OffendES: A New Corpus in Spanish for Offensive Language Research

Flor Miriam Plaza-del-Arco, Arturo Montejo-Ráez,
L. Alfonso Ureña-López and María-Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, amontejo, laurena, maite}@ujaen.es

Abstract

Offensive language detection and analysis has become a major area of research in Natural Language Processing. The freedom of participation in social media has exposed online users to posts designed to denigrate, insult or hurt them according to gender, race, religion, ideology, or other personal characteristics. Focusing on young influencers from the well-known social platforms of Twitter, Instagram, and YouTube, we have collected a corpus composed of 47,128 Spanish comments manually labeled on offensive pre-defined categories. A subset of the corpus attaches a degree of confidence to each label, so both multi-class classification and multi-output regression studies are possible. In this paper, we introduce the corpus, discuss its building process, novelties, and some preliminary experiments with it to serve as a baseline for the research community.

1 Introduction

Offensive language is defined as the text which uses hurtful, derogatory, or obscene terms made by one person to another person (Wiegand et al., 2019). Related terms in the literature are *hate speech* (Waseem and Hovy, 2016), *cyberbullying* (Rosa et al., 2019), *toxic language* (van Aken et al., 2018), *aggression language* (Kumar et al., 2018), or *abusive language* (Nobata et al., 2016). Although there are subtle differences in meaning, they are all compatible with the above general definition.

Due to the well-acknowledged rise in digital social interactions, in particular on social media platforms, the amount of offensive language is also steadily growing. Unfortunately, this type of prejudiced communication can be extremely harmful and could lead to negative psychological effects among online users, especially among young people, causing anxiety, harassment, and even suicide in extreme cases (Hinduja and Patchin, 2010).

At the same time, this issue also implicates governments, online communities, and social media platforms. In order to help fight this problem, these stakeholders are continuously taking appropriate actions to implement laws and policies combating hate speech. For instance, since 2013 the Council of Europe has sponsored the "No Hate Speech" movement[1] seeking to mobilize young people to combat hate speech and promote human rights online. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter, and YouTube to create a "Code of conduct on countering illegal hate speech online"[2]. From 2018 to 2020, platforms such as Instagram, Snapchat, and TikTok adopted the Code. According to a Spanish report in 2019 on the evolution of hate crimes in Spain[3], threats, insults, and discrimination are counted as the most repeated criminal acts, with the Internet (54.9%) and social media (17.2%) as the most widely used media to commit these actions.

To help achieve this goal, automatic systems based on Natural Language Processing (NLP) techniques are required. To train these systems, corpora labeled on offensive language are essential. In recent years, the NLP community has invested considerable effort into resource generation. However, most of them have been directed towards English, even though it is a global concern and there are important cultural differences depending on the language examined. In addition, most of them have been focused on Twitter data, despite the presence of offensive language on other platforms such as YouTube or Instagram, which more widely used by young people.

To contribute to filling this gap, in this paper[4]

[1] https://cutt.ly/sj5EdJ7
[2] https://cutt.ly/Hj5EsAh
[3] https://cutt.ly/ej5EgU7
[4] NOTE: This paper contains examples of potentially ex-

Proceedings of Recent Advances in Natural Language Processing, pages 1096–1108
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_123

we present OffendES, a Spanish collection of comments manually labeled for offensive content using a fine-grained annotation scheme. We collect our data from young influencers from well-known social platforms including Twitter, Instagram, and YouTube. Therefore, a comparative study of offensive behavior in social media and its relationship with the influencers is conducted. Finally, we propose preliminary experiments to serve as a baseline for the NLP community in which we show the validity of the corpus.

The remaining of the paper is organized as follows. Section 2 describes the related work on offensive language including some available datasets. Section 3 introduces our OffendES dataset and some descriptive statistics. Section 4 depicts our baseline evaluation of the novel dataset. A discussion is provided in Section 5. Finally, we conclude with our future studies in Section 6.

2 Related Work

2.1 Offensive Language Detection

In recent years, while offensive language continues to spread on the Internet, the importance of identifying this type of content in textual information has become increasingly significant in the NLP field, with several studies applying different machine learning systems. Most of these studies focus on the detection of offensiveness in social media, usually including a binary classification task to detect the presence of offensive language in the text.

Early studies explored traditional machine learning algorithms including Support Vector Machines, Logistic Regression, Random Forest, or Decision Trees, as well as the combination of different types of syntactic, lexical, semantic, and sentiment features (Chen et al., 2012; Nobata et al., 2016; Orăsan, 2018; Plaza-del-Arco et al., 2019).

As neural network architectures have shown promising results, extensive studies have recently explored a variety of deep learning architectures including Recurrent and Convolutional Neural Networks (Ranasinghe et al., 2019; Sharifirad and Matwin, 2019; Georgakopoulos et al., 2018). More recently, Transformer-based models have made significant progress and represent the state-of-the-art of multiple tasks, including offensive language detection (Plaza-del-Arco, Flor Miriam and Molina-González, M. Dolores and Ureña-López, L. Alfonso and Martín-Valdivia, María-Teresa, 2020; Casula et al., 2020; Wiedemann et al., 2020).

plicit or offensive content which may be offensive to some readers. They do not represent the views of the authors.

2.2 Data Available

Several labeled datasets are publicly available and usually include a binary annotation, indicating whether the content is offensive or not. Most of them have been generated in the context of different shared tasks for different languages.

For instance, the well-known offensive language task OffensEval has held two editions in the International Workshop on Semantic Evaluation (SemEval). In the first edition, Zampieri et al. (2019b) released the OLID dataset which contains over 14,000 English tweets. It was annotated using a three-level hierarchical annotation model by two people using a crowd-sourcing platform (Zampieri et al., 2019a). In order to retrieve tweets, they selected specific keywords and constructions often included in offensive posts related to Twitter accounts. Following the same annotation scheme, in the second edition Zampieri et al. (2020) introduced multilingual datasets comprising five different languages.

The Germeval shared task focused on offensive language identification in German tweets (Wiegand and Siegel, 2018). A dataset of over 8,500 annotated tweets was provided following also a hierarchical annotation. To collect the data, the authors explored the timeline of users that regularly post offensive content. Tweets were manually annotated by one of the three organizers of the task, and to measure inter-annotation agreement, 300 tweets were annotated by the three annotators in parallel. The annotation scheme is similar to the previously shared task, but differs in the following aspects: the number of levels in the hierarchy, the labels in the second level, and the language.

Related to Spanish, most of the datasets within the context of offensive language target hate speech, including AMI (Fersini et al., 2018), HatEval (Basile et al., 2019), and the HaterNet (Pereira-Kohatsu et al., 2019) collections. However, there is a lack of resources regarding the Spanish offensive language. To the best of our knowledge, the first corpus appeared at the 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval) (Carmona et al., 2018). This corpus was also used in the next edition of this workshop in 2019 (Aragón et al., 2019). The dataset focuses on the Mexican variant of Spanish

and contains around 10,475 tweets binary labeled as offensive or non-offensive. This collection has been recently revised (Díaz-Torres et al., 2020). EmoEvent (Plaza-del-Arco, Flor Miriam and Strapparava, Carlo and Ureña López, L. Alfonso and Martín-Valdivia, María-Teresa, 2020) is a multilingual emotion corpus based on different events, it also includes a small proportion of tweets labeled as offensive. Finally, the DETOXIS task[5] recently introduced the first dataset of comments in response to news articles labeled at different toxicity levels. To the best of our knowledge, there is no other Spanish corpus available with fine-grained categories for offensive language focused on young people. As the authors point out in (Aragón et al., 2019), the characterization of the offensiveness level found in a text is complex; therefore, there is a need for a more detailed classification of the tweets.

Our dataset, OffendES, differs from existing Spanish offensive language datasets because *(i)* apart from Twitter, we study the problem of offensive language detection on YouTube and Instagram, platforms that young people are more used to, *(ii)* we collect the data with a focus on young influencers, and *(iii)* we propose an annotation scheme with fine-grained classification.

3 OffendES Dataset

In this section, we describe the context of the dataset, the methodology followed to collect it and the annotation scheme proposed to label offensive content. Besides, we give some descriptive statistics and a detailed analysis of the collected data. OffendES is available upon request to the authors.

3.1 Scope of the Dataset

To understand the rationale behind the design and generation of the corpus, certain contextual information may be useful. As stated in the introduction, dealing with offensive posts in social networks is a growing concern. Several platforms are clear on this issue, as can be read in rules and policies of Twitter[6], Instagram,[7] or YouTube[8]. Indeed, YouTube has disabled comments on videos and channels featuring children (The YouTube Team, 2019). But this is a major concern not only for platform providers but for public administrations, in order to limit the possible side effects of harmful messaging to more vulnerable communities, like children or teenagers. With this in mind, the creation of this resource aims to achieve the following long-term goals:

1. Early detection of offensive language use in social media on the Internet, with a special focus on young people.

2. Identifying improvements in protection systems for young people in social networks.

3. Studying the feasibility of automatic learning systems for offensive language in Spanish.

4. Creating a reference corpus for the study of language technologies applied to the classification of sexist language.

3.2 Data Collection

Instagram, YouTube, and Twitter are among the social media platforms most used by people ages from 18 to 24 (Jenn Chen, 2020). These three have been selected as the main data sources. A total of 12 controversial influencers with a significant number of followers have been identified and their respective accounts in the three targeted social media platforms have been tracked. Table 2 (Appendix) shows the accounts used by the selected influencers in the three selected media. They are Spanish influencers from 24 to 35 years old and, six are men and six are women. The process for collecting comments consisted of two main steps. To collect the data, first, the last 50 posts by each influencer were obtained using the platform API. Then, an *ad hoc* web scraper was launched to extract user comments to each of the posts obtained (limited to 2,000 replies). This script uses scrolling through JavaScript code commands to retrieve further comments. In the case of YouTube, instead of the scraper, its API[9] has been used to retrieve comments.

During two months (from February to March 2020), a total number of 283,622 comments were collected (see Table 1 for detailed information). The comments were then filtered according to two main constraints: the presence of potentially offensive language and lexical diversity.

[5] https://cutt.ly/RkrVTQn
[6] https://cutt.ly/1j5Eut0
[7] https://cutt.ly/yj5Eijc
[8] https://cutt.ly/kj5Eo2d
[9] https://cutt.ly/JkrVSYv

Social network	Offensive terms	Non-offensive terms	Total
YouTube	19,449	184,414	203,863
Instagram	3,142	58,209	61,351
Twitter	1,197	18,728	19,925
Total	23,788	259,865	283,622

Table 1: Presence of offensive terms from lexicons in the retrieve comments.

To avoid the creation of a corpus with few or no offensive comments set, we labeled all the comments with flags determining whether the comment contained any of the words found in five different controlled lexicons (Plaza-del-Arco, Flor-Miriam and Molina-González, M Dolores and Ureña-López, L Alfonso and Martín-Valdivia, M. Teresa, 2020). All comments with potentially offensive language were selected (23,788 comments). We selected 60,000 comments to be labeled in the manual annotation phase. Therefore, we selected 36,212 comments without offensive terms. Applying lexical diversity measures proved to be an interesting approach to ensure a diverse set of comments. Therefore, we first attempted to include those comments that added the highest lexical diversity value to the growing set of collected comments. To that end, we applied the Measure of Lexical Textual Diversity MTLD (McCarthy and Jarvis, 2010), but the expected time to build the corpus with our implementation was unacceptable. Thus, we simply added those comments that produced the highest increase in the vocabulary size to the collection by iterating through all the comments and checking the amount of increase in vocabulary size comment by comment. At each iteration, that comment with the highest contribution of new vocabulary to the final collection was selected. This process was repeated until 60,000 comments were reached.

3.3 Labeling Process

In order to establish the annotation schema, we followed those defined in (Wiegand and Siegel, 2018; Zampieri et al., 2019a), while introducing some additional details that we consider important. Namely, we created a new category to include those posts with inappropriate language but no offense intended. For instance, the comment "eres la puta ama" (*you're the fucking boss*) contains inappropriate but non-offensive language and has a positive polarity. Then, we reformulated the definition of offensiveness to not include such posts.

The previous analysis led us to propose a definition of an offensive comment: one where language is used to commit an explicit or implicitly directed offense that may include insults, threats, profanity or swearing. Based on this definition, we established the following categories:

- **Offensive, the target is a person (OFP)**. Offensive text targeting a specific individual.
- **Offensive, the target is a group of people or collective (OFG)**. Offensive text targeting a group of people belonging to the same ethnic group, gender or sexual orientation, political ideology, religious belief, or other common characteristics.
- **Offensive, the target is different from a person or a group (OFO)**. Offensive text where the target does not belong to any of the previous categories, e.g., an organization, an event, a place, an issue.
- **Non-offensive, but with expletive language (NOE)**. A text that contains rude words, blasphemes, or swearwords but without the aim of offending, and usually with a positive connotation.
- **Non-offensive (NO)**. Text that is neither offensive nor contains expletive language.

The annotation of the collected data was performed via Amazon Mechanical Turk (MTurk)[10], which is a popular crowdsourcing platform. It provides the option of specifying some requirements that human annotators must meet to work on the task, and the time allotted per assignment. In our case, we selected the location as Spain and the time to five minutes due to the presence of some long comments from YouTube. Apart from releasing the annotation scheme with four examples of instances

[10] https://www.mturk.com/

for each class, in the purpose of ensuring clear and concise documentation, we also provided a list of instructions about rules, tips, and FAQs to try to solve any potential problems that could arise during the labeling process. Finally, to ensure the quality of the annotations, we used tracking comments.

We first conducted a round of trial annotation for both types of labeling, 4,500 and 1,500 instances with three and ten annotators, respectively. The goal of the trial annotation was *(i)* to identify any confusion in understanding the annotation schema, *(ii)* to estimate the average time to label the dataset, and *(iii)* to learn about the platform. The launch of these datasets was on September 24th, 2020, and it took two weeks to complete the annotation process on both sets. After analyzing the annotations, we observed through the comments of the annotators that the NOE and OFO classes were the most difficult to identify in the comments by the annotators. For this reason, we improved the definition of each class, providing examples as clear as possible to the annotators. The average agreement (kappa coefficient) grew from 36.85% for trial annotations up to 39.37% for final released comments. Yet, this level of agreement is lower than expected, which reflects the difficulty to discriminate among proposed classes.

Once the trial round was completed, the next step was to release the final dataset. A total of 54,023 instances were released in two subsets: 40,513 labeled by three annotators, and 13,510 labeled by ten annotators. The annotation took place from 17 November 2020 to 2 January 2021. As result, the three annotators subset covered 44,951 comments and the ten annotators subset 14,989 comments.

3.4 Post-processing

In order to check the reliability of the annotators, we analyzed their annotations in the tracking comments, i.e. those comments given as examples in the annotation guide. We observed that one of the annotators had over 60% of error rate in the tracking comments of both types of labeling, so we decided to remove their annotations since they could negatively affect the quality of the dataset. Sadly, this annotator was one of the most prolific, so the removal of his/her annotations resulted in a reduction of the three annotators subset to a number of 44,951 comments. A sample of the collected data is given in Tables 3 and 4 (Appendix).

3.5 Corpus Analysis

Thus, the final dataset is released divided into two subsets: the three annotators subset (3-Ann), with 44,951 comments, and the ten annotators subset (10-Ann), with 14,989 comments. The former is intended for multi-class classification research and the latter for tackling multi-output regression problems. Only 38 comments belong to both subsets. Comments are compiled without processing, therefore, case, punctuation, and emojis are preserved. Every comment is associated with a social network platform (Instagram, Twitter, or YouTube) and directed to one of the 12 selected influencers as the target. In Table 2, the amount of comments associated with each platform and influencer is depicted. Comments on *dalas'* posts are more frequent (over 26% in both subsets). YouTube is the platform where most of the comments were collected (about 75% for both subsets), followed by Instagram (over 18%). Comments from Twitter only represent just over 6% of the collection.

For both subsets, the label is the majority class according to human annotators. For the subset labeled by ten annotators, the majority vote was set to five annotators. An additional *None* label was used when no agreement was reached between annotators. Table 3 shows the number of comments for each label on both subsets. Noticeably, the 10-Ann subset has a much lower percentage of *None* labels than the 3-Ann subset. The more annotators that were involved, the easier it was to decide the final label for a comment.

Table 4 shows statistics on comments length (i.e. the number of characters in the text). As expected, YouTube is the platform with the highest average length (about 190 for both subsets), with high variance; Twitter comments average length is lower (149 characters), with very small variance, and Instagram is the platform where comments tend to be the shortest (with an averaged length of 114).

Figure 1 shows the distribution of comments among influencers and social media platforms in the 3-Ann subset. YouTube is the most frequent platform, followed by Instagram. The influencer *dalas* is the target of more than a quarter of the total amount of comments. A similar distribution of comments is found in the 10-Ann subset.

An interesting analysis is to measure label frequency according to each influencer. Figure 2 shows the proportion of influencer-level labels and reflects the differences among these users as tar-

	3-Ann Subset				10-Ann Subset			
Influencer	*Instagram*	*Twitter*	*YouTube*	*Total*	*Instagram*	*Twitter*	*YouTube*	*Total*
dalas	3,558	1,454	6,813	*11,825 (26.3%)*	1,223	494	2,214	*3,931 (26.2%)*
soyunapringada	582	31	5,412	*6,025 (13.4%)*	172	7	1,745	*1,924 (12.8%)*
windygirk	466	487	3,756	*4,709 (10.5%)*	183	186	1,249	*1,618 (10.8%)*
javioliveira	276	130	3,890	*4,296 (9.6%)*	92	52	1,297	*1,441 (9.6%)*
wismichu	859	327	2,929	*4,115 (9.2%)*	318	101	1,014	*1,433 (9.6%)*
miare	508	167	2,749	*3,424 (7.6%)*	166	63	936	*1,165 (7.8%)*
wildhater	648	0	2,485	*3,133 (7.0%)*	204	0	843	*1,047 (7.0%)*
nauterplay	540	0	2,058	*2,598 (5.8%)*	180	0	685	*865 (5.8%)*
lauraescane	286	152	1,991	*2,429 (5.4%)*	107	50	633	*790 (5.3%)*
dulceida	226	0	1,400	*1,626 (3.6%)*	81	0	440	*521 (3.5%)*
jpelirrojo	69	0	582	*651 (1.4%)*	23	0	187	*210 (1.4%)*
nosoymia	107	13	0	*120 (0.3%)*	42	2	0	*44 (0.3%)*
Total	*8,125*	*2,761*	*34,065*	*44,951*	2,791	955	11,243	*14,989*
	(18.6%)	*(6.4%)*	*(75.0%)*	*(100.0%)*	*(18.1%)*	*(6.1%)*	*(75.8%)*	*(100.0%)*

Table 2: Comments per social media and influencer in the OffendES dataset.

Label	**3-Ann**	**10-Ann**
NO	26,425	9,715
OFP	4,102	2,362
NOE	2,470	1,414
None	11,529	1,283
OFG	425	215

Table 3: Comments per label in the OffendES dataset.

(3-Ann subset)	**Average**	**Std. dev.**	**Min.**	**Max.**
YouTube	189	247	3	9,986
Twitter	149	75	4	413
Instagram	114	124	3	2,200
(10-Ann subset)	**Average**	**Std. dev.**	**Min.**	**Max.**
YouTube	191	277	4	9,812
Twitter	150	74	5	292
Instagram	113	115	3	1,631

Table 4: Statistics over comments length.

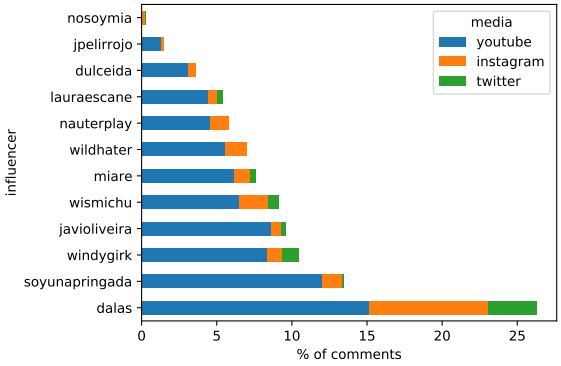

Figure 1: Comments distribution by influencer and social media platform in the 3-Ann subset.

get of offensive comments. In terms of gender, it can be seen that female influencers are subject to a greater number of offensive comments than male accounts. In particular, *soyunapringada*, *miare_love*, and *WindyGirk* are the accounts ranked with the most offensive comments. Regarding male influencers, accounts like *JaviOliveira* and *NauterPlay* contain more offense comments than accounts like *WildHater* and *JPelirrojo*. The profile of the influencer may define more controversy compared to others, or raise more negative emotions to their followers. Therefore, it could be interesting to consider the target profile as a source of information in offensive detection systems.

Inter-annotator agreement using the three annotators subset was measured with Cohen's kappa coefficient. The k value is 0.3579 (fair agreement), which is quite low and reflects how difficult it is for humans to agree between the proposed categories. By analyzing annotations on tracking comments, we found that it was a common mistake to label a comment NOE or OFG when it should have been labeled OFO. Figure 3 shows the percentage of consensus per label in the subset of 3-Ann taking as consensus the majority vote (2-annotators agreement and 3-annotators agreement). As can be noticed, the label OFO exhibits the lowest consensus rate, with all three annotators only agreeing on 33.72% of the time. We found that many OFO comments were wrongly annotated with the NOE label and, actually, this could be reasonable since these offenses are not directly targeted to persons or groups, and they often consist in expletive ex-

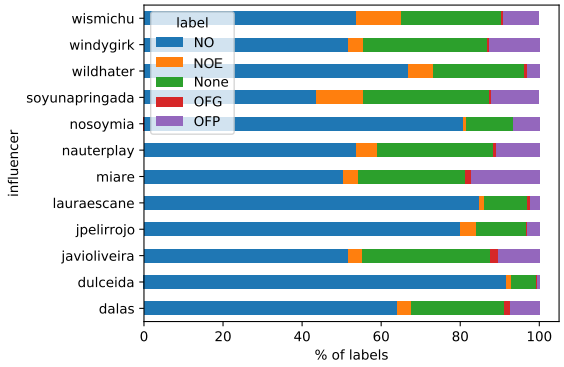

Figure 2: Distribution of labels per influencer in the OffendES dataset.

pressions. Thus, we decided to merge them. After merging the OFO label into the NOE label, the kappa value increases slightly up to 0.3837. Figure 4 shows the final percentage of consensus per label after the merge of NOE and OFO labels.

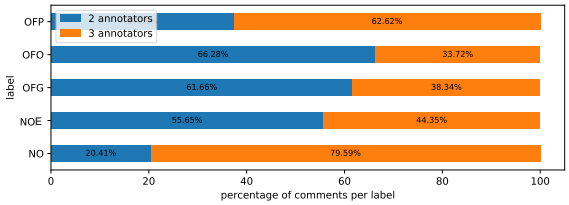

Figure 3: Percentage of consensus per label.

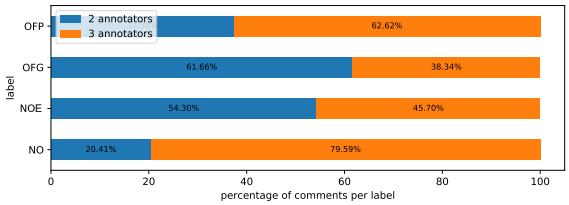

Figure 4: Percentage of consensus per label after including OFO label into NOE.

Another feature we analyzed is the lexical diversity of comments. To this end, we use the MTLD metric already introduced, which allows us to get an insight into lexical variation and avoiding biases due to different text lengths. Table 5 shows the average values for MTLD for comments over labels and platforms, respectively.

As can be noticed, offensive comments targeted to a person (OFP) have low lexical diversity, as well as for those with expletive language (NOE). When the comment is not offensive at all, the lexical diversity is clearly higher. Regarding social networks, we would expect the lowest value of diversity in Twitter, as it limits comment length. On the contrary, Twitter is the platform with the highest lexical diversity, followed by YouTube. Instagram is clearly much poorer in terms of the diversity of vocabulary used. These findings are worth exploring, as they could provide more understanding of how language is used across platforms and how it relates to harmful language use, or on the average profile of their communities. To understand MTLD values, we have to consider that a value of 50 is the average lexical diversity of texts for an average adult text (being 80 for academic writings).

		MTLD
Social network	Instagram	42.14
	Twitter	61.74
	YouTube	60.59
Label	NO	66.36
	NOE	26.41
	None	53.59
	OFG	53.19
	OFP	28.68

Table 5: Average values of measures of lexical textual comments diversity per social network and label.

4 Baseline System

In order to establish a baseline for the OffendES corpus, we conducted experiments based on three different approaches:

Simple majority class model. Our simplest classifier assigns the majority class of the training set, i.e., the NO class, to each instance in the test set. This results in accuracy values of 58.78% and 64.85% respectively for 3-Ann and 10-Ann subsets.

Lexicon-based model. We also developed a lexicon-based approach using the lexical resources described in Section 3.2. In this approach, we only consider a binary classification scenario: whether the comment is offensive or not. For the 3-Ann subset, we obtained 67.13% of accuracy, 21.27% precision, 83.78% recall, and 33.93% F1. For the 10-Ann subset, the values of accuracy, precision, recall and F1 were, respectively, 71.45%, 35.59%, and 81.60%, 49.56%.

Transformer-based model. Finally, we experimented with a Spanish pre-trained BERT model called BETO (Canete et al., 2020) which has shown

promising results in offensive language detection tasks (Plaza-del-Arco et al., 2020). Details about different configurations of the BETO model and the training process are given in the Appendix. In order to evaluate the model, we sampled from the collection two different sets, for training and evaluation. Measures used to report performance are Precision (P), Recall (R), and F1-score (F1) at class level, and macro and weighted average of these metrics. For the multi-output regression task, since we are not dealing with a multi-class scenario, we used one of the most preferred metrics for regression tasks, the mean squared error (MSE), a risk metric corresponding to the expected value of the squared (quadratic) error or loss.

4.1 Multi-class classification

This experiment is performed on the 3-Ann subset. All entries labeled as *None* were discarded (as no final label was assigned to these comments). The set was split into training (95%) and evaluation (5%) partitions, resulting in 30,079 comments in the training set and 3,343 in the evaluation set. Transformers (Wolf et al., 2020) library by Huggingface[11] was used to build the BERT network and the tokenizer from available BETO models (uncased variant).

A sequence classifier was implemented for this multi-class task, with a final linear layer with four outputs (the logits for each possible label). Training the model took 2 hours and 26 minutes.

After seven training epochs, the model was evaluated against the evaluation partition. The results obtained are depicted in Table 6.

Class	P (%)	R (%)	F1 (%)
NO	95.24	87.88	91.42
NOE	57.86	79.31	66.91
OFP	57.48	68.87	62.66
OFG	30.00	52.17	38.10
macro	60.15	72.06	64.77
weighted	86.96	84.39	85.33

Table 6: Multiclass experiment results.

4.2 Binary classification with BETO

Same configuration as the previous model, but using non-weighted cross-entropy as loss function during training. Classes have been merged into two classes as follows: *Non-offensive*, which comprises labels NO and NOE, and *Offensive*, combining OFP and OFG labels. This results in 28,895 non-offensive comments and 4,527 offensive comments. Training the model took 2 hours and 16 minutes. The results obtained are depicted in Table 7.

Class	P (%)	R (%)	F1 (%)
Non-offensive	92.79	95.14	93.95
Offensive	68.06	58.33	62.82
macro	80.42	76.74	78.39
weighted	89.06	89.59	89.26

Table 7: Binary classification experiment results.

4.3 Multi-output regression with BETO

For every sample, a vector of probabilities is computed by counting the number of annotators that selected each label and dividing by the number of annotators. This provides an estimate of the confidence of each label to be assigned to the comment. Training the model took 48 minutes.

The 10-Ann dataset was split into training and validation partitions. After training for seven epochs over a partition of 13,020 samples, the model was evaluated against a partition of 685 test samples, obtaining an MSE of 0.0241.

5 Discussion

One of the main characteristics of the corpus is its imbalance at all levels: comments are not uniformly distributed across labels, influencers, or social platforms. The corpus size allows for stratified random sampling over those dimensions, but we considered that releasing the full set of comments is the best choice to allow researchers to decide on how to prepare their experiments. That is also the reason why comments with *None* class have been kept in the corpus, so different studies on the use of language within groups of young users of social networks can be conducted. Also, the *None* label is of interest by itself, as it reflects the absence of consensus in determining the nature of the comment.

Results show that deep learning models, like BERT, are good estimators of the presence of different kinds of offensive language, but that it is still a challenging task to decide whether a comment is directed to a person or not (so cyber-bullying risk could be measured). Despite the fusion of NOE

[11] https://huggingface.co

and OFO categories, precision values for all labels different from NO are low.

6 Conclusion and Future Work

In this paper, we described OffendES: the first large-scale Spanish dataset of user comments on influencer posts from Instagram, YouTube and Twitter. It consists of 47,128 comments manually labeled for offensive content using a fine-grained annotation scheme. A subset of the corpus (10-Ann) assigns a confidence degree allowing both multi-class classification and multi-output regression studies. Additionally, a preliminary analysis of offensive behavior in social media and its relationship with the selected influencers is presented. Finally, baselines experiments have been performed, showing the validity of the corpus as well as the difficulty of the task.

A number of challenges remain open. On the one hand, we plan to explore systems trained on OffendES to monitor offensive messages in online channels participated by young people. On the other hand, the gender of the commenters and the subject of the comments have been left out for deeper analysis, so further research could be shed light on these matters. Finally, we believe that this dataset enables future work in the NLP community to tackle these interesting issues regarding Spanish language.

Acknowledgments

This work has been partially supported by a grant from European Regional Development Fund (FEDER), the LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Mario Ezra Aragón, Miguel Ángel Álvarez Carmona, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, and Daniela Moctezuma. 2019. Overview of MEX-A3T at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 478–494. CEUR-WS.org.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Fbk-dh at semeval-2020 task 12: Using multi-channel bert for multilingual offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1539–1545.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes-y Gómez, Juan Aguilera, and Luis Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136, Marseille, France. European Language Resources Association (ELRA).

E. Fersini, P. Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2018.

Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–6.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Jenn Chen. 2020. 2020 Social media demographics for marketers. https://sproutsocial.com/insights/new-social-media-demographics/. Accessed: 2020-09-22.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Constantin Orăsan. 2018. Aggressive language identification using word embeddings and sentiment features. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 113–119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2019. SINAI at SemEval-2019 task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 735–738, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2020. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Plaza-del-Arco, Flor-Miriam and Molina-González, M Dolores and Ureña-López, L Alfonso and Martín-Valdivia, M. Teresa. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Plaza-del-Arco, Flor Miriam and Molina-González, M. Dolores and Ureña-López, L. Alfonso and Martín-Valdivia, María-Teresa. 2020. SINAI at SemEval-2020 task 12: Offensive language identification exploring transfer learning models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1622–1627, Barcelona (online). International Committee for Computational Linguistics.

Plaza-del-Arco, Flor Miriam and Strapparava, Carlo and Ureña López, L. Alfonso and Martín-Valdivia, María-Teresa. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (Working Notes)*, pages 199–207.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, João Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Sima Sharifirad and Stan Matwin. 2019. Using attention-based bidirectional lstm to identify different categories of offensive language directed toward female celebrities. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 46–48.

The YouTube Team. 2019. More updates on our actions related to the safety of minors on YouTube. http://web.archive.org/web/20080207010024. Accessed: 2020-01-10.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. Uhh-lt at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Michael Wiegand and Melanie Siegel. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of KONVENS 2018*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Appendix

A.1 Model settings

Hyper-parameters. In the experiments with Transformer the hyper-parameters used for fine-tuning BETO are specified in Table 1. In the multioutput regression task the hyper-parameters are the same, except for the loss function, which is replaced by mean squared error loss, as it is a regression problem.

All experiments (training and evaluation) were performed on a node equipped with two Intel Xeon Silver 4208 CPU at 2.10GHz, 192GB RAM, as main processors, and six GPUs NVIDIA GeForce RTX 2080Ti (with 11GB each).

Hyper-parameter	Value
Batch size	32
Epochs	7
Learning rate (LR)	2e-5
LR linear decrease	Yes
Loss	Weighted cross-entropy
Optimizer	AdamW
Weight-decay	Yes

Table 1: BETO fine-tuning hyper-parameters.

A.2 OffendES dataset

Table 2 shows the accounts used by the selected influencers in the three selected media: Instagram, Twitter, and Youtube.

Table 3 shows examples of labeled comments in the OffendES dataset by social network.

Instagram	Twitter	Youtube
dalasito	DalasReview	Dalasreview
Wismichu	Wismichu	wismichu
jpelirrojo	JPelirrojo	jpelirrojo
nosoymia	s0ymia	Chrismellow
dulceida	dulceida	aidadomenech
lauraescanes	LauraEscanes	eshcanesh
miare_love	MIAREsproject	AchlysProject
javioliveira	javioliveira_	JaviOliveira
nauterplayyt	nauterplay	Nauter100
wildhater	WildHater	WildHater
windygirk	WindyGirk	WindyGirkTV
soyunapringada	soyunapringada_	Soy una pringada

Table 2: Different account identifiers for selected influencers.

	Comment	Social Network	Label
1	UNA MIERDA IGUAL QUE TU CANAL. *SHITTY JUST LIKE YOUR CHANNEL.*	Instagram	OFO
2	El que llora siempre en sus videos por haber sido acosado para dar pena ahora acosa a gente... patético. *The one who always cries in his videos for having been harassed to get pity now harasses people... pathetic.*	Twitter	OFP
3	El feminismo es cáncer y las feministas son mierda. *Feminism is cancer and feminists are shit.*	Youtube	OFG
4	Yo estoy de puta madre en casa... yo nací en cuarentena. *I'm doing fucking great at home... I was born in quarantine.*	Youtube	NOE
5	Si pudiera viajar. Bueno iria a italia. Que tengas un buen dia saludos desde Buenos Aires, Argentina. *If I could travel. Well I would go to Italy. Have a nice day. Greetings from Buenos Aires, Argentina.*	Instagram	NO

Table 3: Examples of comments labeled in OffendES (3-annotators subset), along with English translations.

	Comment	Social Network	OFP	OFG	OFO	NOE	NO
1	Vieja ridícula. *Ridiculous old woman.*	Instagram	1	0	0	0	0
2	Vaya tontería. Es campaña electoral, evidentemente unos le tiran mierda a los otros. *What nonsense. It's an election campaign, of course some of them throw shit at the others.*	Twitter	0	0	0	0.7	0.3
3	Eres un cómico increíble siempre consigues sacarme una sonrisa y se me olvidan las penas. *You are an amazing comedian, you always make me smile and forget my problems.*	Instagram	0	0	0	0	1
4	Mocosos "retrasados", ¿a alguien le ha sorprendido?, creo que no... *Snotty "retards", was anyone surprised? I don't think so...*	Youtube	0.1	0.7	0	0	0.2
5	Vaya mierda de vídeo. Deja de hablar sin saber, gracias. *What a shitty video. Stop talking out of your ass, thanks.*	Youtube	0.3	0	0.5	0	0.1

Table 4: Examples of comments labeled in OffendES (10-annotators subset), along with English translations.