje. ritchie debería empezar a comprar porque todas las películas suyas que he visto me han encantado.

  (@joseluisrubin i downloaded them in hbo max for this trip. I should start buying ritchie's because I've loved every single one of his films.)

- **Suggestion**. It reflects recommendations, suggestions, or invitations that the user makes so that the other person performs an action.

  ¿@mariafher2 viste jugar a maradona?. échale un vistazo a un documental que hay en hbo sobre maradona. (did you see maradona play?. check out a documentary about maradona on hbo.)

- **Command**. The person who reflects this intention urges, presses or forces the other person (or him/herself) to perform an action through commands, duties, rules or moral obligations.

  Es imprescindible que todas las siglas del colectivo LGTBI participemos de la vida política porque ese espacio también nos corresponde. Hemos de aprender del feminismo muchas cosas pero, sobre todo, que lo personal es político. (It is essential that every acronym from the LGTBI community participate in politics because it also concerns us. We have to learn many things from feminism but, above all, that personal matters are political.)

- **Request**. The user requests or asks another person to perform an action.

  *hola estoy realizando una investigación de *netflix* para mi materia de diseño, ¿podrían apoyarme contestando esta encuesta?* (hi i'm doing research on *netflix* for my design course, could you help me by answering this survey?)

- **Question**. It aims at retrieving information or an explanation of a topic through a question.

  ¿es verdad que disney compro la segunda temporada de tokyo revengers? (is it true that disney bought the second season of tokyo revenger?)

- **Threat**. With this intention, the user insinuates that an action will take place in the future in case a condition also expressed in the message is fulfilled or not.

  Como vuelva a leer que el consumo de drogas y las prácticas sexuales de riesgo son un problema específico del colectivo LGBT quemo algo. (If I read again that drug use and risky sexual practices are a problem specific to the LGBT community I'll burn something.)

- **Promise**. It aims at committing oneself to perform an action in the future or confirming the truthfulness of a statement.

  jamás le quitaré la cuenta de spotify premium a mi papá. jamás (I'll never take away my dad's spotify premium account. never.)

- **Praise**. It is used when the user wants to give a positive opinion of the situation described within the text.

  desde niña pinocchio ha sido mi historia favorita. definitivamente la versión de guillermo del toro es la mejor. (since i was a little girl pinocchio has been my favourity story. guillermo del toro's version is definitely the best.)

- **Criticism**. Here, the user gives a negative opinion of the situation described in the text.

  @ramn maroto, te lo digo como persona del colectivo lgtbi+: ¡no puedes dar más asco! (@ramn maroto, I tell you this as part of the lgtbi+ community: you can't get any more disgusting!)

- **Emotional**. It aims to express users' psychological states as feelings, emotions, thanks or excuses, among many others, regarding a situation described within the message.

soy feliz viendo steven universe, gracias a mi novia que me presta hbo (i'm happy watching steven universe, thanks to my girlfriend who lends me hbo.)

- **Desire**. It reflects users' wish for something mentioned in the text to occur.

  este hombre no se puede morir sin adaptar en las montañas de la locura. ojalá en netflix o alguna plataforma encuentre financiación. (this man cannot die without adapting at the mountains of madness. I wish he finds any funding in netflix or any other platform.)

- **Sarcasm / joke**. The user expresses the opposite of what is said by using rhetorical devices like irony or sarcasm and messages with a humorous twist.

  @igualdadlgbt perfecto, un retroceso en "materia lgtbi" es justo lo que necesitamos todos. gracias ayuso por apoyar la buena iniciativa de vox. (@igualdadlgbt perfect, a setback in "lgtbi matters" is exactly what we all need. thanks ayuso for supporting vox's great initiative.)

4 Data & Annotation

Once the annotation scheme on communicative intentions is created, validation is necessary to ensure the consistency and accuracy of the rules for identifying intentions in Spanish texts. To achieve this, we conducted several manual annotation tasks for both intention dimensions. This section details the textual data selected for the annotation task and the results of the manual annotations.

4.1 Data Collection and Filtering

We used Twitter as the textual genre for applying our intentions annotation scheme, inspired by research outlined in Section 2. Given the extensive body of NLP research focused on language use in this social platform, we merged several existing Spanish tweet corpora from the NLP community. Such is the case of the corpus from the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion (Chakravarthi et al., 2022), which consists of tweets related to the LGTBIQ+ community annotated for the task of hope speech detection. Additionally, we included another Spanish tweet corpus published by Alcaraz Mármol et al. (2023), which focuses on detecting hate speech in football-related messages.

By collecting tweets from these datasets, we aimed to create a representative sample of the intentions in our annotation scheme, especially for polarized categories like "praise", "criticism", and "threat", as well as those tied to personal feelings, such as "desire", "promise", and "emotional". Another key goal was to ensure a balanced representation of diverse topics covering all defined intentions. To achieve this, we expanded our dataset using UMUCorpusClassifier (García-Díaz et al., 2020), a tool for automatically extracting tweets based on keyword-defined topics. We focused on tweets about streaming platforms (HBO, Netflix, Amazon Prime, Disney+) to capture informative intentions like "personal opinion", "question", and "suggestion". Additionally, we collected tweets about the LGTBIQ+ community, as we considered that expanding the tweets included in García-Díaz et al. (2020)'s corpus would enrich the dataset with more specific intentions beyond expressions of hope.

Dataset	Tweets
LGTBIQ+	58,620
HBO	97,089
Netflix	117,049
Amazon Prime & Disney+	2,202
Hope Speech Shared Task (Chakravarthi et al., 2022)	1,650
Detección de odio en fútbol (Alcaraz Mármol et al., 2023)	7,483
Total	284,093

Table 1: Total number of tweets downloaded for each topic and included in each dataset for creating our corpus.

During an initial review of the downloaded tweets, we observed that the selected keywords for compiling tweets resulted in tweets in multiple languages. To filter Spanish tweets, we used a Transformer fine-tuned model of the XLM-Roberta-Base for language detection[5]. Tweets identified as Spanish with a confidence score above 0.9 were retained, reducing the dataset from 284,093 to 243,310 tweets. However, a subsequent analysis revealed that some tweets were written in Valencian/Catalan. Thus, we filtered them out using a list of personal pronouns specific to that language, resulting in 229,507 Spanish tweets to annotate.

[5] The model can be found at `https://huggingface.co/papluca/xlm-roberta-base-language-detection`

4.2 Annotation Task

We organized an annotation task with three linguistics experts to validate our annotation scheme. The goal was to assess the inter-annotator agreement when tagging intentions in the tweets collected from the three chosen topics. The experts annotated 454 tweets with their segment intentions using the INCEpTION platform (Klie et al., 2018). The inter-annotator agreement was 0.84 according to Cohen's Kappa (Cohen, 1960) and 0.86 based on Krippendorf's Alpha (Krippendorff, 2018). We used both metrics to evaluate the agreement between each pair of annotators and among the three annotators, considering the possible differences in how each annotator segmented the text, as this could result in variations in the final intention tags assigned.

Given these positive results, we curated the annotated tweets so the creator of the annotation scheme could select the correct annotation from the three provided when no majority was achieved on the labels, establishing a gold standard (Pustejovsky and Stubbs, 2012) of the scheme for this intention typology. This allowed us to use the segment-annotated tweets for the global intention annotation task to assess whether segment information was of help when determining the global intention of a tweet.

Then, we conducted the global intention annotation task with three other linguistic experts. Given the complexity of this task (with 13 mutually exclusive intention tags), we performed several annotation tests to ensure the adequacy of the annotation scheme: (1) an annotation task with 160 tweets to check if the scheme was self-explanatory; (2) an annotation task using the same 160 tweets —without informing the annotators and after a week— to test if providing segment intention tags in the messages to be globally annotated would improve the inter-annotator agreement; and (3) an annotation task with a different set of 160 tweets, already showing their segment intention tags, to assess whether the improvement in agreement was due to annotating the same set of tweets or if the segment intention information actually enhanced accuracy in annotating global intentions. The inter-annotator agreement results for the segment and global intention annotation tests are shown in Table 2.

We used Cohen's Kappa to measure inter-annotator agreement between each annotator pair[6], and Fleiss' Kappa to assess overall agreement among all three annotators.

Segment Intentions		
Test	Cohen's Kappa	Krippendorf's Alpha
1	0.84	0.86
Global Intentions		
Test	Cohen's Kappa	Fleiss' Kappa
1	0.67	0.63
2	0.71	0.67
3	**0.80**	**0.73**

Table 2: Inter-annotator agreement values for segment and global intentions.

As Table 2 shows, the inter-annotator agreement improved with each test. In fact, providing segment intention information in the global intention annotation boosted agreement on the global intention tags. Considering these results, we curated the annotations to determine the official tagging of the tweets used in the annotation tasks, completing our gold standard of tweets annotated with both global and segment intentions.

4.3 Extending the Dataset via Human in the loop

NLP tasks typically require large amounts of annotated data for accurate system training. However, manually creating linguistic resources is time-consuming and requires qualified experts. To address this challenge, we enriched our tweet dataset using a Human-in-the-loop (HITL) approach with Active Learning (AL) strategies to create a representative corpus of Spanish intention annotations in tweets.

Specifically, we followed the guidelines included in Botella-Gil et al. (2024) to annotate our dataset semi-automatically. AL is commonly used to train a ML model by selecting representative examples of the task to fulfill. This process improves the model's performance as well as reduces the amount of data to annotate (Kholghi et al., 2016). The representative examples are chosen from an unlabeled pool so they are annotated by a human and then used to train the ML model, therefore completing the HITL methodology.

For the first annotation stage, we used random sampling as our AL strategy, ensuring that the initial annotation sample had

[6] Table 2 only shows the highest results achieved with Cohen's Kappa due to space restrictions.

the same category distribution as the entire dataset. This batch of annotated texts was divided into a training (1,380) and a test set (1,620) to evaluate the task. The unusual proportion between the training (40%) and test set (60%) is because the training set will be augmented with subsequent annotations using more complex sampling strategies.

The second annotation stage aims to gather more examples for the training set. To achieve this, we applied the uncertainty sampling strategy, a common approach in AL. This strategy allows us to select samples based on the uncertainty of the model about the assigned global intention tag. To implement this, we trained a classification model to predict global intentions. We fine-tuned a pre-trained RoBERTa-base-bne language model on our training set and added a classification layer to its output. We selected RoBERTa as the text encoder due to its high performance in various NLP tasks (Fandiño et al., 2022; Liu et al., 2020).

The selected examples were pre-annotated with global intentions using the previously trained model, and with segment intentions via a segment intention classifier trained similarly. Tweets were split into sentences before classification. In each annotation iteration, human annotators reviewed 100 pre-annotated examples, after which the models were retrained, continuing HITL methodology. For further implementation details, see Botella-Gil et al. (2024). In this second annotation stage, 52 annotation batches were processed, expanding the training set to 6,492 tweets.

5 *INTENT-ES Corpus*

The INTENT-ES corpus[7] contains 8,112 annotated tweets, capturing both global and segment-level intentions according to the defined annotation scheme. INTENT-ES is the first Spanish corpus for intention classification, featuring a wide range of communicative intention categories across two discourse dimensions: the overall intention of a message and the intentions within its segments.

Table 3 shows the distribution of the different intention categories in the corpus. As for segment intentions, the "representative" class is the most represented within the corpus. Meanwhile, the "commissive" intention has the lowest number of tags, perhaps due to the more specific contexts in which users express such intention compared to the rest of the segment intention categories. Regarding global intentions, only the categories of "personal opinion", "criticism", and "emotional" have over 1,000 tags in the corpus, while "command" and "threat" are the least frequent, with fewer than 300 tags.

Segment intention	**Training**	**Test**	**Total**
Representative	10,123	2,089	12,212
Directive	3,204	780	3,984
Expressive	1,223	490	1,713
Commissive	454	219	673
Segment tags	15,004	3,578	18,582
Global intention	**Training**	**Test**	**Total**
Personal opinion	1,168	292	1,460
Criticism	909	227	1,136
Emotional	852	213	1,065
Informative	666	166	832
Suggestion	617	154	771
Request	457	114	571
Sarcasm / joke	383	95	478
Question	328	82	410
Desire	290	72	362
Promise	267	67	334
Praise	264	65	329
Command	203	51	254
Threat	88	22	110
Global tags	6,492	1,620	8,112

Table 3: Distribution of segment and global intention tags in INTENT-ES.

5.1 Intention correlation

One goal of annotating both intention dimensions is to analyze how the global intention of a message is influenced by its segment intentions, either directly or indirectly. To study this relationship, Figure 1 shows each global intention category and its segment intentions' distribution as annotated in the corpus.

The representative segment intention, as expected, is present across all global intentions. Interestingly, the directive intention also appears in all global intentions. A notable case is the "criticism" global intention, where directive segments are more prevalent than expressive ones. This phenomenon may be explained by the diverse linguistic structures in Spanish used to express criticism, many of which rely on imperative sentences or constructions involving command periphrases. This distribution suggests that "criticism" is often conveyed indirectly through directive segments.

[7] The corpus is available online at: https://github.com/Maria3mmm/INTENT_ES_corpus

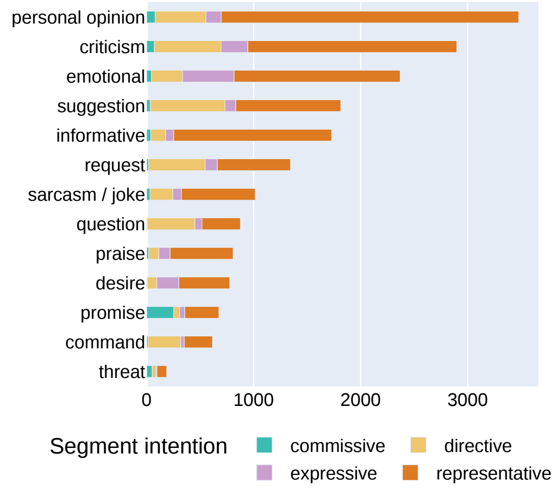

Figure 1: Segment intention correlation with each global intention in INTENT-ES.

Another finding worth commenting on is the widespread presence of the "expressive" segment intent across most of the dataset, except for the "command" and "threat" global intentions, where its representation is minimal. This suggests that these two global intentions are fulfilled via direct means with linguistic structures that express their meaning clearly, with less room for inferences.

6 *Automatic Intention Classification*

A key goal of designing the linguistic corpus INTENT-ES is to evaluate a novel task paradigm within modern NLP systems. Our primary focus is on assessing how effectively these systems can handle classification tasks that require understanding complex inferential aspects of language, such as intentions.

To thoroughly assess the current landscape of NLP regarding intention classification in Spanish, we tested three baseline approaches in our pragmatic corpus: traditional ML, Encoder-based models, and generative LLMs. These approaches were trained on 6,492 examples (training set) and evaluated on 1,620 examples (test set).

6.1 Traditional Machine Learning

We employed the AutoGOAL (Estevez-Velarde et al., 2020) AutoML framework for this category. The system was configured to optimize for the $F1_{macro}$ score over a runtime of three hours. During this period, AutoGOAL explored 214 unique algorithms and hyperparameter configuration combinations. These configurations were drawn from a pool of 123 algorithms sourced from well-established libraries such as NLTK (Bird, 2006), scikit-learn (Pedregosa et al., 2011), and AutoGOAL itself.

6.2 Encoder-based LLMs

We used the same approach created in the AL process, consisting of fine-tuning the RoBERTa-base model. In addition, we included adjusting hyperparameters to optimize performance metrics relevant to our task. The configuration and code to replicate this experiment can be found in **Intention_classifier** repository[8].

6.3 Generative LLMs

In exploring generative models, we conducted zero-shot evaluations using ChatGPT-4 with a manually curated prompt[7], assessing its ability to classify intentions without prior fine-tuning on our dataset. The zero-shot approach offers insights into the generalization capabilities of generative models when applied to new tasks without task-specific training. Additionally, we attempted intention classification with open models like LLaMA (Touvron et al., 2023); however, its 3.1-8B-Instruct version declined to process our request, arguing that the textual content was either harmful or biased and that it could not work with such type of texts.

6.4 Results & Discussion

In this section, we present and discuss the findings from our experiments on the intention classification task. We measured the baselines performance using $F1$ score class-wise and $F1_{macro}$, well-known metrics for multi-class classification problems. Table 4 shows the results obtained by our contestants for each global intention classifier.

AutoGOAL outputted a pipeline combining TF-IDF and a Passive-Aggressive Classifier from scikit-learn as its best solution. In the "desire" category, this solution almost matched the performance of the Encoder-Based contestant. However, overall, traditional ML underperformed in comparison to the other baselines. It seems that traditional models struggle with the complexity of this pragmatic task. This results from the model

[8] Code and prompt available at: `https://github.com/EEstevanell/chat-completions/tree/main` and `https://github.com/rsepulveda911112/Intention_classifier`

	TF-IDF +PAC	RoBERTa	ChatGPT 4o
Global Intentions	$F1$ **Score** (%)		
Informative	27.07	61.22	**82.72**
Personal opinion	32.29	51.76	**89.24**
Suggestion	45.84	63.25	**82.26**
Command	06.59	65.26	**87.23**
Request	46.15	73.27	**84.79**
Question	06.49	67.04	**79.02**
Threat	34.28	58.82	**93.62**
Promise	32.87	62.89	**80.33**
Praise	22.58	48.69	**73.97**
Criticism	32.68	52.85	**85.25**
Emotional	35.51	60.09	**82.98**
Desire	67.55	68.67	**89.21**
Sarcasm / Joke	15.78	28.4	**81.39**
$F1_{macro}$	31.21	58.63	**84.0**

Table 4: $F1$ score and $F1_{macro}$ for the three baselines for each intention tag.

being unaware of the linguistic nuances that differentiate each global intention category, giving these low results.

As for RoBERTa, we can arguably say it is far more competitive concerning ChatGPT than the previous baseline in intention categories as “request”, “desire” and “question”. However, we observed that this model gets generally confused when distinguishing the “personal opinion” intention from semantically similar categories like “informative”, “sarcasm / joke”, and “criticism”. This last result was quite surprising, considering that this intention is way more polarized than a “personal opinion”, which should remain objective and impersonal. We believe the issue arises from problems in fitting the model, as it appears to favor the majority class. RoBERTa also tends to mix “praise” with the “emotional” category. To improve classification results for these two intentions, as well as “criticism”, which are the most polarized categories, incorporating a sentiment analysis feature into the classification models may be beneficial. Furthermore, a multi-task learning approach could be a promising area for future exploration, as it would allow us to assess and utilize the influence of segmental intentions in estimating the overall intention.

Our ChatGPT approach got the best results among the three baselines. Good results were unsurprising as this is a state-of-the-art model in NLP tasks (Bogireddy and Dasari, 2024). However, we did not expect such high performance for the overall intention classification task, considering the results of previous studies in this kind of task (Section 2). We noted that in contrast to RoBERTa, ChatGPT confuses many “emotional” tweets as “praise” but not vice versa.

The significant performance gap between traditional ML approaches and LLMs in intention classification highlights the complexity of pragmatic understanding in computational linguistics. This gap is justifiable, considering the difference in parameter size between RoBERTa (125 million (Fandiño et al., 2022)) and GPT-4o (175+ billion[9]). However, we recommend exploring this gap with more open-source models to find good performers while minimizing resource needs compared to GPT-4o.

7 Conclusions & Future Work

In this paper, we introduced a new annotation scheme for communicative intentions in Spanish that we used to build INTENT-ES: a new corpus of Spanish tweets annotated with their intentions in two textual dimensions.

We evaluated a range of NLP models, testing their performance when addressing the intention classification task. Our results indicate that INTENT-ES envelops complex pragmatic relationships in the Spanish language. We believe this corpus could serve as a milestone in pragmatic benchmarking.

Moreover, by creating these linguistic resources, we contribute to the Spanish NLP research community by addressing complex pragmatically-based tasks that require inferential processes and contextual relations.

Our resource could assist models in detecting disinformation or harmful content, which are crucial in text classification tasks related to political discourse or fake news. Detecting the real intentions behind a political candidate’s post or a potential fake headline is vital in today’s online communication. As for Natural Language Generation (NLG), in tasks like text summarization, detecting the intentions of the original texts could help to generate summaries that better align with the text’s intended meaning.

[9] GPT-4o size has not been publicly disclosed. We estimate that GPT-4o is likely larger than GPT-3.5 (Brown, 2020) based on performance scaling laws for language models (Kaplan et al., 2020).

References

Alcaraz Mármol, G., R. Valencia García, E. Montesinos-Cánovas, F. García-Sánchez, and J. A. García-Díaz. 2023. Spanish hate-speech detection in football. *Procesamiento del lenguaje natural*, 71:15–27.

Algotiml, B., A. Elmadany, and W. Magdy. 2019. Arabic tweet-act: Speech act recognition for Arabic asynchronous conversations. In W. El-Hajj, L. H. Belguith, F. Bougares, W. Magdy, I. Zitouni, N. Tomeh, M. El-Haj, and W. Zaghouani, editors, *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 183–191, Florence, Italy. Association for Computational Linguistics.

Alnajjar, K. and M. Hämäläinen. 2021. ¡Qué maravilla! Multimodal sarcasm detection in Spanish: a dataset and a baseline. In A. Zadeh, L.-P. Morency, P. P. Liang, C. Ross, R. Salakhutdinov, S. Poria, E. Cambria, and K. Shi, editors, *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 63–68, Mexico City, Mexico, June. Association for Computational Linguistics.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford at the Clarendon Press.

Azaria, A., R. Azoulay, and S. Reches. 2024. ChatGPT is a remarkable tool—for experts. *Data Intelligence*, 6(1):240–296.

Bach, K. 2012. Saying, meaning, and implicating. In *The Cambridge Handbook of Pragmatics*. Cambridge University Press, Cambridge, pages 47–68.

Bang, Y., S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November. Association for Computational Linguistics.

Bird, S. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Bogireddy, S. R. and N. Dasari. 2024. Comparative analysis of ChatGPT-4 and LLaMA: Performance evaluation on text summarization, data analysis, and question answering. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Botella-Gil, B., R. Sepúlveda-Torres, A. Bonet-Jover, P. Martínez-Barco, and E. Saquete. 2024. Semi-automatic dataset annotation applied to automatic violent message detection. *IEEE Access*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chakravarthi, B. R., V. Muralidaran, R. Priyadharshini, S. C. Navaneethakrishnan, J. P. McCrae, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, and R. Valencia-García. 2022. Shared task on hope speech detection for equality, diversity, and inclusion - ACL. {`https://competitions.codalab.org/competitions/36393#learn_the_details-organizers`}.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Corpas Pastor, G. 1996. *Manual de fraseología española*. Gredos, Madrid.

Díaz-Torres, M. J., P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes-y Gómez, J. Aguilera, and L. Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136, Marseille, France. European Language Resources Association (ELRA).

Dijk, T. A. v. 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. L.

Erlbaum Associates, Hillsdale, New Jersey.

Estevez-Velarde, S., Y. Gutiérrez, A. Montoyo, and Y. Almeida-Cruz. 2020. Automatic discovery of heterogeneous machine learning pipelines: An application to natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3558–3568.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

Grice, H. P. 1957. Meaning. *The Philosophical Review*, 66(3).

Grice, H. P. 1975. Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, volume 3. Academic Press, New York.

Herring, S. C., 2019. *The Coevolution of Computer-Mediated Communication and Computer-Mediated Discourse Analysis*, pages 25–67. Springer International Publishing, Cham.

Herring, S. C., D. Stein, and T. Virtanen, 2013. *Introduction to the pragmatics of computer-mediated communication*, pages 3–32. De Gruyter Mouton, Berlin, Boston.

Ignat, O., Z. Jin, A. Abzaliev, L. Biester, S. Castro, N. Deng, X. Gao, A. E. Gunal, J. He, A. Kazemi, M. Khalifa, N. Koh, A. Lee, S. Liu, D. J. Min, S. Mori, J. C. Nwatu, V. Perez-Rosas, S. Shen, Z. Wang, W. Wu, and R. Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia, May. ELRA and ICCL.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Kholghi, M., L. Sitbon, G. Zuccon, and A. Nguyen. 2016. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, 23(2):289–296.

Klie, J.-C., M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, Juni.

Krippendorff, K., 2018. *α-agreement for coding.*, chapter Reliability. Sage publications.

Laurenti, E., N. Bourgon, F. Benamara, A. Mari, V. Moriceau, and C. Courgeon. 2022. Give me your intentions, I'll predict our actions: A two-level classification of speech acts for crisis management in social media. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4333–4343, Marseille, France. European Language Resources Association.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Mirzaei, M. S., K. Meshgi, and S. Sekine. 2023. What is the real intention behind this question? Dataset collection and intention classification. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13606–13622, Toronto, Canada, July. Association for Computational Linguistics.

Mu, F., X. Chen, L. Shi, S. Wang, and Q. Wang. 2023. Developer-intent driven code comment generation. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 768–780.

Pascual, D. 2021. Speech acts in travel blogs: users' corpus-driven pragmatic intentions and discursive realisations. *Elia: Estudios de lingüística inglesa aplicada*, (21):52–85.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of machine Learning research*, 12:2825–2830.

Plakidis, M. and G. Rehm. 2022. A dataset of offensive German language tweets annotated for speech acts. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4799–4807, Marseille, France. European Language Resources Association.

Pustejovsky, J. and A. Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.".

Real Academia Española. 1999. *Gramática descriptiva de la lengua española. Las construcciones sintácticas fundamentales. Relaciones temporales, aspectuales y modales.*, volume 2. Espasa Calpe, Madrid.

Real Academia Española. 2009. *Nueva gramática de la lengua española*, volume 2. Espasa, Madrid.

Real Academia Española. 2016. *Manual de la nueva gramática de la lengua española.* Grupo Planeta Spain.

Reinig, I., I. Rehbein, and S. P. Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Resende de Mendonça, R., D. Felix de Brito, F. de Franco Rosa, J. C. dos Reis, and R. Bonacin. 2020. A framework for detecting intentions of criminal acts in social media: A case study on twitter. *Information*, 11(3):154.

Rodrigo, S. R. 2020. Actos de habla expresivos en la red social Facebook. *Onomázein*, (47):225–239.

Rodrigo, S. R. 2021. Actos de habla en redes sociales: perfiles privados versus perfiles públicos. *Literatura y lingüística*, (44):429–446.

Saha, T., S. Saha, and P. Bhattacharyya. 2019. Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Sampietro, A. 2017. Emoticonos y cortesía en los mensajes de WhatsApp en España. *Español en la red. Lingüística iberoamericana*, 68:279–301.

Schopf, T., K. Arabi, and F. Matthes. 2023. Exploring the landscape of natural language processing research. In R. Mitkov and G. Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1034–1045, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Scola, E. and I. Segura-Bedmar. 2021. Sarcasm detection with BERT. *Procesamiento del Lenguaje Natural*, 67:13–25.

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

Searle, J. R. and D. Vanderveken. 1985. *Foundations of illocutionary logic.* Cambridge University Press, Cambridge.

Sperber, D. and D. Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Srikanth, N., R. Sarkar, H. Mane, E. Aparicio, Q. Nguyen, R. Rudinger, and J. Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico, June. Association for Computational Linguistics.

Subramanian, S., T. Cohn, and T. Baldwin. 2019. Target based speech act classification in political campaign text. In R. Mihalcea, E. Shutova, L.-W. Ku, K. Evang, and S. Poria, editors, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*, pages 273–282, Minneapolis, Minnesota. Association for Computational Linguistics.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

Vosoughi, S. and D. Roy. 2016. Tweet acts: A speech act classifier for Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 711–714.

Wierzbicka, A. 1987. *English Speech Act Verbs: A Semantic Dictionary.* Academic Press.

Yule, G. 2022. *The study of language.* Cambridge university press.

Zhang, L. and J. Liu. 2022. Intent-aware prompt learning for medical question summarization. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 672–679.

Zhang, R., D. Gao, and W. Li. 2011. What are tweeters doing: recognizing speech acts in Twitter. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, AAAIWS'11-05, pages 86—-91. AAAI Press.