Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 43-53 recibido 07-03-2024 revisado 12-05-2024 aceptado 16-05-2024

Toxicity in Spanish News Comments and its Relationship with Constructiveness

Toxicidad en Comentarios de Noticias en Español y su Relación con la Constructividad

Pilar López-Úbeda,[1] **Flor Miriam Plaza-del-Arco**[2]
Manuel C. Díaz-Galiano[3] **M. Teresa Martín-Valdivia**[3]
[1]HT Médica
[2]Bocconi University
[3]University of Jaén
p.lopez@htmedica.com, flor.plaza@unibocconi.it,
mcdiaz@ujaen.es, maite@ujaen.es

Abstract: Online news comments are a critical source of information and opinion, but they often become a breeding ground for toxic discourse and incivility. Detecting toxicity in these comments is essential to understand and mitigate this problem. This paper presents a corpus of Spanish news comments labeled with toxicity (NECOS-TOX) and conducts a series of experiments using several machine learning algorithms, including different language models based on transformers. Our findings show that Spanish language models, such as BETO, are capable of detecting toxicity in Spanish news comments. Additionally, we investigated the relationship between toxicity and constructiveness in these comments and found that there is no clear correlation between the two factors. These results provide insights into the complexities of online discourse and highlight the need for further research to better understand the relationship between toxicity and constructiveness in Spanish news comments.
Keywords: Toxicity Detection, Natural Language Processing, Constructivenes, Spanish Linguistic Resources.

Resumen: Los comentarios en plataformas de noticias digitales constituyen una fuente esencial de información y opinión. Sin embargo, frecuentemente se transforman en focos de discurso tóxico e incivilidad. La detección de la toxicidad en dichos comentarios es fundamental para comprender y atenuar este problema. Este artículo introduce un corpus de comentarios de noticias en español, etiquetados por su toxicidad (NECOS-TOX), y realiza una serie de experimentos empleando diversos algoritmos de aprendizaje automático, incluyendo modelos de lenguaje basados en la arquitectura de *transformers*. Los resultados obtenidos demuestran que los modelos de lenguaje específicos para el español, como BETO, poseen la capacidad de identificar la toxicidad en los comentarios de noticias en español. Adicionalmente, se exploró la relación existente entre la toxicidad y la constructividad en estos comentarios, concluyendo que no se aprecia una correlación evidente entre ambos factores. Estos hallazgos aportan luz sobre las complejidades inherentes al discurso en línea y subrayan la necesidad imperante de realizar investigaciones adicionales para comprender de manera más profunda la relación entre la toxicidad y la constructividad en los comentarios de noticias en español.
Palabras clave: Detección de Toxicidad, Procesamiento de Lenguaje Natural, Constructividad, Recursos Lingüísticos en Español.

ISSN 1135-5948 DOI 10.26342/2024-73-3

1 Introduction

Online content moderation is an increasingly complex and challenging task, especially on platforms that allow user participation through comments on news or posts (Nobata et al., 2016; Narang et al., 2022). The number of comments generated on these platforms can be overwhelming, and many of them may contain offensive, discriminatory, threatening, or toxic language that can be detrimental to the user experience and the image of the platform.

Toxicity in news comments refers to rude, disrespectful, unreasonable, or hateful language that is likely to make other users leave a discussion (Risch and Krestel, 2020). Toxic comments may contain insults, profanity, or attacks on the authors or people mentioned in an article. With the rise of online presence and social media, it is increasingly important to ensure that online spaces are safe and respectful for all users. Therefore, detecting toxicity in news comments is an increasingly relevant issue in the information age in which we live.

One interesting approach to address this problem is applying Natural Language Processing (NLP) technologies to develop computational systems that automatically detect toxicity in news comments. These systems use machine learning algorithms to analyze comments and detect language patterns that indicate toxic behavior (Taulé et al., 2024).

The importance of these systems lies in their ability to help human moderators identify and remove toxic comments more quickly and effectively from online platforms. Moderators can be exposed to large amounts of toxic content, which can have a negative impact on their emotional well-being and their ability to do their jobs effectively. By using automated systems to filter out toxic comments, moderators can focus on reviewing the most relevant and useful comments for discussion. Additionally, the quality of comments can be improved, and online discussions can be made more respectful and constructive.

In this paper, we present a corpus of NEws COmments written in Spanish annotated with the level of TOXicity (NECOS-TOX) and explore different NLP and machine learning-based classification algorithms for detecting toxicity in such comments. Furthermore, we conduct an analysis of the relationship between constructiveness and toxicity in Spanish news comments, aligning with prior research that has explored similar connections. This investigation allows for the identification of factors contributing to toxic behavior (Kolhatkar and Taboada, 2017; Nguyen, Van Nguyen, and Nguyen, 2021) and enables targeted interventions to prevent or mitigate its effects. In this paper we use an existing corpus NECOS annotated with constructiveness (López-Úbeda et al., 2021). However, in this study, we have adopted a different approach and annotated the same corpus with toxicity levels using a newly created version called NECOS-TOX, which is available for research purposes.

The following are the primary contributions of this paper:

- We have annotated the NECOS Spanish dataset with toxicity levels generating the NECOS-TOX corpus.
- We have established benchmark experiments for the NECOS-TOX dataset by exploring different NLP models for toxicity detection.
- We have examined the relationship between toxicity and constructiveness in news comments.

The paper is structured as follows[1]. Section 2 provides an overview of the existing research in the field of detecting toxic comments using NLP and machine learning techniques. Section 3 introduces the NECOS-TOX corpus and discusses how it has been annotated, including relevant statistics. Section 4 describes the experiments we conducted, including the use of advanced machine learning algorithms and pre-processing techniques. We report on the results of our experiments, including the performance of the models. In Section 5 we have analyzed and discussed the relationship between toxicity and constructiveness. Finally, in Section 6, we summarize our findings and discuss potential areas for future research.

2 Related Work

The toxicity detection in comments is usually treated as a text classification problem, mainly dealt with by machine learning methods (Zaheri, Leath, and Stroud,

[1]Warning: This paper discusses and contains content that may be deemed offensive or upsetting.

2020). The recent trend for text classification tasks uses language models (LMs) that are pre-trained on large unlabeled corpora (Xenos, Pavlopoulos, and Androutsopoulos, 2021; Chvasta et al., 2022; Plaza-del Arco et al., 2022; Bose, Perera, and Dorr, 2023).

Numerous studies have been conducted to detect toxicity, offensive language, and hate speech in text across various platforms and social networks (Kogilavani et al., 2021). However, there is a lack of research on detecting toxicity in news comments specifically. (Li, Mao, and Liu, 2019) developed three models to detect toxic comments with high accuracy using Kaggle's Civil Comments dataset. The best-performing individual model achieved an F1 score of 81.19% and two ensemble methods further improved the results to 84.28%. The team used Naïve Bayes-SVM as the base model and explored two deep learning models, LSTM and BERT, while addressing the issue of imbalanced data using a weighted loss function. (Kolhatkar et al., 2020b) presents the SFU Opinion and Comments Corpus (SOCC), a collection of opinion articles and the comments posted in response to the articles. This corpus is an interesting resource that is freely available and is labeled with constructiveness and toxicity, among other features[2]. Several works have been done on this line, including previous papers such as (Kolhatkar and Taboada, 2017), which studies the toxicity of comments and suggests that it is necessary to consider constructiveness along with toxicity when moderating news comments because some toxic comments may still be constructive. In (Garlapati, Malisetty, and Narayanan, 2022), the authors study the application of NLP techniques to classify various types of toxicity in online comments. The project aims to predict the toxicity class of each comment accurately using data from online platforms labeled as toxic or non-toxic. The research is divided into two phases: phase I involves evaluating the toxicity in comments using various techniques, while phase II involves analyzing the data to organize the comments into two categories (toxic and non-toxic). The experiments include the training of an LSTM model, which successfully achieved an accuracy rate of 94%. The literature review reveals that a limited number of studies have explored the use of pre-trained language models for the classification of toxic comments, and there has been a minimal investigation into languages other than English (Zhao, Zhang, and Hopfgartner, 2021).

Regarding works in languages other than English, not many can be found. For example, (Dinkov, Koychev, and Nakov, 2019) created a new dataset by mining a Bulgarian website to collect news articles for five years that were manually classified into eight toxicity groups. They then trained a multi-classifier with nine categories: eight toxic and one non-toxic. The experiments involved different representations based on ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and XLM (Lample and Conneau, 2019), as well as a variety of domain-specific features. (Nguyen, Van Nguyen, and Nguyen, 2021) created UIT-ViCTSD (Vietnamese constructiveness and Toxic Speech Detection dataset), a dataset for constructiveness and toxic speech detection. Through their analysis, authors uncovered insights into the relationship between these two phenomena, demonstrating that constructive comments can also exhibit toxicity. Focused on Spanish, some works have appeared (Plaza-del-Arco et al., 2021; Subies, 2021), mainly inspired by the DETOXIS (DEtection of TOxicity in comments In Spanish) workshop held in 2021 (Taulé et al., 2021). The aim of the DETOXIS task is the detection of toxicity in comments posted in Spanish in response to different online news articles related to immigration. The DETOXIS task is divided into two related classification subtasks: toxicity detection task and toxicity level detection task[3]. Our study differs from previous work in Spanish because i) we annotate toxicity in a corpus previously labeled with constructiveness, and ii) we have examined the relationship between these two phenomena (toxicity and constructiveness).

3 NECOS-TOX Corpus

In this section, we present NECOS-TOX, a corpus of online news comments enriched with toxicity annotations. We have used the NECOS corpus as a base, which was previously labeled for constructiveness (López-Úbeda et al., 2021). This will allow us to con-

[2] `https://github.com/sfu-discourse-lab/SOCC` (Last accessed July, 2024)

[3] `https://detoxisiberlef.wixsite.com/website` (Last accessed July, 2024)

duct a study to compare and analyse the relationship between constructiveness and toxicity in news comments.

3.1 Data Collection

The NECOS corpus contains 1,419 comments from 10 articles from the *El Mundo* newspaper[4]. The corpus contains approximately 150 random comments from each news article. The news articles selected were published from April 3rd to April 30th, 2018, and were specifically chosen for their potential to stimulate lively discussions and disagreements among readers.

3.2 Data annotation

For the annotation of the corpus, we have followed the annotation guide of a similar study (Kolhatkar et al., 2020b). The study uses 4 levels of toxicity which are defined as follows:

1. **Not toxic**: comments that do not use harmful, offensive, or abusive language.
2. **Mildly toxic**: comments that express frustration, anger, or that some people might consider toxic in some contexts.
3. **Toxic**: comments that are non-constructive and potentially harmful, including those that are sarcastic, involve ridiculing or teasing others, and are characterized by aggressive disagreement or inappropriate joking.
4. **Very toxic**: comments that use abusive and offensive language, including personal attacks, insults, and derogatory or demeaning remarks that cause embarrassment and disrespect. These types of comments would be removed by a moderator.

The corpus is annotated using the four labels listed above. The annotation is accomplished by 3 annotators, who are male undergraduate students of Computer Science between 20 and 24 years old. Each annotator receives a spreadsheet in which each row contains: news item ID, link to the news item, comment ID, comment text and an empty cell to type the label code. The spreadsheets were independent. In this way, one annotator could not see what was annotated by the others. The annotator is required to read the news item and then assign a label to each comment of that news item. The whole process is divided into 3 phases. The first and second phases were carried out to better define the annotation guide. First, the three annotators were involved and three subsets of 150 comments each were created. Each subset was labeled by two annotators. Each subset is randomly formed and stratified, that is, they contain comments on the 10 news items. The allocation of subsets to each annotator can be seen in Table 1.

Annotator/Subset	S1	S2	S3
Annotator A	X	X	
Annotator B	X		X
Annotator C		X	X

Table 1: Assignment of subsets to annotators. S1: subset one, S2: subset two, S3: subset three.

In the second phase, all comments whose annotations were not matched were reviewed by annotators. They included examples and new comments in the annotation guide to improve the description of each toxicity level.

Finally, in the third phase, the three annotators labeled the entire corpus according to the annotation guide and the findings of phase two. All spreadsheets were merged into a single spreadsheet to calculate the final label for each comment.

In reviewing the final spreadsheet, we noted that only four comments were assigned the label 4 (very toxic) by a single annotator. All other annotators labeled these comments with label 3 (toxic). For this reason, no comments have been labeled 4 when calculating the final toxicity value (see section 3.3). Therefore, this label has not been taken into account in the experimentation. In summary, Table 2 shows three different comments, each annotated with a level of toxicity.

3.3 Inter-annotator Agreement

Each comment in the corpus is annotated by the 3 annotators (A, B and C) with a value of 1 (no toxic), 2 (mildly toxic) or 3 (toxic). To assign the final toxicity label, the value closest to the mean of the annotations was calculated.

The final result of the annotation can be observed in Table 3, which shows the number of each type of label assigned to the com-

[4] `https://www.elmundo.es/` (Last accessed July, 2024)

Comment	Toxicity level
Eso es porque su vivero de votos está en las provincias vascas. Yo ya no los votaba pero no los volveré a votar jamás.	1
That's because their breeding ground for votes is in the Basque provinces. I didn't vote for them any more, but I will never vote for them again.	
Así que: lo del 155 solo era una pantalla para sacar más y más al cobarde Sr. Rajoy. ¡Dan ganas de llorar! y Sánchez estará feliz y orgulloso de su NO, aunque el país se desangre. Ninguno de los dos merece un solo voto.	2
So: 155 was just a smokescreen to bring out more and more of the coward Mr. Rajoy. It makes you want to cry! and Sánchez will be happy and proud of his NO, even if the country bleeds to death. Neither of them deserve a single vote.	
Estos no saben con quien se la están jugando... Alguno va a aparecer con una piedra en el cuello	3
They don't know who they are playing with.... Someone is going to show up with a rock around their neck.	

Table 2: Examples of comments labeled in the NECOS-TOX corpus, along with English translations.

ments for each news item. The table also shows the total percentage of comments labeled with each toxicity level. The proportion of toxic comments is about 59% (52.92% + 5.99%), while the proportion of non-toxic comments is 41%.

To measure the level of agreement between the three annotators, we determined the percentage of agreement between each pair of annotators and Cohen's kappa coefficient (Cohen, 1960). Table 4 shows the agreement between each pair of annotators and the percentage of coincidence. In this table, we can see that the average kappa is 43.55. This score represents moderate agreement. Similarly, the agreement between annotators A & B, and then between annotators A & C. However, the agreement value between B & C corresponds to a fair level of agreement. On the other side, the percentage of agreement with the annotations is between 63% and 70%, and the average, with an average of 67.46%. Both metrics achieve acceptable values for annotation and agreement in the NECOS-TOX corpus.

In addition to Cohen's Kappa coefficient, the Fleiss' Kappa metric has been calculated. This metric allows us to obtain an overall value of agreement when there are more than two annotators (Landis and Koch, 1977). The obtained Fleiss' Kappa value is 0.4343 (43.43%). This means that a moderate value of the inter-annotator agreement is obtained.

These agreement values indicate that the task is moderately subjective and that the value assigned is influenced by the annotator's point of view. However, this labelled corpus will allow the performance of different classification algorithms to be compared.

4 Benchmark Experiments

This section provides a summary of the different machine learning and transformer models used for detecting toxic comments.

4.1 Models

We have considered five machine learning models for this study: Support Vector Machine (SVM) (Burger, 1998), two transformer-based language models trained on Spanish texts (BETO (Cañete et al., 2020) and BERTIN (la Rosa et al., 2022)), and two cross-lingual language models (multilingual BERT (Pires, Schlinger, and Garrette, 2019) and XLM-RoBERTa (Conneau et al., 2019)).

SVM For the development of this algorithm, we use the *scikit-learn* library with the default SVM parameters together with the TF-IDF feature extractor (Salton and Buckley, 1988). In this study, this traditional algorithm is used as the baseline.

Spanish language models The Spanish pre-trained models used for this study are

News item	(1) No toxic	(2) Mildly toxic	(3) Toxic	# of comments
1	50	84	15	149
2	74	71	3	148
3	61	82	7	150
4	84	45	20	149
5	73	70	6	149
6	73	63	6	142
7	46	93	11	150
8	31	89	8	128
9	53	93	4	150
10	38	61	5	104
Total	583	751	85	1419
Percentage	41.08%	52.92%	5.99%	

Table 3: Analysis of toxicity annotation for each news item in the NECOS-TOX corpus.

Annotators	Cohen's kappa	Agreement
A & B	47.72	68.82%
A & C	47.67	70.43%
B & C	35.26	63.14%
Average	43.55	67.46%

Table 4: Inter-annotation agreement in the NECOS-TOX corpus.

BETO (*bert-base-spanish-wwm-cased*[5]) and BERTIN (*bertin-roberta-base-spanish*[6]). On the one hand, BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT base and was trained with the whole word masking technique.

On the other hand, BERTIN is a series of BERT-based models for Spanish. The current model hub points to the best of all RoBERTa-base models trained from scratch on the Spanish portion of mC4 using Flax.

Cross-lingual language models Since there is a shortage of language-specific models, we also tested some multilingual language models to observe how they performed. First, multilingual BERT (mBERT) (*bert-base-multilingual-cased*[7]) is a model trained on the top 104 languages with the largest Wikipedia using a Masked Language Modeling (MLM) objective.

Subsequently, we tested the XLM-RoBERTa cross-lingual model (*xlm-roberta-base*[8]). XLM-RoBERTa is a multilingual version of RoBERTa. This model is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

4.2 Hyperparameter optimization

During the experimental development phase, 10-fold cross-validation and grid search has been employed as hyperparameter optimization techniques. As for the hyperparameters used, we carried out a grid search to find out the combination that maximized the task's metric. For the epsilon parameter, we tested with $1e-5$, $1e-6$, $1e-7$, $1e-8$, $2e-5$, $2e-6$, $2e-7$, $2e-8$. Concerning the learning rate, the values we tested during the grid search were $1e-3$, $1e-4$, $1e-5$, $2e-3$, $2e-4$, $2e-5$. The batch size values tested during the optimization were 8, 16 and 32. Finally, the number of epochs tested ranged from 5 to 20. Table 5 summarizes the best-selected hyperparameters for fine-tuning each model along with the total number of training epochs.

4.3 Results

In this section, we discuss the results obtained by the different models mentioned in the previous section.

We evaluate the machine learning models using 10-cross validation. The metrics used for evaluation are those commonly used to measure the quality of text classification al-

[5] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased` (Last accessed July, 2024)

[6] `https://huggingface.co/bertin-project/bertin-roberta-base-spanish` (Last accessed July, 2024)

[7] `https://huggingface.co/bert-base-multilingual-cased` (Last accessed July, 2024)

[8] `https://huggingface.co/xlm-roberta-base` (Last accessed July, 2024)

Model	Optimizer	Epsilon	Learning rate	Batch size	Training epochs
BETO	AdamW	1e-8	1e-5	16	17
BERTIN	AdamW	1e-5	1e-3	32	9
mBERT	AdamW	2e-6	1e-3	32	14
XLM-RoBERTa	AdamW	1e-7	2e-5	16	15

Table 5: Hyperparameters selected for each model.

Model	Precision	Recall	F1 score	Accuracy
BETO	**0.616**	**0.571**	**0.561**	**0.678**
BERTIN	0.545	0.520	0.501	0.665
XLM-RoBERTa	0.578	0.539	0.526	0.661
mBERT	0.493	0.492	0.473	0.635
SVM	0.396	0.368	0.316	0.555

Table 6: Performance comparison of different machine learning algorithms on the NECOS-TOX corpus using 10-cross validation. Global metrics using the macro-avg variant. The highest results among those tested are highlighted in bold.

gorithms, i.e., precision, recall, F1-score, and accuracy.

The results of all the models tested are shown in Table 6. Regarding the baseline adopted for this study (SVM), the overall performance showed that it performed relatively poorly on the NECOS-TOX dataset, with a precision value of 0.396, a recall of 0.368, and an F1-score of 0.316 using the macro avg metric, although the accuracy reached 0.555.

The pre-trained corpus-specific language models for Spanish performed well on the dataset. Specifically, BETO achieved the best performance with an accuracy of 0.678 and BERTIN with an accuracy of 0.665.

On the other hand, the results obtained by the cross-linguistic models showed that they are at a similar level to the Spanish-specific models for toxicity detection in our NECOS-TOX corpus. XLM-RoBERTa obtained an accuracy of 0.635 and mBERT of 0.635. However, we can highlight that among the transformer-based models, mBERT was the worst of the transformer-based models in the precision, recall, and F1-score metrics since they obtained values below 0.5 in contrast to the others.

Since the dataset is composed of three toxicity variants, Figure 1 presents the precision, recall, and F1-score results for each of the toxicity levels used. As we can see, there is no difference in terms of the language models used. The figure also shows the low results in the detection of toxicity given the small number of examples in the corpus. Specifically, SVM has obtained values of 0 in this category and BETO outperforms the other models.

5 Toxicity and Constructiveness relationship

According to the definition of constructiveness, we expect constructive comments to present well-reasoned arguments and offer meaningful debate. On the other hand, toxic comments contain insults and derogatory or denigrating attacks against the authors or the persons mentioned. In this section, we analyze the NECOS-TOX corpus by studying how the two phenomena interact with each other. Figure 2 illustrates the distribution of comments labeled with constructive and non-constructive classes along with the toxicity levels. The most important result of this annotation experiment is that there was no significant difference in toxicity levels between constructive and non-constructive comments, i.e., constructive comments were equally likely to be toxic (in all three levels) or non-toxic as non-constructive comments.

According to the analysis conducted on the NECOS-TOX corpus, a non-constructive and non-toxic comment does not contribute to a meaningful debate, but neither does it offend or insult. Some examples of these

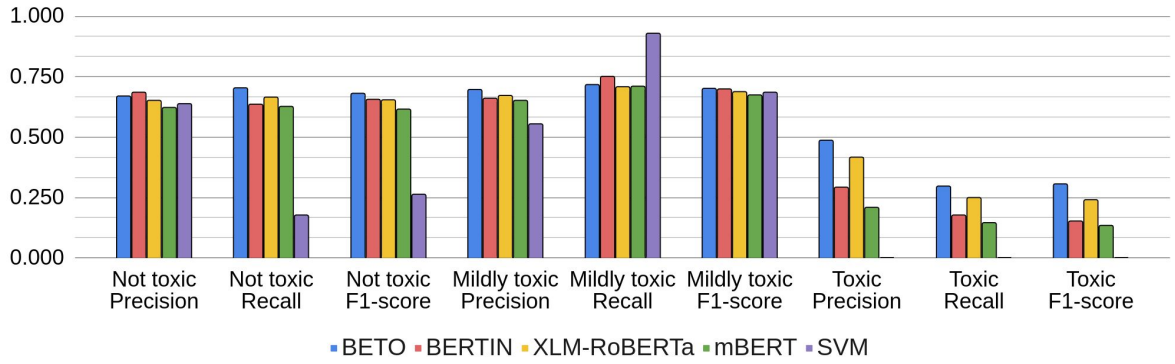

Figure 1: Performance results for each label according to toxicity level in NECOS-TOX dataset.

	Comment
1	*Es una vergüenza. Si nos están extorsionando, es una maldita vergüenza tener que caer en esto. Luchad por vuestro país, por el bien de todos y no para seguir en el poder con esta escoria desangrándonos. Vayamos a elecciones para que en un futuro cambie la constitución y desaparezcan todas las autonomías.*
	It is a disgrace. If we are being extorted, it's a damn shame to have to fall for this. Fight for your country, for the good of all and not to stay in power with this scum bleeding us dry. Let's go to elections so that in the future the constitution will change and all the autonomies will disappear.
2	*Una vez mas, el Jefe del Estado ha tenido que ir a respaldar a una institución. Rajoy, haciendo el ridículo, en ese autohomenaje que se dan los peperos, donde no dicen nada mas que tonterías. ¡vete a tu casa! y deja que este país pueda salir adelante sin vosotros. ¡cobardes!*
	Once again, the Head of State has had to go to support an institution. Rajoy, making a fool of himself, in that self-homage that the *peperos* give themselves, where they say nothing but nonsense. go home! and let this country move forward without you. cowards!
3	*Mira tonto... ¿Balay? ¿Avantia? Tenemos Gestamp, Cie Automotive, Ingeteam, Gamesa, CAF, Irizar,... Todas estas empresas vascas, fruto del trabajo y el espíritu emprendedor tienen sus principales mercados en el exterior. Si en vez de tanto chiringuito y tanta plantación de pepinos, se hubieran creado tantas empresas exportadoras por cada mil habitantes, otro gallo cantaría.*
	Look silly... Balay? Avantia? We have Gestamp, Cie Automotive, Ingeteam, Gamesa, CAF, Irizar,... All these Basque companies, the fruit of hard work and entrepreneurial spirit, have their main markets abroad. If instead of so many *chiringuito* and so many cucumber plantations, so many exporting companies had been created for every thousand inhabitants, another song would sing.

Table 7: Examples from the NECOS-TOX corpus annotated with toxicity (level 3) and constructiveness, along with English translations.

types of comments are usually defined by an agreement or disagreement expressed in a polite way and referenced in the article. Constructive but non-toxic comments are usually posts that contribute ideas and provide convincing arguments. Regardless of whether the author agrees or disagrees with the article, no offensive language is used at any time. On the contrary, a non-constructive and toxic comment is used by users who express malice without making a good argument. Finally, the most interesting category to explore is toxic but constructive comments. Table 7 provides some examples of comments anno-

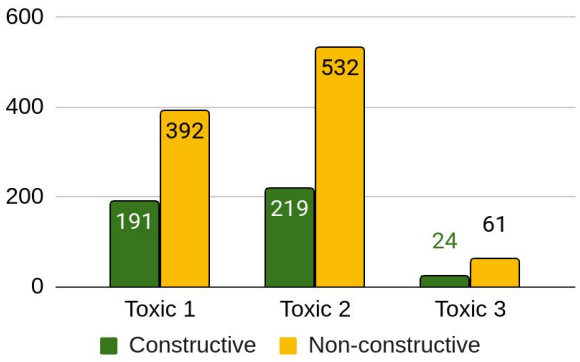

Figure 2: Correlation between toxicity and constructiveness in NECOS dataset.

tated with this category. As we can see, users tend to argue in a noxious way using harmful words such as "cowards", "damn" "shame", "scum", etc. Paradoxically, however, the authors indicate how the country's future could be improved by providing some reasoning.

The NECOS corpus was manually annotated with constructiveness with a high percentage of agreement among the annotators. Thus, considering the percentage of agreement in the annotation of constructiveness and toxicity, we found another important difference. While for constructiveness, the authors reached 91.03% agreement (López-Úbeda et al., 2021), for toxicity, it drops to 67.46%. We believe that the agreement was low because the three classes were treated as mutually exclusive and the toxic class only covered 5.99% of the corpus (85 comments). Moreover, the authors of the NECOS dataset provide results for the constructiveness classification, so we can compare these results in order to analyze whether the two phenomena are comparable in terms of performance. In both cases, the best result was achieved by the BETO language model, however, in the constructiveness classification the system reached 77.59% accuracy, 78.54% in recall, and 77.24% F1 score. In contrast, for the detection of toxicity levels, we obtained 61.6%, 57.1%, and 56.1% precision, recall, and f1 respectively.

In summary, as discussed in the related literature and in line with other works (Kolhatkar and Taboada, 2017; Kolhatkar et al., 2020a; Nguyen, Van Nguyen, and Nguyen, 2021), so far, no relationship has been found between toxicity and constructiveness since a comment may be toxic or bad sounding but may offer solutions, new perspectives and insights. In the NECOS-TOX dataset, as we verified through examples and annotated cases, the same situation exists; based on our analysis we consider there is not a clear relationship between the two phenomena.

6 *Conclusion*

This paper presents a corpus of Spanish news comments labeled with toxicity and investigates the relationship between toxicity and constructiveness in these comments. Through various experiments, we found that Spanish models, such as BETO, are effective in detecting toxicity in Spanish news comments. Our results indicate that there is no clear correlation between toxicity and constructiveness in these comments. This suggests that these two factors may not be closely related, as other studies in different languages have also found constructive comments that are toxic. Therefore, the level of toxicity does not necessarily indicate the level of constructiveness in comments.

Our study has important implications for future research. On the one hand, our findings suggest that the relationship between toxicity and constructiveness in online comments is complex and warrants further investigation to better understand how to foster constructive online dialogues in Spanish news comment sections. On the other hand, there is a need for new corpora that annotate both constructiveness and toxicity, not only in news comments but also in other online content, such as social media conversations on controversial topics. This would allow for a more comprehensive understanding of the relationship between these factors in online discourse. Another aspect to take into account is that the Spanish adaptation of the annotation guide was carried out by a team of two women and one man, but the corpus was only annotated by male annotators. The moderate agreement value obtained indicates that the annotators have internalized different sensitivities when interpreting the annotation guide. However, there is a possibility that some comments could be considered offensive by women or people of different ages. As future work, we consider expanding the number of annotators to reflect different sensitivities.

Acknowledgements

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, and grant number PTQ2021-012120 funded by Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033).

References

Bose, R., I. Perera, and B. Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14, Toronto, Canada, July. Association for Computational Linguistics.

Burger, C. 1998. A tutorial on support vector machines for pattern recognition, data mining and knowledge discovery. WORKSHOP ON DATA MINING AND KNOWLEDGE DISCOVERY.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020):1–10.

Chvasta, A., A. Lees, J. Sorensen, L. Vasserman, and N. Goyal. 2022. Lost in distillation: A case study in toxicity modeling. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 92–101, Seattle, Washington (Hybrid), July. Association for Computational Linguistics.

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

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dinkov, Y., I. Koychev, and P. Nakov. 2019. Detecting toxicity in news articles: Application to bulgarian. *arXiv preprint arXiv:1908.09785*.

Garlapati, A., N. Malisetty, and G. Narayanan. 2022. Classification of toxicity in comments using nlp and lstm. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 16–21. IEEE.

Kogilavani, S., S. Malliga, K. Jaiabinaya, M. Malini, and M. M. Kokila. 2021. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*.

Kolhatkar, V. and M. Taboada. 2017. Constructive language in news comments. In *Proceedings of the first workshop on abusive language online*, pages 11–17.

Kolhatkar, V., N. Thain, J. Sorensen, L. Dixon, and M. Taboada. 2020a. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Kolhatkar, V., H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada. 2020b. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155–190.

la Rosa, J. D., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Lample, G. and A. Conneau. 2019. Cross-lingual language model pretraining.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Li, H., W. Mao, and H. Liu. 2019. Toxic comment detection and classification. In *CS299 Machine Learning*. Standford University.

López-Úbeda, P., F. M. Plaza-del Arco, M. C. Díaz-Galiano, and M. T. Martín-Valdivia. 2021. Necos: An annotated corpus to identify constructive news comments in spanish. *Procesamiento del Lenguaje Natural*, 66:41–51.

Narang, K., A. M. Davani, L. Mathias, B. Vidgen, and Z. Talat. 2022. Proceedings of the sixth workshop on online abuse and harms (woah). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*.

Nguyen, L. T., K. Van Nguyen, and N. L.-T. Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 572–583. Springer.

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations.

Pires, T., E. Schlinger, and D. Garrette. 2019. How multilingual is multilingual bert?

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. U. López, and M. T. M. Valdivia. 2021. SINAI at iberlef-2021 DETOXIS task: Exploring features as tasks in a multi-task learning approach to detecting toxic comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 580–590. CEUR-WS.org.

Plaza-del Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Risch, J. and R. Krestel. 2020. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, pages 85–109.

Salton, G. and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Subies, G. G. 2021. Guillemgsubies at iberlef-2021 DETOXIS task: Detecting toxicity with spanish BERT. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 591–598. CEUR-WS.org.

Taulé, M., A. Ariza, M. Nofre, E. Amigó, and P. Rosso. 2021. Overview of detoxis at iberlef 2021: detection of toxicity in comments in spanish. *Procesamiento del lenguaje natural*, 67:209–221.

Taulé, M., M. Nofre, V. Bargiela, and X. Bonet. 2024. Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish. *Language Resources and Evaluation*, pages 1–41.

Xenos, A., J. Pavlopoulos, and I. Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online, August. Association for Computational Linguistics.

Zaheri, S., J. Leath, and D. Stroud. 2020. Toxic comment classification. *SMU Data Science Review*, 3(1):13.

Zhao, Z., Z. Zhang, and F. Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507.