CoWoYTP1Att: A Social Media Comment Dataset on Gender Discourse with Appraisal Theory Annotations

Valentina Tretti-Beckles[1], **Adrián Vergara-Heidke**[2], **Natalia Molina-Valverde**[2]

[1]University of Potsdam, [2]Universidad de Costa Rica

Correspondence: tretti@uni-potsdam.de, adrian.vergara@ucr.ac.cr, natalia.molinavalverde@ucr.ac.cr

Abstract

This paper presents the *Corpus on Women in YouTube on Performance with Attitude Annotations* (CoWoYTP1Att), developed based on Appraisal Theory (Martin and White, 2005). Between September 2020 and May 2021, 14,883 comments were extracted from a YouTube video featuring a compilation of the performance *Un violador en tu camino* (*A Rapist in Your Path*) by the feminist collective *LasTesis*, published on the channel of the Costa Rican newspaper La Nación. The extracted comments were manually and automatically classified based on several criteria to determine their relevance to the video. As a result, 5,939 comments were identified as related to the video. These comments were annotated with the three attitude subdomains (affect, judgement, and appreciation) proposed on the Appraisal Theory (Martin and White, 2005), as well as their polarity, target, fragment, and whether the attitude was implicit or explicit. The statistical analysis of the corpus highlights the predominant negative evaluation of individuals present in the comments on this social media platform.

1 Introduction

In December 2019, the Costa Rican newspaper La Nación published a video featuring a compilation of the performance *Un violador en tu camino* (*A Rapist in Your Path*) by the Chilean feminist collective *LasTesis*. In this video, a group of women sang and danced in protest against sexism and the violation of women's rights. Although the performance was first presented on November 25th in Santiago, Chile, women from all over the world later joined to present it in countries such as Spain, Germany, France, United Kingdom, Dominican Republic, Argentina, Colombia and Mexico (BBC News Mundo, 2019). The impact of this performance was so significant that TIME Magazine included *LasTesis* in the list *The 100 Most Influential People of 2020*.

These performances have sparked a series of reactions that have been widely shared on social media. In this study, only the comment section of the video published on YouTube is considered. A total of 14,883 comments were extracted from the video using MAXQDA, with 5,939 comments automatically identified as related to the video. Even though there are multiple videos and manifestations regarding social movements, *Un violador en tu camino* (*A Rapist in Your Path*) represents the start of a series of performances that were held along 33 countries worldwide.

The aim of this paper is to present the *Corpus on Women in YouTube on Performance with Attitude Annotations* (CoWoYTP1Att)[1], developed based on Appraisal Theory (Martin and White, 2005). This framework allows for the identification and classification of evaluations of individuals or entities, specifically those associated with feminism, women, or the performance itself.

CoWoYTP1Att was annotated with the subdomains of attitude, polarity, and the implicitness of the evaluation. This paper outlines the labels, the annotation process, and the characteristics of the annotated corpus.

2 Related Work

Appraisal Theory has been widely applied to diverse corpora, including diplomatic speeches (Anisimova and Zikánová, 2022; Anisimova and Šárka Zikánová, 2024), social media (Parameswaran et al., 2022a; Carrió-Pastor, 2025), newspaper commentaries (Arunsirot, 2012; Cavasso and Taboada, 2021; Tian et al., 2023), news articles (Tian et al., 2023), fake news (Trnavac and Põldvere, 2024), and reviews (Read et al., 2007; Mora and Lavid-López, 2018), demonstrating its versatility in analyzing evaluative language

[1] Available at: https://github.com/valentina-tretti/CoWoYTP1Att-Dataset

across textual genres.

Research indicates that the choice of appraisal domains or subdomains is influenced by the textual genre. In diplomatic speeches, judgement with a positive polarity is most common (Anisimova and Šárka Zikánová, 2024). In social media, affect dominates (Carrió-Pastor, 2025), while newspaper commentaries primarily use judgement with frequent use of negative language to intensify emotions (Cavasso and Taboada, 2021; Arunsirot, 2012). In reviews, appreciation is the second most frequent subdomain (Mora and Lavid-López, 2018). Both genuine and fake news also predominantly use judgement (Trnavac and Põldvere, 2024).

Annotating appraisal presents challenges, including difficulty identifying categories (Anisimova and Zikánová, 2022), annotator subjectivity, and disagreement among annotators, often resulting in low inter-annotator agreement (Read et al., 2007; Parameswaran et al., 2022a; Zeng et al., 2024). The need for extensive contextual understanding further complicates the process (Cavasso and Taboada, 2021; Anisimova and Zikánová, 2022; Parameswaran et al., 2022a). Consequently, some argue that automatic annotation is unreliable (Cavasso and Taboada, 2021; Parameswaran et al., 2022b), underscoring the role of linguistically trained human annotators (Parameswaran et al., 2022a).

To improve annotation consistency and facilitate future research, scholars suggest publishing the datasets (Parameswaran et al., 2022a) and sharing key annotation decisions and guidelines (Parameswaran et al., 2022a; Trnavac and Põldvere, 2024).

In recent years, research has also focused on automating the identification of appraisal. Some studies have employed lexicon-based approaches (Neviarouskaya et al., 2010) and explored methods for identifying appraisal targets (Bloom and Argamon, 2010). Additionally, Large Language Models (LLMs) have been used to detect judgement in tweets (Lan et al., 2019; Aroyehun and Gelbukh, 2020) and to classify media attitudes towards China in newspaper articles (Gao and Feng, 2025). Furthermore, a recent study by Imamovic et al. (2024) investigated the use of ChatGPT for annotating attitude subdomains in English texts.

As demonstrated in the reviewed studies, judgment is the most frequent subdomain across various genres, which we can reasonably expect to be the case in our social media corpus. Several studies highlight the inherent difficulty of annotating attitude due to its subjective nature and how this impacts inter-annotator agreement (IAA) results. Finally, despite efforts to automate the annotation of attitude, it remains necessary to have annotations reviewed by linguistically trained annotators and to develop high-quality datasets.

3 Method

3.1 Data

Between September 2020 and May 2021, 14,883 comments were extracted using MAXQDA2020[2] (VERBI Software, 2021) from a YouTube video[3] featuring a compilation of the performance *Un violador en tu camino* (*A Rapist in Your Path*) by the Chilean feminist collective *LasTesis*. The video, published in December 2019 on the YouTube channel of the Costa Rican newspaper La Nación, had gained significant public attention. The dataset consists of comments written in Spanish, encompassing various regional variations, including Latin American and Peninsular Spanish. Comments in other languages were excluded from the dataset[4].

To isolate comments relevant to the video's content, a Spanish-language transformer-based model, BETO[5] (Cañete et al., 2020), was used to classify comments as either related to the video or not related. Two sequential experiments[6] were conducted. Despite sharing the same classification objective, the experiments differed in the size and composition of their training data.

3.1.1 Objective and Label Definition

In both experiments, the classification task involved assigning one of two labels:

1. **Related to the video** ("yes"): Comments were labeled as related if they met at least one of the following criteria:

[2] More information about the software available at https://www.maxqda.com/.

[3] Video available at: https://www.youtube.com/watch?v=tB1cWh27rmI.

[4] Dataset was preprocessed including: removing emojis, punctuation marks, converting numbers to their written form, and replacing usernames with "@user"

[5] A BERT model pre-trained on a large corpus of Spanish text. Additional information available at: https://github.com/dccuchile/beto

[6] In both experiments, models were fine-tuned with batch size of 64, 8 epochs and a random seed. Training was evaluated with Cross Entropy Loss from PyTorch Library.

- Discussed aspects of the **performance** (e.g., the participants, the song and the lyrics).
- Referred to **gender-related themes** (e.g., gender differences, gender rights, privileges, the LGBTQ+ community, gender-associated occupations and military enlistment).
- Mentioned **feminists, feminism** or related themes, such as abortion.

2. **Not related to the video** ("no"): Comments were labeled as not related if they did not meet any of the above criteria and instead addressed unrelated content, such as interpersonal interactions between users or general remarks.

These two classes served as the sole labels used across both experiments. The enumerated criteria outlined above were directly employed during manual annotation and automated classification.

3.1.2 Experiment 1: Initial Model Fine-Tuning

In the first experiment, a sample of 1,200 comments was manually annotated according to the criteria described above. BETO was then fine-tuned using this annotated dataset to perform binary classification. To enhance model performance and improve data quality, a subset of the model's predictions was manually reviewed. This process resulted in the creation of a balanced dataset comprising 4,830 comments[7], with an equal number of examples labeled as "yes" and "no".

3.1.3 Experiment 2: Large-Scale Classification

The second experiment employed the balanced dataset produced in Experiment 1 to further fine-tune BETO. The resulting model was applied to the complete set of 14,883 comments, assigning each comment one of the two predefined labels. To assess model performance and inform subsequent analysis, a manual evaluation of 8,471 classified comments was conducted. This evaluation included all comments predicted as "yes" (n = 5,562) and a random sample of those predicted as "no" (n = 2,909). The distribution of predictions and manual evaluations is shown in Table 1.

Label	**Total**	**Manually Evaluated**
related	5,562	5,562
not related	9,321	2,909
Total	14,883	8,471

Table 1: Classification Results from the Second Experiment with BETO (Cañete et al., 2020).

Manual evaluation revealed that some comments initially classified as “no” had been misclassified. Following correction, the number of comments determined to be topically related to the video increased to 5,939.

3.1.4 Selection for Further Annotation

From the manually reviewed dataset, a subset of 1,500 comments was randomly selected for a subsequent phase of annotation based on the Attitude domain of Appraisal Theory. This annotation phase considered both explicit and implicit expressions of evaluative stance, including those in which the attitude target was inferred from context, previous comments, or references to the video.

3.2 Annotations

3.2.1 Annotation Framework: Appraisal Theory

Appraisal Theory (Martin and White, 2005) systematizes the subjective evaluative expression found in texts, as well as their respective gradation and presence of monoglossia or heteroglossia. It consists of three domains: attitude, engagement and gradation (see Oteíza and Pinuer, 2019). The *CoWoYTP1Att* corpus was annotated solely with the attitude domain, without considering its internal classification, which will be further developed in future studies.

The attitude domain refers to the evaluative expressions present in a text. It is divided into three subdomains: affect, judgement and appreciation. Affect pertains to the enunciator's affective reactions or dispositions toward a given propositional content. Some examples from the annotated dataset are:

- Example 1:
  - **Comment:** "@user **lástima**[8] [affect] por bestias por que "mujeres" ni a reclasificación llegan" (*@user* **pity** *[affect]*

[7] This sample and the previous one with 1,200 comments were divided into train-validation-test sets with the following distribution 75%-15%-10%.

[8] The attitudinal fragment is underlined and the evaluative word(s) are in bold .

  for beasts because "women" don't even make it to reclassification)
  - **Explanation:** The enunciator conveys disappointment about something through the use of the word "lástima" (*pity*).

- Example 2:
  - **Comment:** "Me **encanta** [affect] bailar el remix de esto a la noche (y no es por machista osea es broma pero me gusta bailar eso)" (*I **love to** dance to the remix of this at night (and not to be sexist I'm just kidding but I like to dance to that).*)
  - **Explanation:** The enunciator expresses a desire with the verb "encanta" (like) in relation to "bailar el remix de esto a la noche" (to dance to the remix of this tonight).

The judgement subdomain includes evaluations of people, objects, or institutions presented as social agents. In this study, an object or institution is considered as a social agent when it is depicted as an actor that interacts with members of society, performing actions that affect people. Such examples from the corpus are:

- Example 3:
  - **Comment:** "Eres **patetica** [judgement]" (*You're **pathetic***)
  - **Explanation:** A user evaluates another user as pathetic, using the verb "eres" (you are) to address the interlocutor.

- Example 4:
  - **Comment:** "Menuda porquería de canción. Hasta para hacer canciones somos **mejores** [judgement] los hombres..." (*What a crappy song. Even for making songs we men are **better***...)
  - **Explanation:** "Men" ("los hombres") are implicitly valued as superior to women in terms of their ability to create songs.

Lastly, the appreciation subdomain encompasses evaluative expressions about inanimate objects. This category includes appraisals based on aesthetics, quality, effect, utility, and other perspectives. Some examples from the annotated dataset are:

- Example 5:
  - **Comment:** "**Pinche** [appreciation] cumbion **bien loco** [appreciation]" (***Fucking crazy** song*)
  - **Explanation:** The "cumbion" (performance song) is evaluated as "pinche" (of poor quality) and "bien loco" (crazy).

- Example 6:
  - **Comment:** "@user y si la educacion esta **mal mal mal** [appreciation] y no se si alguien fomente la falta de respeto a los demas, yo creo que eso lo vamos aprendiendo mas por las personas..." (*@user and yes education is **wrong wrong wrong** and I do not know if anyone encourages disrespect to others, I think that we are learning more by people....*)
  - **Explanation:** The educational situation is evaluated as incorrect, with the use of "si" as an affirmative (not conditional), which also contains a spelling mistake.

In Martin and White's (2005) proposal, each of these subdomains includes a set of predefined categories. However, as demonstrated in previous studies Oteíza and Pinuer (2019), Molina Valverde and Tretti Beckles (2021), and Vergara Heidke and Tretti Beckles (2024), this internal classification is open-ended, as new categories may emerge from a fine-grained analysis of texts. Given this, we have opted not to annotate the categories within each subdomain at this stage, as a thorough analysis of the results is required.

Appraisals in a text can be explicit or implicit[9], referred to by Martin and White (2005) as inscribed and invoked, respectively. Additionally, appraisals express a polarity, meaning each fragment can be classified as either positive or negative. This is presented in examples 7 and 8.

- Example 7:
  - **Comment:** "Me **encanta** [affect-negative-yes] bailar el remix de esto a la noche (y no es por machista osea es broma pero me gusta bailar eso)" (*I **love to** dance to the remix of this at night (and*

[9] In the annotations, explicit is labeled as "no" and implicit as "yes".

not to be sexist I'm just kidding but I like to dance to that).)

- Example 8:
  - **Comment:** "Menuda porquería de canción. Hasta para hacer canciones somos mejores [judgement-positive-no] los hombres..." (*What a crappy song. Even for making songs we men are **better**...*)

The annotated fragment in example 7 is implicit (invoked). The user conveys irony (as negative polarity) through their expressed desire to dance to the remix. This interpretation is supported by the content of the comment itself, as indicated by the user's use of parentheses.

3.2.2 Annotation Process

The first annotation trial was conducted by two native Spanish speakers with a background in linguistics, who had previously worked with Appraisal Theory (Annotators A and B [10]). The annotators followed the theoretical descriptions provided by Martin and White (2005), Oteíza (2017), and Oteíza and Pinuer (2019) and annotated a set of 40 comments using the following labels:

1. **Attitude type:** affect, judgement, and appreciation.

2. **Attitude target:** the target of the annotated attitude fragment.
   - Explicit target: as it appears in the annotated fragment.
   - *Undetermined:* includes cases where there is an evaluation of the target, but the target cannot be clearly identified either by the text or the context.
   - *Implicit:* cases where the target:
     - Is mentioned in another sentence or is within the same comment, but not in the segment containing the annotated attitude fragment.
     - Is referenced in a previous comment.
     - Is inferred from context (e.g., video or theme).
   - *@user:* cases where the target is explicitly mentioned with their username in the comment.
   - *Ending:* cases where the target is not explicitly stated (e.g., through a noun or pronoun) but can be inferred from the verb conjugation, particularly in the first and second person singular and plural forms.

3. **Attitude fragment:** the span of the comment containing attitude, which could be: single words, two or more words, and entire sentences.

4. **Attitude polarity:** the sentiment of the attitude, which could be positive or negative.

Following this trial, both annotators discussed ambiguous cases. Given the nature of the comments, an additional label - **implicitness**- was introduced to indicate whether the attitude was expressed explicitly or implicitly.

In the second trial, 150 comments were annotated using the updated labeling scheme. This time, three [11] native Spanish speakers with a background in linguistics, all of whom had prior experience with Appraisal Theory (Annotators A, B, and C), participated[12]. The annotators were instructed to annotate following the attitude descriptions in Martin and White (2005), Oteíza (2017), and Oteíza and Pinuer (2019), and adhered to the following:

1. Read the comment and identify appraisals.

2. Identify the fragment spans containing attitude. A span may contain more than one attitude, and all must be annotated. If multiple attitudes exist within a span, the smallest relevant span should be annotated.

3. Assign an attitude type to each span.

4. For each identified attitude type, annotate the following:

   (a) Attitude target (explicit or implicit):
   - If explicit, annotate it as it appears in the text.
   - If implicit, follow the criteria outlined above.

   (b) Attitude polarity (positive or negative).

[10] Both are authors.

[11] Only 3 annotators participated in the annotation processes do to a lack in funding. However, following the annotation guidelines available at: `https://github.com/valentina-tretti/CoWoYTP1Att-Dataset`, more annotators could be trained to further annotate a larger sample.

[12] All annotators are authors.

(c) Attitude implicitness (explicit or implicit).

Following this second trial, annotators engaged in discussions to resolve doubtful cases, and the annotation guidelines were refined accordingly. Finally, one of the annotators reviewed all the annotations to ensure that they were similar.

3.2.3 Annotation Tool

Annotations were conducted using an Excel sheet containing both the original comments, including emojis, and their preprocessed versions, in which:

- Usernames were replaced with "@user".
- Numbers were replaced with their corresponding written version.
- Emojis and punctuation marks were removed.

The annotations were made on the preprocessed comments; however, when necessary, annotators were allowed to refer to the original comments for clarification. This was particularly useful in cases where identifying the polarity or implicitness of the attitude was challenging due to information conveyed through emojis.

The structure of the annotation file and an example are presented in Figure 1.

The number of columns in the file varied depending on the number of appraisals identified within each comment. Excel was chosen as an annotation platform for two main reasons:

1. The results from MAXQDA2020 were exported as Excel files, with each row containing a single comment.
2. Given the short length of the comments (typically one sentence or just a few words), it was deemed unnecessary to convert all files into txt format for use with the INCEpTION tool, which had initially been considered.

4 Corpus

4.1 Annotation Statistics

This section presents the statistical analysis of the corpus. The dataset comprises 1,521 comments, with a minimum length of one word and a maximum length of 345 words. These comments fall into two categories: *base comments*, which are posted directly to the YouTube video, and *response comments*, which engage with other user's remarks. The corpus consists of 564 base comments and 957 response comments.

Among the 1,521 annotated comments, 149 (9.8%) do not express an attitude, while 1,372 (90.2%) do (see Figure 2). These results indicate that the corpus is characterized by a high presence of attitudinal expressions. Moreover, they suggest that users commenting on this type of YouTube content primarily aim to express evaluations and opinions.

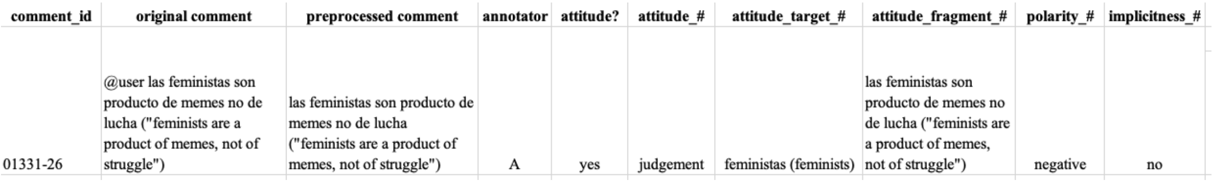
The 1,372 comments expressing attitudes contain a total of 3,107 attitudinal fragments, with an average of 2.04 instances per comment. Table 6 presents the length distribution of these annotated fragments. Fragments expressing affect are the shortest, with a maximum of 40 words, followed by appreciation with 76 words and judgement with 170 words. These results suggest that evaluations of individuals tend to require more words in Spanish, possibly because such evaluations often involve describing or explaining actions and their consequences to assess the agent performing them.

Attitude label	Min	Max	Mean	Median
affect	1	40	6.59	5
appreciation	1	76	8.50	6
judgement	1	170	10.09	9

Table 2: Number of Words per Attitude Fragment.

The distribution of attitude subdomains in the annotated comments is as follows: 2,033 (65.5%) with judgement, 720 (23.2%) with appreciation, and 353 (11.4%) with affect (see Figure 3). These findings indicate that judgement is the most prevalent subdomain, suggesting that people or animate beings are more frequently evaluated within the corpus.

Each annotated fragment was also annotated with polarity. The corpus contains 458 (14.74%) positive fragments and 2,647 (85.19%) negative fragments. The fragments with positive affect polarity are 78 (22.10%), judgement 232 (11.43%) and appreciation 1147 (20.42%). On the other hand, the fragments with negative polarity are: affect 275 (77.90%), judgement 1,798 (88.57%) and appreciation 573 (79.58%). The percentage distribution of positive and negative fragments across the attitude subdomains is shown in Figure 4. The results indicate that most attitudinal fragments are negative, regardless of whether they evaluate individuals, objects or express emotions.



comment_id	original comment	preprocessed comment	annotator	attitude?	attitude_#	attitude_target_#	attitude_fragment_#	polarity_#	implicitness_#
01331-26	@user las feministas son producto de memes no de lucha ("feminists are a product of memes, not of struggle")	las feministas son producto de memes no de lucha ("feminists are a product of memes, not of struggle")	A	yes	judgement	feministas (feminists)	las feministas son producto de memes no de lucha ("feminists are a product of memes, not of struggle")	negative	no

Figure 1: Excel Annotations File Structure.

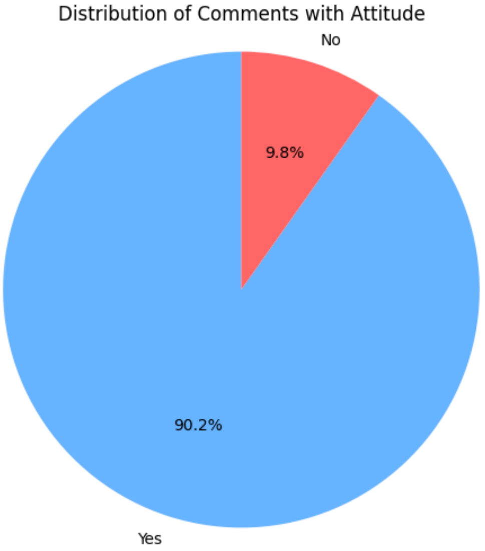


Figure 2: Distribution of Comments with Attitude.

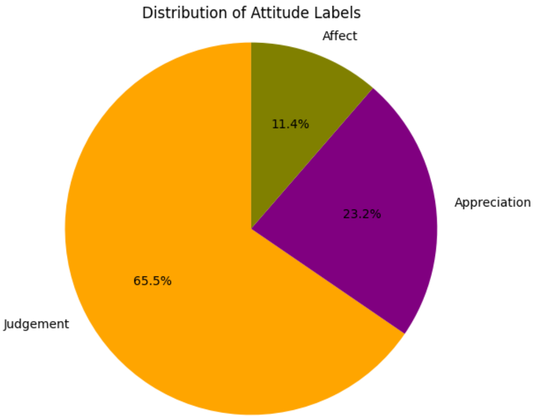


Figure 3: Distribution of Attitude Labels.

The corpus was also annotated for implicitness, yielding the following distribution: fragments with explicit valuations 2,414 (77.69%) and with implicit valuations 691 (22.24%). The fragments with explicit valuation are present in 259 (73.37.7%) of affect, in 1,547 (76.13%) of judgement and 608 (84.44%) of appreciation. The distribution of the fragments with implicit valuation is 94 (26.63%) of affect, 485 (23.87%) of judgement and 112 (15.56%) of appreciation. The percentage distribution of explicit and implicit evaluations across attitude subdomains is illustrated in Figure 5. The results indicate that most evaluations are expressed explicitly, despite the highly context-dependent meaning of social media comments.

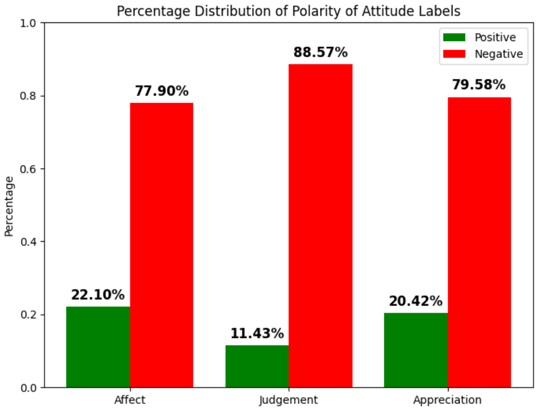


Figure 4: Percentage Distribution of Polarity and Attitude Labels.

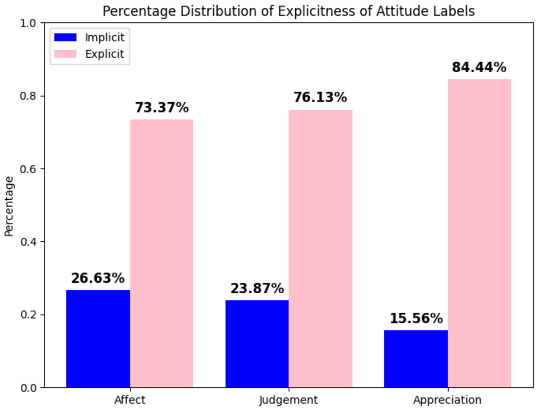


Figure 5: Percentage Distribution of Implicitness and Attitude Labels.

The relationship between polarity and implicitness is distributed as follows:

- Positive Evaluations:
  - Explicit: 416 (90.83%)
  - Implicit: 42 (9.17%)
- Negative Evaluations:

- Explicit: 1998 (75.48%)
- Implicit: 649 (24.52%)

The percentage distribution of implicitness within positive and negative polarities is shown in Figure 6. The results show that positive evaluations tend to be explicit, meaning they do not rely on contextual cues or prior knowledge of the readers. In contrast, negative evaluations exhibit a higher degree of implicitness, which may be explained by the frequent use of irony and sarcasm in the comments.

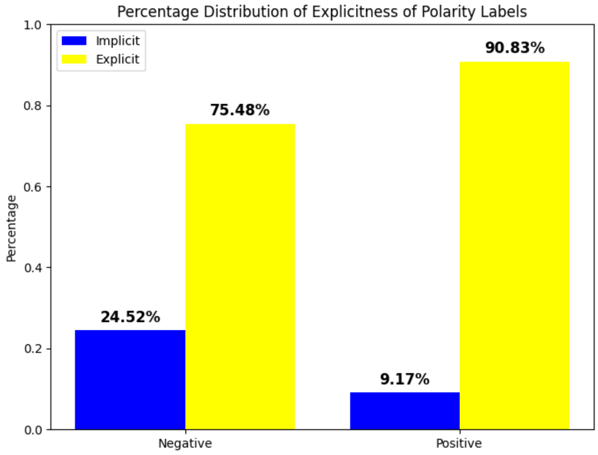


Figure 6: Percentage Distribution of Explicitness and Polarity Labels.

4.2 Inter-Annotator Agreement

4.2.1 Application of Krippendorff Alpha

To calculate inter-annotator agreement (IAA), we applied Krippendorff Alpha, because this measure allows to evaluate the agreement between more than two annotators and with categories that are not mutually exclusive (Krippendorff, 2004; Hayes and Krippendorff, 2007) [13]. In our annotated sample, the number of attitudes per comment varied among annotators. This means that one comment could have two attitudes assigned by annotator A and three attitudes assigned by annotators B and C. Additionally, the same comment could include multiple instances of the same category (e.g., *judgement*, *judgement*, *affect*), leading to discrepancies in the number of categories assigned per comment by each annotator. As a result, certain annotations contained missing values for some annotators.

Given the complexity of our annotations, we computed Krippendorff's Alpha across four different settings for all three annotators:

1. **Attitude:** Presence or absence in the comment (binary label).
2. **Attitude type:** Affect, judgement, and appreciation (see Subsection 4.2.2)
3. **Polarity:** Positive and negative (binary label).
4. **Implicitness:** Yes and no (binary label).

Table 3 presents the IAA results using Krippendorff's Alpha for the categories[14] and with an R script [15]. The results indicate a low agreement for attitude (0.38), meaning that annotators had a bad agreement on the presence or absence of valuation. For polarity (negative/positive) and implicitness (yes/no), the agreement was better, but yet low (0.46). For the attitude type the result was low too (0.35). These results suggest that there was no strong consensus between annotators, likely due to the interpretative nature of attitude identification, where each annotator's subjectivity influenced their annotations.

4.2.2 Problems with Krippendorff's Alpha

Several tests were conducted using two available tools for calculating Krippendorff's Alpha:

Test 1: We used the online K-Alpha Calculator[16] (Marzi et al., 2024). However, this tool was not suitable for our annotation scenario because it only supports mutually exclusive categories, making it inapplicable for the attitude subdomains (affect, judgement, and appreciation).

Test 2: We tested NLTK agreement metric[17] and confirmed its functionality by calculating it over a test sample with perfect agreement between three annotators, using mutually exclusive labels while allowing for missing values per comment. However, when we applied it to a test dataset structured like ours—where a comment contained multiple labels and missing values with perfect inter-annotator agreement—the metric returned a score of 0.449999 instead of 1. This result indicates that

[13] Other measures such as Cohen's Kappa, Fleiss' Kappa and Scott's Pi Coefficient were not used because they did not allow evaluation between more than two annotators or with categories that were not mutually exclusive.

[14] Calculated the score with NLTK library: https://www.nltk.org/api/nltk.metrics.agreement.html

[15] The script is available in https://github.com/valentina-tretti/CoWoYTP1Att-Dataset

[16] Available at: https://www.k-alpha.org/.

[17] Calculated the score with NLTK library: https://www.nltk.org/api/nltk.metrics.agreement.html.

the values obtained for the attitude subdomains (affect, judgement, and appreciation) were not accurate. However, the metric was valid for the binary categories (attitude, polarity and implicitness)[18].

Test 3: Thanks to the collaboration of a statistician[19], a script in R was developed to calculate the Krippendorff's Alpha for the characteristics of our dataset. This code allowed us to extract an alpha value of 0,35. However, we found a new problem. We ran a test by changing the order of the items and noticed that the result varied. This showed that the Krippendorff's Alpha is sensitive to the degree of similarity of the items. Items must be homogeneous so that the probability of being assigned certain categories is similar. As social media comments are heterogeneous, we consider that Krippendorff's Alpha is not an optimal measure to assess the degree of inter-annotator agreement when the sample is annotated with more than two annotators and multiple mutually exclusive labels.

Label	**Krippendorff Alpha Value**
attitude	0,38
attitude type	0,35
polarity	0,46
implicitness	0,46

Table 3: Inter-Annotator Agreement with Krippendorff's Alpha Metric for General Attitude, Polarity and Implicitness Labels.

4.2.3 Normalized Categorical Coincidence Index (NCCI)

Given these challenges, we developed the normalized categorical coincidence index, a formula to assess the agreement among the three annotators. This formula is only intended to show the percentage of coincidence between the annotators, to identify in which categories there might be more differences and the possible causes of these differences (e.g., problems in the guideline, subjectivity of the annotators). The formula follows these steps:

1. As in Krippendorff's Alpha (Krippendorff, 2004), each category was counted only once per item (**ci**).

2. The total number of different categories assigned per item was counted, yielding the number of categories per item (**nci**).

3. Categories annotated by more than one annotator per item were identified.

4. The occurrences of categories annotated by more than one annotator per item were counted (**nrci**).

5. The maximum number of occurrences of each category per item was calculated (**nci** × number of annotators = **mnoci**).

6. The percentage of agreement computes as

$$\text{NCCI} = \left(\frac{\text{nrci}}{\text{mnoci}} \times 100 \right)$$

Example for one item:
Annotations:

- Annotator 1: *judgment, appreciation, judgment, appreciation*

- Annotator 2: *judgment, appreciation, appreciation*

- Annotator 3: *affect, appreciation, appreciation*

Step-by-Step Calculation:

- **Categories per item (ci):**
  - Annotator 1: *judgment, appreciation*
  - Annotator 2: *judgment, appreciation*
  - Annotator 3: *affect, appreciation*

- **Number of categories per item (nci):** 3 (*judgement, appreciation, affect*)

- **Number of repeated categories per item (nrci):** 5 (*judgement, appreciation, judgement, appreciation, appreciation*)

- **Maximum possible occurrences of each category per item (mnoci):** 9 (nci x 3)

- **NCCI=**

$$\left(\frac{5}{9} \times 100 \right) = 55.5\%$$

[18] The issues encountered with the K-Alpha Calculator and the NLTK's agreement metric were also observed in other tests.

[19] We would like to thank Dr. Ricardo Alvarado-Barrantes, University of Costa Rica.

Label	NCCI
affect	55%
judgement	87%
appreciation	69%

Table 4: Inter-Annotator Agreement with NCCI Formula for Affect, Appreciation, and Judgment.

Table 4 presents the results of the inter-annotator agreement obtained using the custom formula.

The results presented in Table 4 indicate a high level of agreement among annotators regarding the presence of judgement (87%) in the comments, whereas agreement was lower for affect (55%). Based on these findings, we proceeded to analyze the distribution per label per annotator (see Table 5).

Annot.	Affect	Judgement	Appreciation
A	21 (7%)	220 (72%)	65 (21%)
B	41 (12%)	252 (73%)	54 (15%)
C	40 (16%)	160 (63%)	54 (21%)

Table 5: Percentages and Distribution per Label per Annotator.

Table 5 reveals differences in the number of annotated fragments per label among annotators: Annotator A annotated 306 fragments, Annotator B 347 fragments, and Annotator C 254 fragments. Additionally, variations in the distribution of labels across annotators are observed:

- **Annotator A:** 7% *affect*, 72% *judgement*, 21% *appreciation*.
- **Annotator B:** 12% *affect*, 73% *judgement*, 16% *appreciation*.
- **Annotator C:** 16% *affect*, 63% *judgement*, 21% *appreciation*.

The results indicate that the greatest discrepancies among annotators occur in the *judgement* and *affect* labels.

To further investigate these differences, we analyzed the length of the annotated fragments for each annotator. Tables 6, 7, and 8 present these results. The median values suggest that the fragment lengths for *affect* and *appreciation* are relatively consistent across annotators. However, *judgement* annotations exhibit notable differences, particularly in the case of Annotator C. This discrepancy suggest that annotator C tended to annotate longer *judgement* fragments, whereas Annotators A and B may have divided similar content into multiple smaller annotations. For instance, where Annotator C marked a single *judgement* fragment, Annotators A and B may have identified two separate *judgement* fragments. This would explain why Annotator C annotated significantly fewer *judgement* fragments (160) compared to Annotator A (220) and Annotator B (252).

Label	Min	Max	Mean	Median
affect	1	18	7.38	6
appreciation	1	76	11.61	8
judgement	1	170	11.98	8

Table 6: Number of Words per Attitude Fragment of Annotator A.

Label	Min	Max	Mean	Median
affect	1	65	8.48	5
appreciation	1	32	7.72	6
judgement	1	63	8.86	7

Table 7: Number of Words per Attitude Fragment of Annotator B.

In summary, the inter-annotator agreement (IAA) results highlight the influence of annotator subjectivity on the classification of evaluative categories. These findings have informed the refinement of our annotation guidelines to enhance consistency in future annotations.

5 Conclusion

This paper presents the *CoWoYTP1Att* corpus, comprising 1,521 Spanish-language internet comments on the performance *Un violador en tu camino* (*A Rapist in Your Path*), annotated using the Attitude domain of Appraisal Theory (Martin and White, 2005).

The corpus offers detailed annotations on attitude, polarity, and implicitness. The comments

Label	Min	Max	Mean	Median
affect	1	18	6.87	6
appreciation	1	32	7.57	6
judgement	1	98	15.50	11

Table 8: Number of Words per Attitude Fragment of Annotator C.

focus on gender roles and evaluations of individuals, providing valuable data for research on discourse (e.g., gender) and pragmatic phenomena (e.g., (im)politeness and speech acts).

Grounded in Appraisal Theory (Martin and White, 2005), the dataset distinguishes *affect*, *judgement*, and *appreciation*, yet aligns well with standard sentiment analysis. Polarity labels (positive/negative) match traditional sentiment classes, while Attitude types add granularity—for instance, differentiating emotions (*affect*), moral judgments (*judgement*), and aesthetic values (*appreciation*). This enables the creation of multi-label or hierarchical sentiment models that go beyond basic polarity.

By annotating both explicit and implicit attitudes, the corpus supports complex tasks such as sarcasm and stance detection, where conventional datasets often lack depth. Thus, *CoWoYTP1Att* is a valuable resource for transfer learning, domain adaptation, and building models that capture nuanced sentiment.

Corpus analysis reveals *judgement* as the most frequent subdomain, with a predominance of explicit and negative evaluations. These findings indicate that the comments are primarily concerned with negatively evaluating people (*judgement*), rather than objects or ideas (*appreciation*). Despite this negativity, linguistic strategies for implicit or mitigated evaluations are common.

Future work includes releasing the corpus in multiple formats, extending annotations, exploring automatic data augmentation, and conducting further analyses to uncover its full research potential.

Limitations

The characteristics of social media comments and the use of Appraisal Theory introduce an inherent subjectivity to the annotations. In addition, the corpus is currently only available as json and csv files, though we plan to provide alternative formats in the near future. Finally, the corpus is unbalanced, but we aim to address this imbalance in future updates.

Ethical Considerations

The comments were legally collected through MAXQDA. The content may be offensive and reflect harmful attitudes towards individuals or social groups. To ensure privacy, user names were removed to maintain the anonymity of those who posted the comments.

References

Mariia Anisimova and Šárka Zikánová. 2022. Attitude in diplomatic speeches: a pilot study. *Information technologies – Applications and Theory*.

Mariia Anisimova and Šárka Zikánová. 2024. Attitudes in diplomatic speeches: Introducing the codipa unsc 1.0. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 17–26, Torino, Italia. ELRA and ICCL.

S.T. Aroyehun and A. Gelbukh. 2020. Automatically predicting judgement dimensions of human behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 131–134.

Sudrutai Arunsirot. 2012. The use of appraisal theory to analyze thai newspaper commentaries. *Manusya Journal of Humanities*, 15(1):70–89.

BBC News Mundo. 2019. "el violador eres tú": el potente himno feminista nacido en chile que resuena en méxico, colombia, francia o españa. Accessed: 2025-03-08.

K. Bloom and S. Argamon. 2010. Unsupervised extraction of appraisal expressions. In *Canadian Conference on Artificial Intelligence*, pages 290–294. Springer.

María Luisa Carrió-Pastor. 2025. A functional classification of the aggressive digital replies to pedro sánchez' posts on x. *SSRN*.

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*.

Qingyu Gao and Denzheng William Feng. 2025. Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes. *PLoS One*, 20(1):e0313932.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Mirela Imamovic, Silvana Deilen, Dylan Glynn, and Ekaterina Lapshinova-Koltunski. 2024. Using ChatGPT for annotation of attitude within the appraisal theory: Lessons learned. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 112–123, St. Julians, Malta. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications, Thousand Oaks, CA.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

J.R. Martin and P. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator—krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Natalia Molina Valverde and Valentina Tretti Beckles. 2021. Evaluación en tiempos electorales: un acercamiento al proceso electoral desde el sistema de valoración. In *Imaginarios, subjetividades y democracia*, pages 70–99. Evaluation in electoral times: an approach to the electoral process from the appraisal theory.

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.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010)*, pages 806–814.

Teresa Oteíza. 2017. The appraisal framework and discourse analysis. In Tom Bartlett and Gerard O'Grady, editors, *The Routledge Handbook of Systemic Functional Linguistics*, chapter 28. Routledge.

Teresa Oteíza and Claudio Pinuer. 2019. El sistema de valoración como herramienta teórico-metodológica para el estudio social e ideológico del discurso. *Logos*, 29(2):207–229.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2022a. Reproducibility and automation of the appraisal taxonomy. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3731–3740, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2022b. Using aspect-based sentiment analysis to classify attitude-bearing words. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 41–51.

Jonathon Read, David Hope, and John Carroll. 2007. Annotating expressions of appraisal in english. In *Proceedings of the Linguistic Annotation Workshop*, pages 93–100, Prague, Czech Republic. Association for Computational Linguistics.

Lin Tian, Xiuzhen Zhang, Myung Hee Kim, and Jennifer Biggs. 2023. Task and sentiment adaptation for appraisal tagging. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1960–1970, Dubrovnik, Croatia. Association for Computational Linguistics.

TIME. 2020. The 100 most influential people of 2020: Lastesis. Accessed: 2025-03-08.

Radoslava Trnavac and Nele Põldvere. 2024. Investigating appraisal and the language of evaluation in fake news corpora. *Corpus Pragmatics*, 8:107–130.

VERBI Software. 2021. Maxqda 2022 [computer software]. Available from maxqda.com.

Adrián Vergara Heidke and Valentina Tretti Beckles. 2024. Actitudes semióticas a partir de grafitis en costa rica: valoraciones sobre el espacio urbano, las personas y los textos. In Gabriela Cruz Volio, Lisa Eibensteiner, Jan Harjus, and Sandra Issel-Dombert, editors, *Urban Linguistics und Linguistic Landscapes in der Romania*, Reihe Romantische Dossiers. AVM Edition, München.

Jiamei Zeng, Min Dong, and Alex Chengyu Fang. 2024. Annotating evaluative language: Challenges and solutions in applying appraisal theory. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 144–151, Torino, Italia. ELRA and ICCL.