Transformer-based cynical expression detection in a corpus of Spanish YouTube reviews

Samuel González-López
Technological University of Nogales
Nogales, Sonora, México
sgonzalez@utnogales.edu.mx

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

Abstract

Consumers of services and products exhibit a wide range of behaviors on social networks when they are dissatisfied. In this paper, we consider three types of cynical expressions – negative feelings, specific reasons, and attitude of being right – and annotate a corpus of 3189 comments in Spanish on car analysis channels from YouTube. We evaluate both token classification and text classification settings for this problem, and compare performance of different pre-trained models including BETO, SpanBERTa, Multilingual Bert, and RoBERTuito. The results show that models achieve performance above 0.8 F1 for all types of cynical expressions in the text classification setting, but achieve lower performance (around 0.6-0.7 F1) for the harder token classification setting.

1 Introduction

Consumers of services and products actively engage through social networks when they are dissatisfied, exhibiting a wide range of behaviors. Encinas and Cavazos (2021). Encinas presents a classification of dysfunctional consumer behaviors: mild behaviors such as rudeness, complaints, skepticism, or tantrums; moderate behaviors such as manifestations of cynicism, attempts at manipulation, or inappropriate comments and foul language; and intense consumer behaviors such as fraud, theft, verbal aggression, or revenge.

We focus on cynical expressions of consumers, specifically in comments written in videos on the Youtube platform. Cynicism is a negative attitude with a broad or specific focus and comprises cognitive, affective, and behavioral components (Chylinski and Chu, 2010). Consumer 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 (Encinas and Cavazos, 2021). Within expressions of cynicism, we focus on the following specific expressions:

Negative Feelings where consumers reflect negatively on a product, usually in a subjective way that is influenced by their personal experiences.

Specific Reasons where consumers identify the specific aspects or components of a product to which their negative feelings are directed, for instance, fuel efficiency or seating comfort.

Attitude of being right where consumers express their rejection of the product and in contrast assert their own correctness.

Such expressions come in many forms, written both by users who have directly experienced the products on which they are commenting, and by users who have yet to consume or use the product being discussed. Table 1 provides some examples of these three types of cynical expressions.

The contributions of our research are as follows:

- We collected and annotated 3189 comments in Spanish from the Youtube platform, achieving kappa of 0.834, 0.859, and 0.752 for negative feeling, specific reasons, and attitude of being right, respectively.
- We explore detection of cynical expressions both as a token classification task and as a text classification task.
- We compare a variety of pre-trained models to be fine-tuned for this task, including SpanBERTa, BETO, Multilingual BERT, and RoBERTuito.

2 Related work

The analysis of feelings is a broad field of research. Some behaviors in social media, such as offensive language, sarcasm, irony, and aggressiveness, correspond to the negative sentiment side. Cynical expressions are related to the negative aspect and

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 194–201

Spanish Example	English Translation	Expression
"La probé y se maneja bien, tiene buenos acabados, pero No me convenció su diseño, como que es difícil de digerir, siento que va ser de esos que dan el viejazo muy pronto ".	'I tried it, and it handles well, and has good finishes, but I was not convinced by its design , as it is difficult to digest, I feel that it will be one of those that give the old age very soon. '	Negative feeling and specific reason
"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 ".	' 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. '	Negative feeling and Attitude of being right
"La suspensión trasera la cagaron , mejor una suspensión trasera independiente como las generaciones anteriores. Pero los Mazdetos felices con cualquier cosa".	'The rear suspension they crapped up, better an independent rear suspension like previous generations. But Mazdetos are happy with anything.'	Negative feeling and Attitude of being right
"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, camara de retroceso etc."	'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, rearview camera, etc. '	Negative feeling and specific reason

Table 1: Examples of cynical expressions: red color corresponds to Specific Reason expression: green color refers to Negative Feeling; blue color corresponds to Attitute to being right cynical expression.

are specific elements that determine consumer cynicism.

In the field of Irony, we found a study (Al-Mazrua et al., 2022) on an annotated corpus of tweets with 8089 positive texts in the Arabic language. This work uses machine learning and deep learning models and reports a 0.68 accuracy with the SVM algorithm. The Fleiss's Kappa agreement value was 0.54, a moderate level. One of the challenges in this work was detecting implicit phrases as part of the Irony. In (Maladry et al., 2022) a corpus of 5566 tweets for the Dutch language, 2783 were 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 compared to the BERT model. Under approaches such as CNN with Embeddings (FastText, Word2vec) (Ghanem et al., 2020), the Irony was worked on. This study analyzed monolingual and multilingual architectures in three languages, with the monolingual configuration performing better. A second approach, RCNN-RoBERTa, consisting of a pretrained 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).

Very close to Irony, we find Sarcasm in the text. A paper (Alnajjar and Hämäläinen, 2021)for the Spanish language shows a dataset of text aligned to audio and video. This paper reports SVM matching results of 89% using the text alone, 91% combining audio and text, and 93.1% combining text, audio, and video. This multi-modal task is interesting since sarcasm analysis becomes domain-specific. However, adding video could generalize sarcasm detection by movements and gestures. In (Peled and Reichart, 2017) the identification of sarcasm is based on the ability to generate a non-sarcastic text from an original sarcastic text. e.g., from the sarcastic text "how I love Mondays" is obtained "how I hate Mondays" or "I really hate Mondays". In this work, the sarcasm dataset contains 3000 sarcastic tweets, each with five different non-sarcastic interpretations, and the algorithm based on Machine translation places particular emphasis on feeling words.

At a higher level, we find the feeling of aggres-

sion. Aggression can be direct or indirect and is a feeling of anger that results in hostile behavior. An analysis (Lepe-Faúndez et al., 2021) with 22 models combining the lexical and machine learning approach was performed on three corpora for Spanish (Chilean, Mexican, and Chilean-Mexican). The results show that the best performance was for the Chilean corpus with 0.89 F1, while for the Mexican corpus, it was 0.839 and 0.850 for the Chilean-Mexican combination. However, this paper highlights a higher agreement of the corpus with Chilean terms. With BERT models and an assembly strategy, a dataset tagged as non-aggressive, covertly aggressive, and overtly aggressive was classified. The assemblies achieved two percentage points higher F1-score than single models (Risch et al., 2019). Employing the same dataset but with other training features, for instance, the amount of abusive/aggressive/offensive words or the presence of hash-tags, obtain an accuracy of 73.2 % (Kumar et al., 2020).

Our research focused on consumer cynicism, annotating a new corpus for several previously unexplored cynical expressions. And unlike most previous work, which focused on the English language, our analysis of consumer cynicism focuses on the Spanish language.

3 Dataset

The corpus was generated from YouTube comments downloaded from new car analysis channels[1]. The comments were filtered, taking into account two requirements: comments must contain at least ten words, and comments must have a minimum of 5 likes. The goals of these constraints were to ensure sufficient text to judge the presence or absence of cynical expressions, and to focus on comments deemed to be relevant to the discussion. The result was a total of 3189 comments[2]. Table 2 shows some statistics of the corpus.

Two annotators were given a set of annotator guidelines containing examples of each type of cynical expression. One of the annotators was a master's student in computer science, and the second was a university teacher in computer science. The annotation guidelines had three sections: an introduction to the topic of consumer cynicism, examples of each type of cynical expression, and examples of what the annotation should look like using different colors to mark the text. The annotators were also given a description of the context of the research and a video tutorial[3]

on how to use the annotation tool. Figure 1 shows a screenshot of the annotation tool.

A group of 50 pre-training comments was used to familiarize the annotators with the annotation task. To calculate agreement between annotators, we counted two annotations as matching if the text segment of one annotator was contained within the segment the other annotator. A coverage of 90% of the matching was established. If it was lower, the text was considered a disagreement, and the document was not considered for the machine learning models. Table 2 shows agreement for the different types of cynical expressions.

We discarded the cynical expression Suspicions for having a low level of agreement, and then had the annotators annotate the remaining comments.

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

Table 2: Dataset statistics. Of 3189 annotated comments, only 1785 were coincidences among the annotators, distributed in each category.

4 Methodology

We considered two cynicism detection tasks:

Token classification We frame the cynicism detection task using the standard inside-outside-beginning format for token-by-token classification.

For evaluation, a 10-fold cross-validation method was performed. For each cynical expression, the following BERT models were run: SpanBERTa, mBERT, and BETO. The parameters with the best performance were: 160 epochs, 3 $\times$ 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.

Text classification We assigned a label to each YouTube comment as positive for a class if

[1] @autodinamico, @autossergiooliveira, @autocosmos-mx, @gonzalo_conducir, @AlonsoMaldonado0

[2] Cynical Expression Corpus for Spanish Language

[3] https://turet.com.mx/educationcorpus/TutorialEtiquetado.mp4

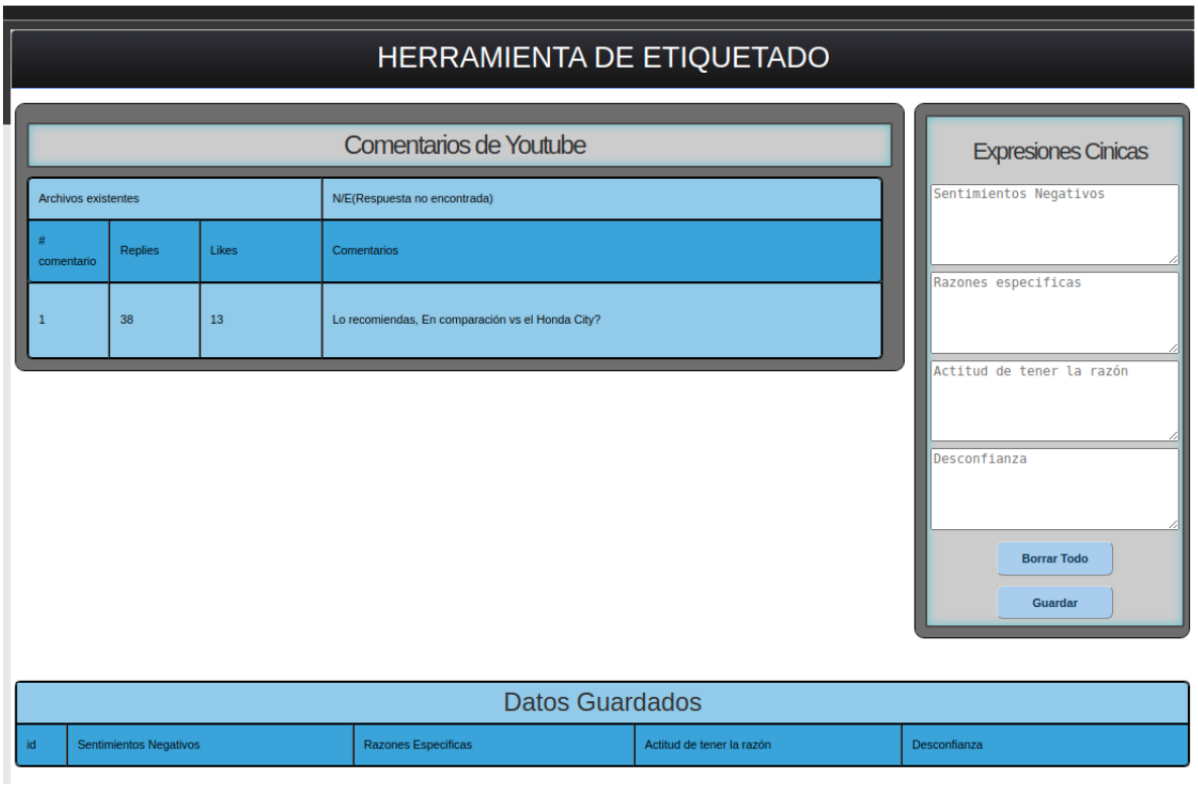

Figure 1: Interface annotation tool.

any part of the comment was annotated for that class, and as negative if none of the comment was annotated for that class.

For evaluation, we used the model (mBERT [4]). The training (75%), validation (12.5%), and test(12.5%) collections were constructed. The 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. The EarlyStopping was also included. We also applied the pysentimiento/robertuito model directly, without fine-tuning.

We considered several different pre-trained models to be fine-tuned and evaluated on our dataset:

BETO The BETO [5] model (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 The SpanBERTa model[6] 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) The Multilingual-BERT (mBERT) model was trained on the concatenation of monolingual Wikipedia corpora from 104 languages. Despite being trained on separate monolingual corpora without a multilingual target, mBERT performs well on multilingual tasks (Pires et al., 2019).

We also consider a model trained specifically for hate speech detection, which is related to negative feelings and thus has potential to be usable without fine-tuning on our cynicism corpus.

RoBERTuito The RoBERTuito model[7] is based on the RoBERTa model architecture and the BETO tokenizer (Pérez et al., 2022). It was trained on 622M tweets from 432k users for hate speech detection, sentiment and emotion analysis, and irony detection.

[4] https://github.com/google-research/bert/
[5] https://github.com/dccuchile/beto
[6] https://github.com/chriskhanhtran/spanish-bert

		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

Table 3: Detailed results on treating cynicism detection as a token classification task, for negative feelings (NF), specific reasons (SR), and attitude of being right (AR).

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
	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

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

5 Results

Table 3 shows detailed results of the token classification task. The first token (B) of specific reasons were the most difficult for models to detect, with models achieving around 0.55 F1, while the inner tokens (I) of attitude of being right were the easiest, with models achieving around 0.75 F1. The different transformer models performed roughly similarly, with all F1s between comparable models within 0.04 F1 of each other.

Table 4 shows overall results for both the token classification task (using a macro-average over the B/I/O labels) and the text classification task. As with the detailed token classification results, we see that there are only small differences between the different pre-trained models when fine-tuned

[7] https://github.com/pysentimiento/robertuito

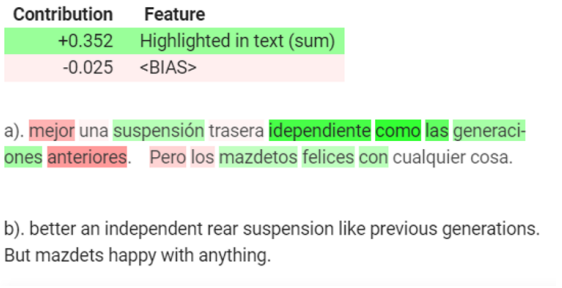

Figure 2: Specific reason example. a)Original text in Spanish, b) English translation. The green words contribute to the model prediction.

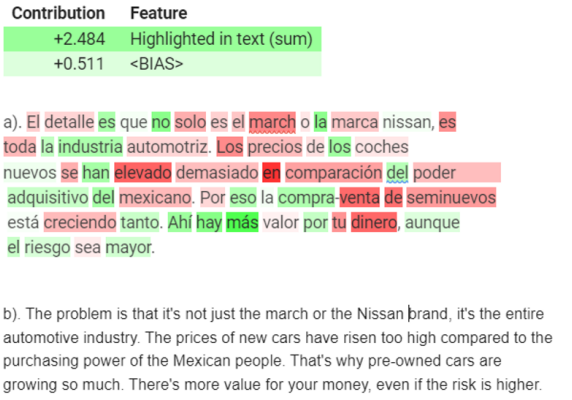

Figure 3: Attitude of being right example. a)Original text in Spanish, b) English translation. The green words contribute to the model prediction.

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.

Because of the minimal differences between the models for the token classification task, we ran only the mBERT model for text classification task. We can see from table 4 that the text classification cynicism detection task is easier than the token classification cynicism detection task, with mBERT achieving > 0.8 F1 for all cynicism types. Applying the RoBERTuito without fine-tuning to this text classification task as expected results in lower performance than our fine-tuned models 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 Explaining Cynicism Classifications

To give some insights into the behavior of our trained models, we apply LIME (Ribeiro et al., 2016) to the mBERT text classification models. In the following figures, green words contribute positively to the model prediction, and red contribute negatively to the model prediction.

Figure 2 shows an example of specific reason classification. Words like “suspension” and “independent” that relate to a car specification contribute positively, as does ‘mazdetos”, a Spanish term for owners of Mazda cars, while words like “previous” and “better” contribute negatively.

Figure 3 shows an example of attitude of being right classification. The phrase “There’s (0.237) more (0.484) value(0.017) for(0.123)” that indicates value assessment contributes positively, while words like "prices(-0.300)" and "money(-0.410)" that are characteristic of the cars have a negative impacts on the model.

Figure 4 shows an example of negative feeling classification. Words that are strongly related to negative sentiment, such as “crap(0.165)”, contribute positively to the model, but terms like “people(0.373)” and “money(0.204)” also contribute positively. Place of origin of car manufacture, “Brazil”, and the word “brands” also negatively impact the model.

7 Discussions

The results achieved in the experiment show that it is possible to detect the three cynical expressions with reasonable reliability. Some of the results are discussed below.

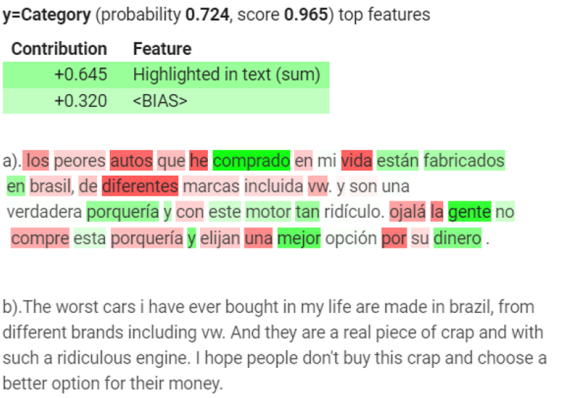

Figure 4: Negative Feeling example. a)Original text in Spanish, b) English translation. The green words contribute to the model prediction.

7.1 Token vs. text classification

Performance was higher on the easier text classification task and lower on the more challenging token classification task. However, token classification is closer to the objective of this work, detecting exactly which part of the comment represents the cynical expression. To extend the success of the text classification setting to the token classification setting, it may be useful to investigate two-stage approaches, where text classification is first used to identify the broad region of the cynical expressions and token classification is then used to narrow down to the specific phrases.

7.2 Expression keywords and boundaries

For negative feelings, the starts of the expressions (B) were easiest to identify, likely because they often start with terms used to describe dissatisfaction. For specific reasons and attitudes of being right, the middles of the expressions (I) were easiest to identify, likely because these types of cynicism include phrase-internal car-specific terms that might be easier to identify. Future work could investigate whether jointly learning such models might help to better establish the boundaries of the different types of cynical expressions.

7.3 Architecture comparison

We evaluated BERT-based architectures, of which three have been trained with Spanish corpora (SpanBERTa, BETO, and RoBERTuito) and one was trained on multiple languages (mBERT). Our expectations from some research (Cañete et al., 2020), (González-López et al., 2021) were that the language specific models would outperform the multilingual model, however, the gap between them was small. We thus conclude that the exact pretrained model selected is not a critical hyperparameter when fine-tuning models for Spanish cynical expression detection.

7.4 Cynicism vs. hate speech

The experiments with RoBERTuito highlight that simply using a model trained for hate speech detection will not provide a solution for cynical expression detection, even in the related category of negative feelings: a non-fine-tuned RoBERTuito achieves only 0.671 F1, while a fine-tuned mBERT achieves 0.925 F1. Nonetheless, these results indicate that there is some overlap between the two tasks, and cynical expression detection might benefit from hate speech detection models, for example, by using the predictions of the hate speech model as features in the cynical expression detection model.

Conclusions

The analysis of cynicism is important as the feelings and opinions of vocal customers can drive the decisions of other customers. We investigated cynicism in consumer opinions in comments on the YouTube platform. We annotated a corpus for three types of cynical expressions: negative feelings, specific reasons, and attitude of being right. We trained models on this corpus for both text classification and token classification settings. The results indicate that it is possible to train models to accurately detect cynical expressions in this domain.

We see our work as a building block towards technologies that detect and display the percentage of cynicism in YouTube videos. Such analyses could assist companies seeking to position their products based on what potential consumers think of their products. In future work, we aim to expand the corpus in size, in variety of components covered, and in types of cynical expressions included (e.g., sarcasm or irony).

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.

Khalid Alnajjar and Mika Hämäläinen. 2021. ¡Qué maravilla! multimodal sarcasm detection in Spanish: a dataset and a baseline. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 63–68, Mexico City, Mexico. Association for Computational Linguistics.

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.

F. C. Encinas and J. Cavazos. 2021. *Comportamientos disfuncionales El lado oscuro de los consumidores de servicios*, volume 1. Mc Graw-Hill Interamericana Editores, Ciudad de México.

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.

Samuel González-López, Steven Bethard, Francisca Cecilia Encinas Orozco, and Adriıan Pastor López-Monroy. 2021. Consumer cynicism identification for spanish reviews using a spanish transformer model. *Procesamiento del Lenguaje Natural*, 66(0):111–120.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Manuel Lepe-Faúndez, Alejandra Segura-Navarrete, Christian Vidal-Castro, Claudia Martínez-Araneda, and Clemente Rubio-Manzano. 2021. Detecting aggressiveness in tweets: A hybrid model for detecting cyberbullying in the spanish language. *Applied Sciences*, 11(22).

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.

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.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

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.

Juan Manuel Pérez, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. 2022. Robertuito: a pre-trained language model for social media text in spanish.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 405–410, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.