Applying Transformer Architectures to Detect Cynical Comments in Spanish Social Media

Samuel González-López
Tecnológico Nacional/Nogales
Nogales, Sonora, México
samuel.gl@nogales.tecnm.mx

Steven Bethard
University of Arizona
Tucson, Arizona, USA
bethard@email.arizona.edu

Rogelio Platt-Molina
Tecnológico Nacional/Nogales
Nogales, Sonora, México
m22340761@nogales.tecnm.mx

Francisca Cecilia Encinas Orozco
Universidad de Sonora
Nogales, Sonora, México
cecilia.encinasorozco@unison.mx

Abstract

Detecting cynical comments in online communication poses a significant challenge in human-computer interaction, especially given the massive proliferation of discussions on platforms like YouTube. These comments often include offensive or disruptive patterns, such as sarcasm, negative feelings, specific reasons, and an attitude of being right. To address this problem, we present a web platform for the Spanish language that has been developed and leverages natural language processing and machine learning techniques. The platform detects comments and provides valuable information to users by focusing on analyzing comments. The core models are based on pre-trained architectures, including BETO, SpanBERTa, Multilingual BERT, RoBERTuito, and BERT, enabling robust detection of cynical comments. Our platform was trained and tested with Spanish comments from car analysis channels on YouTube. The results show that models achieve performance above 0.8 F1 for all types of cynical comments in the text classification task but achieve lower performance (around 0.6-0.7 F1) for the more arduous token classification task.

1 Introduction

The exponential growth of social networks has created an environment where cynical comments, such as sarcasm, negative sentiments, and dogmatic attitudes, can significantly impact discussions and public perception. In this work, we have focused on negative comments that could generate dysfunctional behaviors among social media users. Cynical behavior is a negative attitude with a broad or specific focus and comprises cognitive, affective, and behavioral components. Cynicism refers to customers' disbelief of companies or the market due to customers' perception of dishonesty and integrity on the seller's part (Indibara et al., 2023). Also, cynicism can generate feelings of betrayal and deception, leading to anger and the desire to stop purchasing products or services from the source that generates their anger (Chylinski and Chu, 2010). In this work, we have focused our efforts on the following elements: sarcasm, negative feelings, specific reasons, and attitude toward being right.

- **Sarcasm** includes mocking, biting, and cruel irony that offends or mistreats someone. Detecting sarcasm in online conversations is complex due to its subjective and contextual nature. What may be evident to a human being may be challenging to a machine. Failure to identify sarcasm can lead to misunderstandings, disagreements, and loss in quality of the online interaction (Gibbs, 2000).
- **Negative Feelings** are where users reflect negatively on a product, usually in a subjective way, influenced by their personal experiences.
- **Specific reasons** are when users identify particular aspects or components of a product, as long as the comment contains negative sentiment, sarcasm, or attitude of being right—for instance, seating comfort linked to a comment with sarcastic content.
- The **Attitude of being right** is where users express their rejection of the product and, in contrast, assert their correctness.

Such expressions come in many forms, written by users who have directly experienced the products they are commenting on and by users who have yet to consume or use the product being discussed. The automotive industry is relevant to emerging

Proceedings of the Tenth Workshop on Noisy and User-generated Text, pages 68–77
May 3, 2025 ©2025 Association for Computational Linguistics

economies (Stone and Cabrera, 2024), consumer decision-making, and the strong influence of online opinions on brand perceptions, which impacts the sales of automotive brands. By focusing on this specific domain, we seek to identify linguistic and expressive patterns characteristic of cynicism in digital communication. Furthermore, this analysis has broader implications, as the methods developed can be applied to other datasets involving product reviews, services, or online content, allowing for a better understanding of the impact of negative emotions on public opinion.

The contributions of our research are as follows:

- We collected and annotated 3705 comments in Spanish from the YouTube platform, achieving kappa of 0.841, 0.834, 0.859, and 0.752 for negative feelings, specific reasons, attitude of being right and sarcasm, respectively.
- We explore detecting cynical comments both as a token classification task and as a text classification task.
- We compare various pre-trained models to be fine-tuned for this task, including SpanBERTa, BETO, Multilingual BERT, and RoBERTuito.
- We implemented a web platform that automatically analyzes video comments using the trained models, and allows users to view each comment's predictions from each of the four models. Our models are hosted on the Hugging Face Platform.

Figure 1 shows examples of the elements analyzed in our platform. Each comment is shown in the language of study, Spanish, with its English translation.

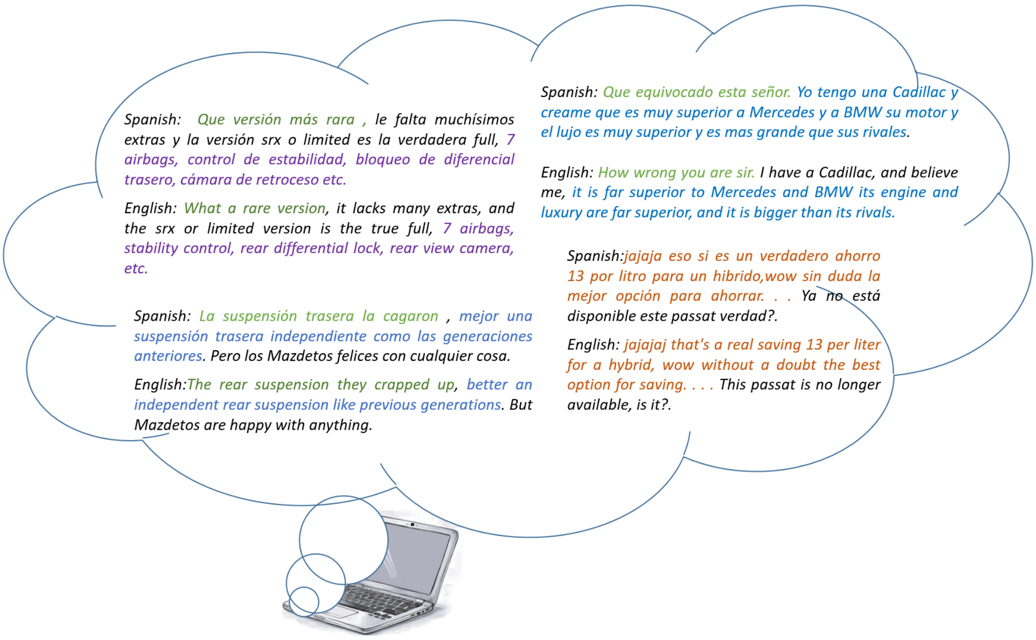
2 Related work

Cynical comments are related to negative aspect and are specific elements that characterize the dark side of consumers of products or services. The closest related work are tasks on irony and sarcasm.

Although both irony and cynicism are close because of the negativity of the content, cynicism can be understood as an extreme form of irony, in which criticism is not only insinuated but used to challenge morality and social conventions openly (Räwel, 2007). For irony detection, AlMazrua et al. (2022) created an annota d corpus of tweets with 8089 positive texts in the Arabic language. The Fleiss's Kappa agreement value was 0.54, a moderate level. This work uses machine learning and deep learning models and reports a 0.68 accuracy with the SVM algorithm. One of the challenges in this work was detecting implicit phrases as part of the irony. Maladry et al. (2022) annotate a corpus of 5566 tweets for the Dutch language, with 2783 labeled as irony. This work reported for a binary classification task a 78.98% for implicit irony and 78.88% for explicit and implicit sentiment. The SVM model performed better than the BERT model. Irony has also been approached with CNNs and Embeddings (FastText, Word2vec) (Ghanem et al., 2020). This study analyzed monolingual and multilingual architectures in three languages, with the monolingual configuration performing better. A second approach, RCNN-RoBERTa, consisting of a pre-trained RoBERTa transformer followed by bidirectional long-term memory (BiLSTM), achieved 0.80 F1 on the SemEval-2018 dataset and 0.78 F1 on the Reddit Politics dataset (Potamias et al., 2020). In a binary classification task performed on Spanish variants for Irony detection (Ortega-Bueno et al., 2019), different representation approaches, such as word embeddings (Word2Vec, FastText) and N-grams, were presented. Our research used contextual transformer representations (BETO, SpanBERTa, RoBERTuito).

Sarcasm detection has received recent NLP research, particularly within sentiment analysis, as sarcasm often leads to misinterpretations of the intended sentiment. Early models relied on traditional machine learning techniques, such as Support Vector Machines (SVM), which utilized handcrafted features like word frequency and sentiment polarity to detect sarcasm (Băroiu and Trăușan-Matu, 2022). However, these methods needed help to capture sarcasm's subtleties and context-dependent nature. Recent advancements have led to the adoption of deep learning models, including Long-Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), which have improved performance. These models can better understand the context in which sarcasm occurs, such as hyperbole, tone, or contrast between expectations and reality(Zhou, 2023). For instance, models like Cascade use context-driven approaches to capture sarcasm more accurately by analyzing dialogues on platforms like Reddit (Hazarika et al., 2018).

Further developments have seen the rise of multimodal approaches that incorporate both text and audio and visual data, which enhance detection accuracy by providing additional cues like facial expres-



Spanish: Que versión más rara , le falta muchísimos extras y la versión srx o limited es la verdadera full, 7 airbags, control de estabilidad, bloqueo de diferencial trasero, cámara de retroceso etc.

English: What a rare version, it lacks many extras, and the srx or limited version is the true full, 7 airbags, stability control, rear differential lock, rear view camera, etc.

Spanish: La suspensión trasera la cagaron , mejor una suspensión trasera independiente como las generaciones anteriores. Pero los Mazdetos felices con cualquier cosa.

English:The rear suspension they crapped up, better an independent rear suspension like previous generations. But Mazdetos are happy with anything.

Spanish: Que equivocado esta señor. Yo tengo una Cadillac y creame que es muy superior a Mercedes y a BMW su motor y el lujo es muy superior y es mas grande que sus rivales.

English: How wrong you are sir. I have a Cadillac, and believe me, it is far superior to Mercedes and BMW its engine and luxury are far superior, and it is bigger than its rivals.

Spanish:jajaja eso si es un verdadero ahorro 13 por litro para un hibrido,wow sin duda la mejor opción para ahorrar. . . Ya no está disponible este passat verdad?.

English: jajajaj that's a real saving 13 per liter for a hybrid, wow without a doubt the best option for saving. . . . This passat is no longer available, is it?.

Figure 1: Examples of cynical comments: purple corresponds to Specific Reason expression; green refers to Negative Feeling; blue corresponds to Attitude to being right cynical comments; orange corresponds to Sarcasm.

sions or intonation. Ensemble learning techniques, combining multiple models, have also improved performance in sarcasm detection by leveraging the strengths of different algorithms (Lemmens et al., 2020). Despite these advances, challenges remain, especially when identifying sarcasm in short texts (e.g., tweets) or highly nuanced expressions (Son et al., 2019). Future research will likely focus on improving model robustness in such environments and integrating more sophisticated contextual understanding (Khodak et al., 2018).

The use of AI for detecting cynicism also intersects with ethical concerns. Algorithms designed to filter harmful content sometimes over-censor, inadvertently suppressing freedom of speech by removing comments that are not genuinely harmful but might be misinterpreted by the model (Dietrich, 2024); this delicate balance between moderating harmful content and preserving free expression is a continuing challenge for AI developers. Recent work explores sentiment prediction in online communities, where AI models attempt to predict the likelihood of cynical comments based on previous patterns of behaviors (Kumar and Bhushan, 2023). While promising, these predictive models are still in the early stages and require more refinement to effectively capture the nuances of negative emotional expression. Artificial intelligence has come a long way in detecting explicitly harmful content in social networks, however, it is still difficult to accurately identify cynical negative sentiments.

3 Dataset

Our corpus was constructed in several stages. First, Spanish-language YouTube channels were selected, primarily from Latin America and focusing on new car reviews, and their video comments were downloaded. These comments were then filtered to include only those with at least ten words and five likes, ensuring sufficient text for cynicism analysis and focusing on relevant discussion. This initial filtering resulted in 3705 comments. Two human annotators independently tagged the filtered comments, freely identifying text segments containing any elements analyzed in this study. To prepare them for this task, we developed a comprehensive visual guide, including: an introduction to consumer cynicism and cynical comments: Examples of different types of cynical comments; Visual examples demonstrating the annotation process, using color coding to mark the text. The annotators, a computer science master's student, and a computer science professor, also received a description of the research context and an explanatory video. To ensure consistent annotation, a calibration stage was conducted using 50 comments from the initial pool (which were subsequently excluded from the final corpus). Inter-annotator agreement was measured by checking if one annotator's marked text segment was contained within the other's. A 90%

Cynical expressions	Count	Kappa
Negative Feelings	644	0.834
Specific Reasons	381	0.859
Attitude of being right	605	0.752
Suspicions	155	0.550
Sarcasm	256	0.841

Table 1: Dataset of Cynical Comments.

overlap was considered a match. Comments with less than 90% overlap were deemed disagreements and were excluded from the final labeled corpus, which consisted of 2041 comments. Finally, comments tagged as "Suspicions" were also excluded from the experiments due to their scarcity. Table 1 details the results of the collection.

4 Methodology

We consider two tasks for detecting cynical comments. For token classification, we use the standard inside-out-inside format for token-by-token classification. For text classification, we assigned a label to each YouTube comment as positive for a class if any part of the comment was annotated for that class and as negative if no part of the comment was annotated for that class.

We explored several pre-trained models as potential candidates for fine-tuning and subsequent evaluation on our dataset:

BETO[1] (Cañete et al., 2020) was trained following the BERT paradigm (Devlin et al., 2019), but only on Spanish documents. It is similar in size to bert-based-multilingual-cased.

SpanBERTa[2] was trained following the RoBERTa paradigm (Liu et al., 2019), but trained on 18 GB of OSCAR's Spanish corpus. It is similar in size to BERT-Base.

mBERT[3] was trained on the concatenation of monolingual Wikipedia corpora from 104 languages. Even though mBERT was trained on separate monolingual corpora without a specific multilingual training objective, it still exhibits impressive performance on a variety of multilingual tasks (Pires et al., 2019).

We further investigate a model that was specifically trained for hate speech detection. This model, which is designed to identify expressions of negativity and hostility, could potentially be directly applied to our cynicism corpus without requiring additional fine-tuning:

RoBERTuito[4] is based on the RoBERTa model architecture and the BETO tokenizer (Pérez et al., 2022). It was trained on 622M tweets in Spanish from 432k users for hate speech detection, sentiment and emotion analysis, and irony detection.

For token classification evaluation, a 10-fold cross-validation method was performed. For each cynical comment, the following BERT models were run: SpanBERTa, mBERT, and BETO. The parameters with the best performance were: 160 epochs, 3 × 10-5 of the learning rate, and a batch size of 16. The number of epochs during the fine-tuning was 20, 80, 160, and 200. The batch was computed with 16 and 32 sizes.

For text classification evaluation, training (75%), validation (12.5%), and test (12.5%) collections were constructed. For each cynical comment, the following models were run: mBERT (fine-tuned on our annotated data) and pysentimiento/robertuito (not fine-tuned on our data). We fine-tuned only mBERT because, as will be seen in the results section, there were minimal differences between mBERT and the other pre-trained models. The mBERT parameters with the best performance were: 10 epochs and a batch size of 16. However, the number of epochs during the fine-tuning was 10 and 20. EarlyStopping was also included.

After the experimentation, the best-performing models were deployed to the HuggingFace model hub, and we proceeded with the implementation of a web platform. The objective was to create an online platform where the user only places the link to the YouTube video, and the analysis is performed automatically. The framework is illustrated in Figure 2. The extraction and data processing models are executed every time a new YouTube link needs to be analyzed. The YouTube comments are extracted with Python using the "youtubecommentdownloader" API. The comments are then subjected to a cleaning, tokenization, and preprocessing process using Python. The TensorFlow models are used in the web platform through the HuggingFace API, which allows models to make predictions using the resources of that platform.

5 Results

Table 2 shows detailed results of the token classification task. The first token (B) of specific reasons

[1] https://github.com/dccuchile/beto

[2] https://github.com/chriskhanhtran/spanish-bert

[3] https://github.com/google-research/bert/

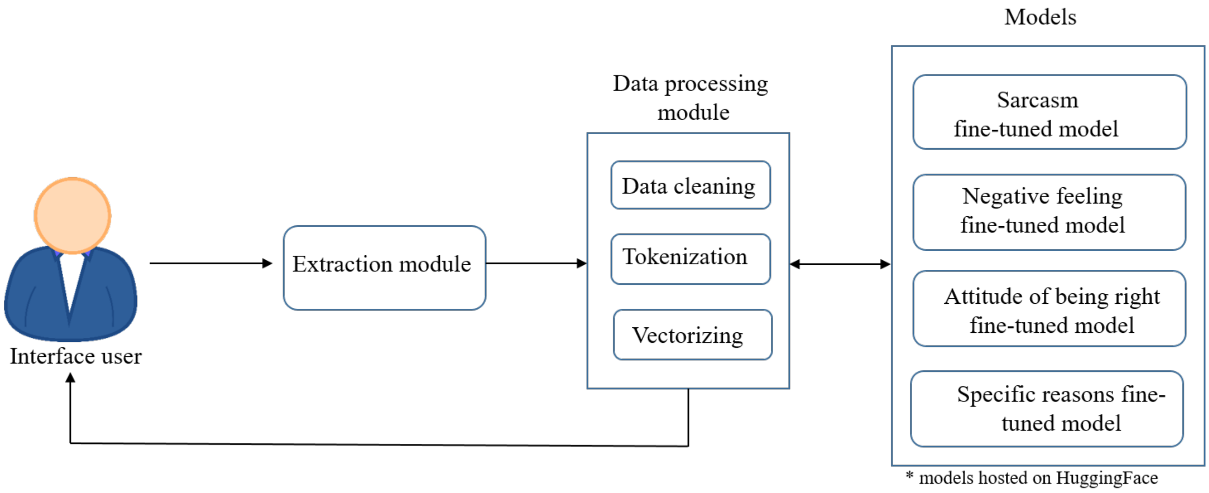

Models
Data processing module
Sarcasm fine-tuned model
Negative feeling fine-tuned model
Attitude of being right fine-tuned model
Specific reasons fine-tuned model
Data cleaning
Tokenization
Vectorizing
Extraction module
Interface user
* models hosted on HuggingFace

Figure 2: Framework for the implementation of the platform, "CODISCO".

		B			I			O		
Cynicism	Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NF	SpanBERTa	0.689	0.715	0.705	0.656	0.657	0.660	0.741	0.740	0.737
NF	BETO	0.670	0.688	0.674	0.674	0.644	0.665	0.750	0.766	0.745
NF	mBERT	0.666	0.683	0.673	0.668	0.636	0.646	0.736	0.765	0.747
SR	SpanBERTa	0.505	0.590	0.544	0.706	0.806	0.745	0.576	0.468	0.488
SR	BETO	0.507	0.642	0.565	0.742	0.841	0.778	0.612	0.470	0.500
SR	mBERT	0.510	0.575	0.538	0.711	0.816	0.749	0.610	0.480	0.502
AR	SpanBERTa	0.593	0.720	0.666	0.745	0.868	0.800	0.620	0.421	0.497
AR	BETO	0.593	0.720	0.666	0.745	0.868	0.800	0.620	0.422	0.497
AR	mBERT	0.602	0.717	0.682	0.770	0.862	0.775	0.637	0.477	0.547
SC	SpanBERTa	0.558	0.679	0.612	0.578	0.706	0.635	0.581	0.382	0.461
SC	BETO	0.558	0.685	0.615	0.580	0.745	0.665	0.620	0.383	0.473
SC	mBERT	0.567	0.676	0.616	0.572	0.770	0.656	0.610	0.438	0.509

Table 2: Detailed results on treating cynicism detection as a token classification task, for negative feelings (NF), specific reasons (SR), attitude of being right (AR), and sarcasm (SC).

were the most difficult for models to detect, with models achieving around 0.538 F1, while the inner tokens (I) of attitude of being right were the easiest, with models achieving around 0.800 F1. The different transformer models performed roughly similarly, with all F1s between comparable models within 0.04 F1 of each other. We can see that the high F1 values are distributed between the BETO and the SpanBERTa models. Sarcasm and specific reasons obtained the lowest F1 values. One possibility for this behavior was the corpus size. We can observe that tokens with label (I) for the SR and AR elements are better results than those with label (B).

Tables 3 and 4 show overall results for the token classification task (using a macro-average over the B/I/O labels) and the text classification task, respectively. As with the detailed token classifica-

Cynicism	Model	Precision	Recall	F1
		Token classification task		
NF	SpanBERTa	0.697	0.703	0.696
NF	BETO	0.694	0.700	0.693
NF	mBERT	0.691	0.695	0.690
SR	SpanBERTa	0.598	0.622	0.592
SR	BETO	0.621	0.650	0.614
SR	mBERT	0.610	0.625	0.597
AR	SpanBERTa	0.625	0.668	0.648
AR	BETO	0.653	0.668	0.649
AR	mBERT	0.668	0.685	0.670
SC	SpanBERTa	0.572	0.589	0.569
SC	BETO	0.586	0.604	0.584
SC	mBERT	0.583	0.628	0.594

Table 3: Overall results of cynical comment detection as a token classification task, for negative feelings (NF), specific reasons (SR), attitude of being right (AR), and Sarcasm (SC).

Cyn.	Model	Precision	Recall	F1
		Text classification task		
NF	mBERT (fine-tuned)	0.902	0.948	0.925
NF	RoBERTuito (not fine-tuned)	0.620	0.731	0.671
SR	mBERT (fine-tuned)	0.912	0.981	0.945
SR	RoBERTuito (not fine-tuned)	0.500	0.128	0.204
AR	mBERT (fine-tuned)	0.728	0.981	0.849
AR	RoBERTuito (not fine-tuned)	0.461	0.089	0.150
SC	mBERT (fine-tuned)	0.678	0.928	0.783
SC	RoBERTuito (not fine-tuned)	0.416	0.075	0.127

Table 4: Overall results for detecting cynicism, as a text classification task, for negative feelings (NF), specific reasons (SR), and attitude of being right (AR).

tion results, we see that there are only small differences between the different pre-trained models when fine-tuned for token classification, with SpanBERTa being slightly higher on negative feelings, BETO being slightly higher on specific reasons, and mBERT being slightly higher on attitude of being right. The hardest cynicism type to detect in a token classification task is specific reasons, while the easiest is negative feelings.

Table 4 shows that cynicism detection is easier as text classification than as token classification, with the mBERT text classifier achieving > 0.8 F1 for all cynicism types. Applying RoBERTuito without fine-tuning to this text classification task results in lower performance than our fine-tuned models, as expected. However, the fact that RoBERTuito is able to achieve 0.671 F1 on negative feeling detection without any fine-tuning on our corpus indicates that there is significant overlap between hate speech detection and negative feeling detection.

6 CODISCO Platform Interface

We evaluated several BERT-based architectures, of which three have been trained on Spanish corpora (SpanBERTa, BETO and RoBERTuito) and one was trained on multiple languages (mBERT). Our prior research suggested that models tuned for the Spanish language would obtain the best results (Gonzalez-Lopez and Bethard, 2023). However, on the current dataset, mBERT, SpanBERTa, and BETO all performed similarly. For implementing the platform we thus arbitrarily selected BETO.

We have named our platform CODISCO[5], after its acronym in Spanish (Spanish: Comportamientos Disfuncionales de los Consumidores). The APIs generated by the HuggingFace platform are the following:

- Negative Feelings HuggingFace Model
- Specific Reasons HuggingFace Model
- Attitude of being right HuggingFace Model
- Sarcasm HuggingFace Model

Figure 3 shows graphs of the results of the analysis of the comments, together with a word cloud. Figure 4 shows the percentages of each comment in detail.

As previously defined in section 4, we wanted to make the interface as easy to use as possible. So, we decided to develop a single screen where the input and output processing are performed when the user enters the internet address of a YouTube video.

6.1 Platform Output Graphics

6.1.1 Results

This section shows a global summary of the platform's analysis results: the video's title, the total number of comments extracted, and a detailed summary of the analysis results, including the number of comments classified in each evaluated characteristic (sarcasm, negative sentiments, specific reasons, and attitude of being right). This overview provides a clear perspective of the scope and nature of the comments detected in the video.

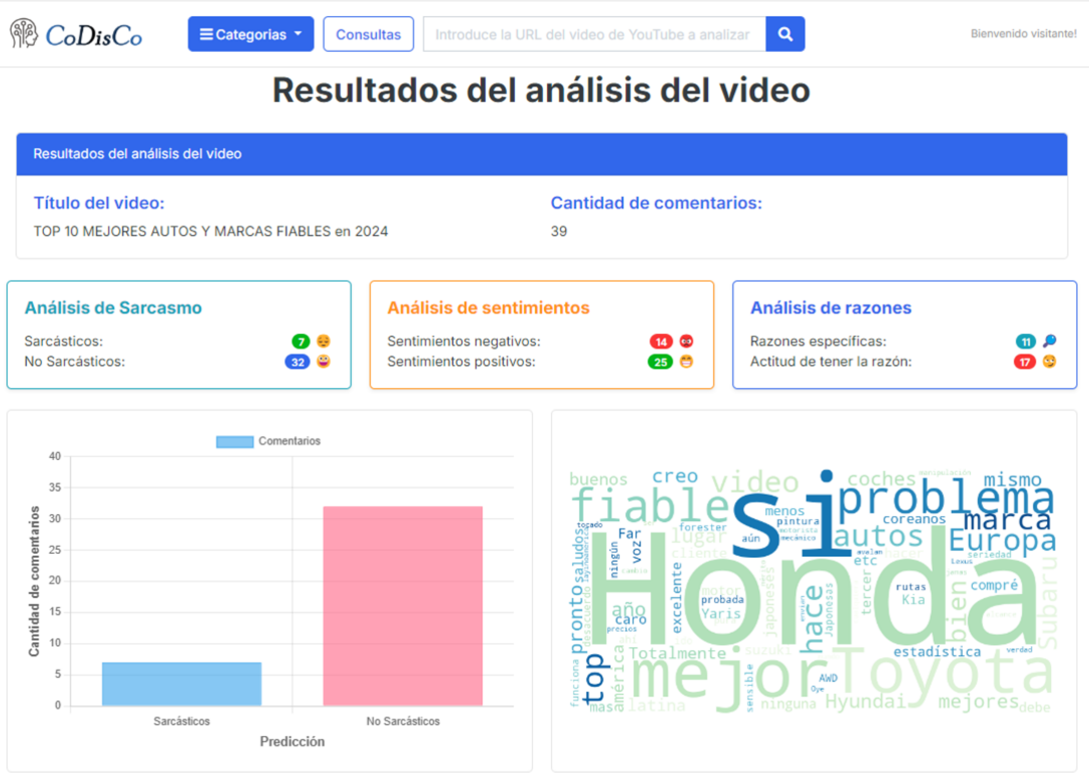
6.1.2 Bar Graph

The bar chart visualizes the number of comments classified as sarcastic versus those without sarcasm. This graphical representation allows us to quickly identify the prevalence of sarcasm in the analyzed data set. It is a valuable tool for understanding the extent of this dysfunctional behavior in the extracted comments.

6.1.3 Word Cloud

The word cloud below highlights the most frequent words found in comments classified as cynical. This visualization helps to identify linguistic patterns and recurring themes in comments containing cynicism, providing additional insights into the nature of the content analyzed. The words with the largest size in the cloud appear most frequently in this type of comment.

[5] https://www.youtube.com/watch?v=3m9I81EnLrg



Resultados del análisis del video

Resultados del análisis del video

Título del video:	Cantidad de comentarios:
TOP 10 MEJORES AUTOS Y MARCAS FIABLES en 2024	39

Análisis de Sarcasmo

Sarcásticos: 7

No Sarcásticos: 32

Análisis de sentimientos

Sentimientos negativos: 14

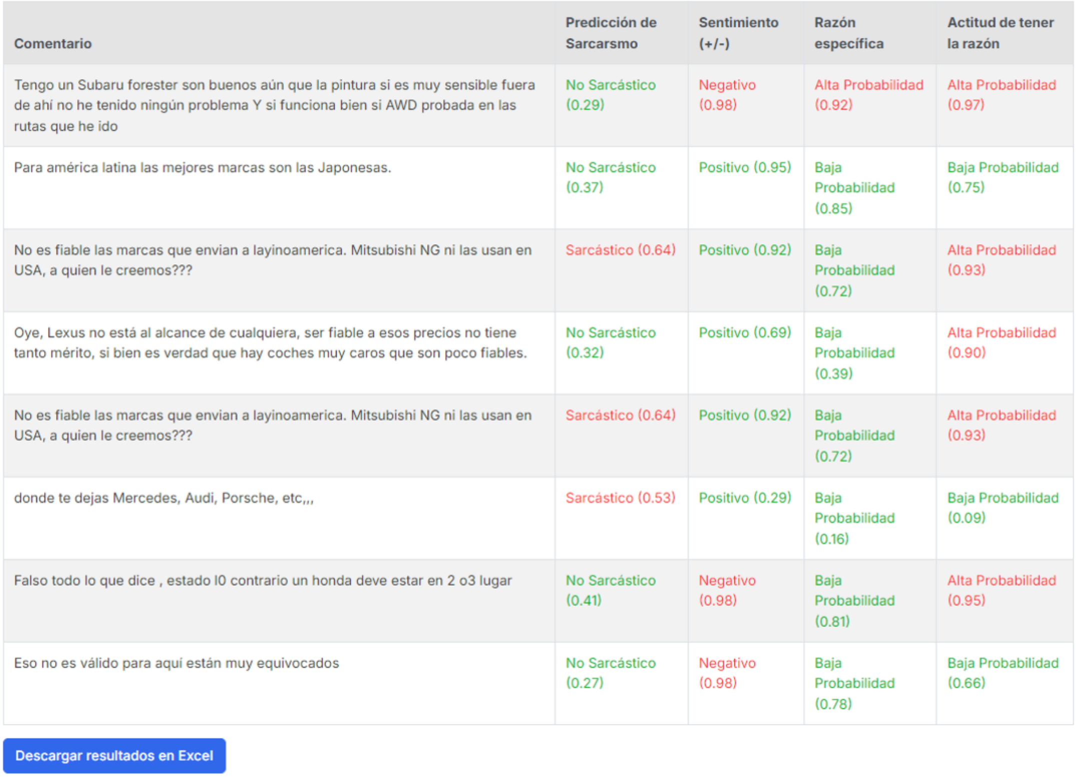
Sentimientos positivos: 25

Análisis de razones

Razones especificas: 11

Actitud de tener la razón: 17

Figure 3: Output of the analysis with General Results, Bar Graph, and Word Cloud.



Comentario	Predicción de Sarcarsmo	Sentimiento (+/-)	Razón específica	Actitud de tener la razón
Tengo un Subaru forester son buenos aún que la pintura si es muy sensible fuera de ahí no he tenido ningún problema Y si funciona bien si AWD probada en las rutas que he ido	No Sarcástico (0.29)	Negativo (0.98)	Alta Probabilidad (0.92)	Alta Probabilidad (0.97)
Para américa latina las mejores marcas son las Japonesas.	No Sarcástico (0.37)	Positivo (0.95)	Baja Probabilidad (0.85)	Baja Probabilidad (0.75)
No es fiable las marcas que envian a layinoamerica. Mitsubishi NG ni las usan en USA, a quien le creemos???	Sarcástico (0.64)	Positivo (0.92)	Baja Probabilidad (0.72)	Alta Probabilidad (0.93)
Oye, Lexus no está al alcance de cualquiera, ser fiable a esos precios no tiene tanto mérito, si bien es verdad que hay coches muy caros que son poco fiables.	No Sarcástico (0.32)	Positivo (0.69)	Baja Probabilidad (0.39)	Alta Probabilidad (0.90)
No es fiable las marcas que envian a layinoamerica. Mitsubishi NG ni las usan en USA, a quien le creemos???	Sarcástico (0.64)	Positivo (0.92)	Baja Probabilidad (0.72)	Alta Probabilidad (0.93)
donde te dejas Mercedes, Audi, Porsche, etc,,,	Sarcástico (0.53)	Positivo (0.29)	Baja Probabilidad (0.16)	Baja Probabilidad (0.09)
Falso todo lo que dice , estado l0 contrario un honda deve estar en 2 o3 lugar	No Sarcástico (0.41)	Negativo (0.98)	Baja Probabilidad (0.81)	Alta Probabilidad (0.95)
Eso no es válido para aquí están muy equivocados	No Sarcástico (0.27)	Negativo (0.98)	Baja Probabilidad (0.78)	Baja Probabilidad (0.66)

Figure 4: Detailed Output of each Comment with its Value obtained in each Category.

6.2 Usability Survey for CODISCO

We performed a survey of 40 users of the CODISCO platform. Most users found the platform responsive and effective. The scale used for the questions was 1 to 10, with 10 being a positive result. The usability survey questions were:

1. How easy was it for you to understand how to use this interface on your first attempt?
2. Did you find the interface visually appealing?
3. How satisfied are you with the response and speed of the interface?
4. How long did it take you to complete your task using this interface?
5. How intuitive did you find the functions available in the interface?

Figure 5 shows the results. The colors in the graph correspond to the five questions asked to the users, the x-axis corresponds to the users who answered the survey, and the y-axis shows the scale used.

Some users reported problems when using the platform on mobile devices, citing difficulties with the devices, mentioning difficulties with the side menu “categories”, and visualization problems. This aspect is critical as it affects the user experience and usability of the platform in mobile contexts. The speed of the interface needs improvement since it obtained low values with respect to the rest of the questions. This could have been caused by the speed of the university internet since those who used the platform and answered the survey were students from school computers. The results allowed us to make improvements to the platform.

7 Discussion

The results obtained in the experiment show that it is possible to detect the four types of cynical comments in Spanish with reasonable reliability. However, we found some points for reflection. Regarding the two tasks analyzed, we found that the performance was higher for the easier text classification task and lower for the more difficult token classification task. However, token classification is closer to the goal of this work, which is to detect exactly which part of the comment represents the cynical comment. It may be helpful to investigate two-stage approaches, in which text classification is first used to identify the general region of cynical comments, and token classification is then used to delineate specific sentences.

For comments labeled as negative feelings, the beginnings of utterances (B) were the easiest to identify, probably because they often begin with terms used to describe dissatisfaction. For comments labeled as specific reasons and attitudes of being right, the middle of utterances (I) were the easiest to identify, probably because these types of cynicism include car-specific terms that might be easier to identify. Future work could investigate whether joint learning of these models could help better establish the boundaries of the different types of cynical comments.

Experiments with RoBERTuito highlight that simply using a trained model for hate speech detection will not provide a solution for detecting cynical comments, even in the related category of negative sentiment: an adjusted RoBERTuito achieves only 0.671 F1, whereas an adjusted mBERT achieves 0.925 F1. Nevertheless, these results indicate some overlap between the two tasks, and the detection of cynical comments could benefit from the hate speech detection models, for example, by using the predictions of the hate speech model as features in the cynical comment detection model.

8 Conclusions

The analysis of cynical comments is crucial, as the sentiments and opinions of vocal customers can significantly influence decisions. Even cynical comments may induce undesirable behavior in other people. We annotated a corpus with four types of cynical comments: negative feelings, specific reasons, an attitude of being right, and sarcasm. We trained models on this corpus for text and token classification tasks.

Our results demonstrate the feasibility of training models to detect cynical comments accurately in this domain. We envision our work as a foundational step toward technologies that can quantify the level of cynicism in YouTube videos. Such analyses could empower companies to position their products strategically based on consumer perceptions. Our implementation with pre-trained models in Spanish represents a substantial advancement in comment moderation on platforms like YouTube. However, areas for improvement include expanding the corpus to encompass more dialectal variations and enhancing the model’s robustness in ambiguous contexts. We plan to fine-tune the model

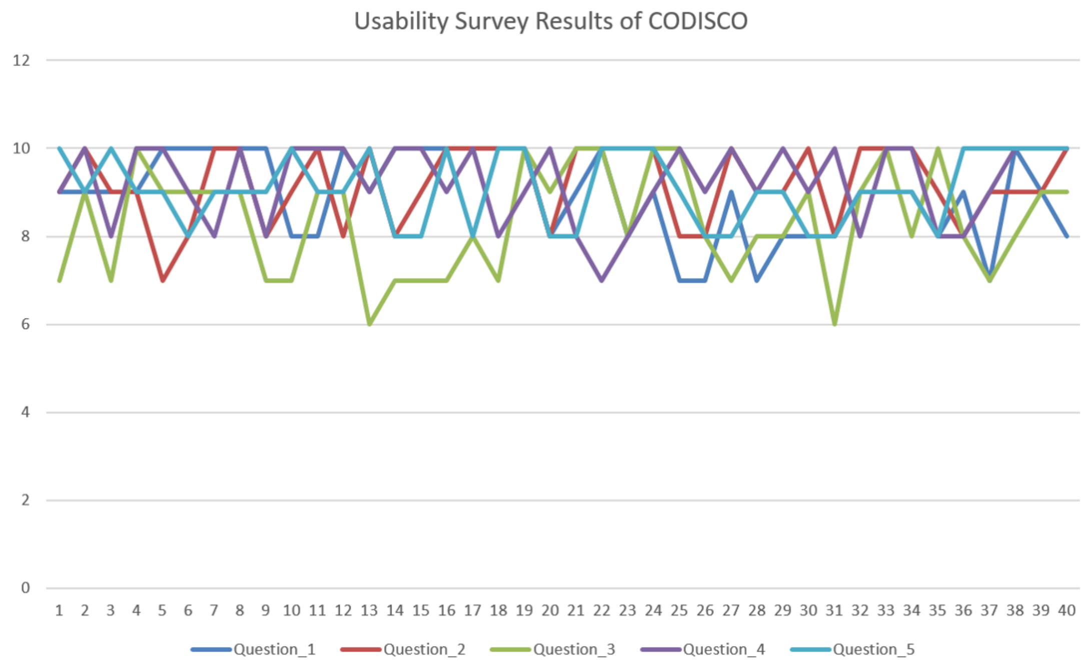


Figure 5: Usability Survey for CODISCO.

with a complementary corpus for future work. The platform has the potential to be adapted for other languages and applications beyond comment moderation, such as sentiment analysis or fake news detection.

Limitations

First, the exclusive focus on the Spanish language restricts the direct generalization of the results to other languages. While Spanish is a global language with many speakers, it is essential to recognize that the linguistic resources and language models available for Spanish do not yet reach the same scale and sophistication as those available for English. This disparity in resource availability could influence the performance and accuracy of the models evaluated in this study. In addition, specific linguistic features distinctive to Spanish, such as its richer morphology and flexible syntax, might require specific adaptations and adjustments to the language models to achieve optimal performance. Second, this study is limited to models with modest computational requirements and precludes evaluating the potential performance of the larger and more advanced language models currently available. The choice of models with modest computational requirements is justified by the need to ensure the reproducibility and accessibility of the research, allowing other researchers to replicate and extend the results obtained. The scientific community should interpret the results presented in this study in the context of the models used. It should not be considered an exhaustive evaluation of the potential of natural language processing in Spanish.

References

Halah AlMazrua, Najla AlHazzani, Amaal AlDawod, Lama AlAwlaqi, Noura AlReshoudi, Hend AlKhalifa, and Luluh AlDhubayi. 2022. Sa‘7r: A saudi dialect irony dataset. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur‘an QA and Fine-Grained Hate Speech Detection*, pages 60–70, Marseille, France. European Language Resources Association.

Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. 2022. Automatic sarcasm detection: Systematic literature review. *Information*, 13(8).

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*.

M. Chylinski and A. Chu. 2010. Consumer cynicism: antecedents and consequences. *European Journal of Marketing*, 44(6):796–837.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Frank Dietrich. 2024. Ai-based removal of hate speech from digital social networks: chances and risks for freedom of expression. *AI and Ethics*.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. 2020. Irony detection in a multilingual context. In *Advances in Information Retrieval*, pages 141–149, Cham. Springer International Publishing.

Raymond W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1-2):5–27.

Samuel Gonzalez-Lopez and Steven Bethard. 2023. Transformer-based cynical expression detection in a corpus of Spanish YouTube reviews. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 194–201, Toronto, Canada. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Indirah Indibara, Deepa Halder, and Sanjeev Varshney. 2023. Consumer cynicism: Interdisciplinary hybrid review and research agenda. *International Journal of Consumer Studies*, 47(6):2724–2746.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ankit Kumar and Bharat Bhushan. 2023. Ai driven sentiment analysis for social media data. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 1201–1206.

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 264–269, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2022. Irony detection for Dutch: a venture into the implicit. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 172–181, Dublin, Ireland. Association for Computational Linguistics.

Reynier Ortega-Bueno, Francisco Range, Delia Irazu Hernandez Farias, Paolo Rosso, Manuel Montes y Gomez, and Jose E. Medina-Pagola. 2019. Overview of the task on irony detection in spanish variants.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

R.A. Potamias, G. Siolas, and A. Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, pages 1433 – 3058.

Jörg Räwel. 2007. The Relationship between Irony, Sarcasm and Cynicism. *Z Literaturwiss Linguistik*, 37:142–153.

Le Hoang Son, Akshi Kumar, Saurabh Raj Sangwan, Anshika Arora, Anand Nayyar, and Mohamed Abdel-Basset. 2019. Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. *IEEE Access*, 7:23319–23328.

Vladimir Márquez Stone and Seyka Verónica Sandoval Cabrera. 2024. Effects of Automation on Mexican Automotive Employment: 2013–2022. *The Indian Journal of Labour Economics*, 67(3):661–680.

Juliann Zhou. 2023. An evaluation of state-of-the-art large language models for sarcasm detection. *Preprint*, arXiv:2312.03706.