Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1307–1316
Marseille, 20-25 June 2022

SHARE: A Lexicon of Harmful Expressions by Spanish Speakers

Flor Miriam Plaza-del-Arco[1], Ana Belén Parras Portillo[1], Pilar López-Úbeda[2], Beatriz Botella Gil[3], María Teresa Martín-Valdivia[1]

[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Spain
{fmplaza, abparras, maite}@ujaen.es
[2]R+D+I department. HT medica. Carmelo Torres nº2, 23007, Jaén, Spain
p.lopez@htmedica.com
[3]Department of Software and Computing System, University of Alicante, Alicante, Spain
beatrizbotellagil@gmail.com

Abstract

In this paper we present SHARE, a new lexical resource with 10,125 offensive terms and expressions collected from Spanish speakers. We retrieve this vocabulary using an existing chatbot developed to engage a conversation with users and collect insults via Telegram, named Fiero. This vocabulary has been manually labeled by five annotators obtaining a kappa coefficient agreement of 78.8%. In addition, we leverage the lexicon to release the first corpus in Spanish for offensive span identification research named OffendES_spans. Finally, we show the utility of our resource as an interpretability tool to explain why a comment may be considered offensive.

Keywords: offensive lexicon, offensive language detection, offensive span identification, interpretability

1. Introduction

Natural language processing (NLP) is one of the most promising areas for social media data processing. This growth is mainly due to the widespread use of social networks, which become the main channel for people to communicate, work and enjoy entertainment content. Moreover, Spanish is positioned as the second most important language worldwide for communication on the Internet.

However, these platforms are also generating a steady increase in court cases for offensive comments. Sharing aggressive and offensive content impacts negatively on society to a large extent due to its devastating effects. Depending on age, religion, and other demographic characteristics, offensive language can affect a user's psychological state or incite bullying on social networks. Therefore, the development of NLP-based automated tools and resources in languages such as Spanish would allow for the generation of alerts on aggressive and offensive comments posted (Chen et al., 2012).

The number of insults used in offensive comments can be unlimited depending on the imagination of the speakers, fashions, the influence of other languages, or the geographical context. Thus, although the Royal Spanish Academy (RAE: Real Academia Española) includes in its current dictionary a large number of insults such as *merluzo* (hake) or *ceporro* (dimwit), the richness of the language allows the creation of new words through composition. Spanish emerges as a great inventor of insults due to the continuous evolution of the language and the emergence of new grammatical forms of verbal violence that are not included in the RAE (Celdrán, 2009). For instance, the predilection for creating insults based on the word *cara* (face) is an example of this, which reaches current words such as *caranchoa* (anchovy face), passing through more subtle uses like *carajaula* (cage face). Furthermore, these insults can be formed by consecutive words, such as *chupa cabras* (suck goats) and *feo de mierda* (ugly shit).

The nature of some languages such as Spanish makes large-scale offensive lexicon development a difficult challenge. Since manual development is very costly and time-consuming, automatic and collaborative construction of computational lexical resources are real alternatives (Gala and Lafourcade, 2010). Moreover, lexical resources, such as lexicons are considered necessary to improve the performance of Named Entity Recognition (NER) and interpretability task (Etzioni et al., 2005; Toral and Muñoz, 2006; Lin et al., 2020). In making pre-trained models transparent and interpretable, it is often necessary to identify features that contribute significantly to a prediction.

In this study, we use the Fiero chatbot to collect potentially offensive words and expressions provided by the Spanish speakers to generate a lexical resource with offensive terms (Botella-Gil et al., 2021). The main contributions of this work are as follows: *i*) we generate SHARE (Spanish HARmful Expressions), a lexical resource composed of insults and offensive expressions manually labeled by 5 annotators (Section 3); *ii*) we use this resource to automatically annotate offensive entities in an available corpus for offensive language detection (Section 4); and *iii*) we explore the usefulness of the lexicon as an interpretability tool for offensive comments by comparing it with a BERT-based

fine-tuning model (Section 5). Both SHARE[1] and OffendES_spans[2] are publicly available.

The remainder of the paper is structured as follows: Section 2 presents some previous studies related to offensive language research. The process of collecting offensive words and expressions and the procedure for annotation is described in Section 3. Section 4 shows the annotation of the OffendES corpus using SHARE. An interpretation tool is employed in the BERT classification model and compared with the resource in Section 5. Finally, conclusions and future work are presented in Section 6.

2. Related Work

In this section, we review the different offensive lexical resources in different languages, focused mainly on Spanish, as well as the main tasks addressed in the offensive language research field.

Offensive Lexical Resources. Lexicons play an important role in offensive language research to analyze and identify problematic content on the Web. In the last years, great efforts have been conducted by the NLP community to generate lexical resources annotated with offensiveness. Most of them have been developed for English (Wiegand et al., 2018; Bassignana et al., 2018; Qian et al., 2019; Olteanu et al., 2018). We refer the reader to recent surveys on the topic to explore more resources (Fortuna and Nunes, 2018; Poletto et al., 2021). However, other languages such as Spanish have received less attention. To the best of our knowledge, only one study has been specifically focused on Spanish to build lexical resources with offensive words (Plaza-Del-Arco et al., 2020), including a lexicon of misogynistic terms and a lexicon of xenophobic words containing 184 and 45 terms, respectively. This resource has been used in an unsupervised approach obtaining results comparable with supervised systems. There are more resources for this language in the literature, but they have been developed multilingually. For instance, HurtLex (Bassignana et al., 2018) is a multilingual lexicon of hate words that covers over 50 languages and is organized into 17 categories such as derogatory words, physical disabilities and diversity, negative stereotypes, and ethnic slurs. Authors started from a preexisting Italian lexical resource (De Mauro, 2016) to perform a semi-automatic multilingual extension using MultiWordNet (Pianta et al., 2002) and BabelNet (Navigli and Ponzetto, 2012). Another resource is Hatebase[3], a collaborative repository of multilingual hate speech. It has been developed to assist companies, government agencies, NGOs, and research organizations to moderate online conversations. It comprises a broad multilingual vocabulary based on nationality, ethnicity, religion, gender, sexual discrimination, disability, and class to monitor incidents of hate speech across countries, specifically it is composed of 3,879 terms, 98 languages, and 176 countries. For Spanish, 142 terms can be found.

Although there are different lexicons available for Spanish, they have mainly two limitations: the number of terms is low and/or they are obtained through a semi-automatic translation from another language. However, the vocabulary used to express hatred or offensiveness is highly dependent on cultural and regional factors. Thus, we consider it important to develop quality resources focused on the specific language based on the vocabulary natively used by the population.

Offensive Language Detection. This challenge involves the use of computational methods to identify offense, aggression, and hate speech in user-generated content. Both binary (eg. offensive, non-offensive) and multiclass classification (eg. automatic categorization of offense types and offense target identification) have been proposed in recent shared tasks such as OffensEval 2019 (Zampieri et al., 2019), OffensEval 2020 (Zampieri et al., 2020), GermEval (Wiegand and Siegel, 2018), HatEval (Basile et al., 2019), and MeOffendEs (Plaza-del-Arco et al., 2021a). Most studies rely on supervised models, from traditional statistics and deep learning models (Davidson et al., 2017; Malmasi and Zampieri, 2018; Plaza-Del-Arco et al., 2020; Badjatiya et al., 2017) to more recent state-of-the-art *Transformers* including BERT, mBERT or XML (Ranasinghe and Zampieri, 2020; Sarkar et al., 2021; Plaza-del Arco et al., 2021b). Some authors have embraced hybrid methodologies incorporating external knowledge from language resources such as lexicons into supervised models, moving beyond the simple lexicon-based approaches. For instance, Koufakou and Scott (2020) explores the use of two types of lexicons (semantic and sentiment) to enhance embedding-based methods for the detection of personal attacks in online conversations. Based on their experimental results, authors claim that semantic lexicon methods outperform baseline methods with at least 4% macro-average F_1 improvement. Vargas et al. (2021) proposed a method that incorporates offensive and sentiment lexicons annotated with contextual information to classify abusive language on Brazilian Portuguese comments, their results show that the proposed approach outperforms baseline methods for Portuguese.

Offensive Span Identification. While most of the efforts have been focused on the offensive language classification, little attention has been paid to the identification of terms that make a text offensive, a task commonly referred to as toxic or offensive spans detection. To the best of our knowledge, only one recent shared task held at SemEval 2021 attempts to address this challenge, namely Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021) in English. For this task, different annotators manually annotate the Civil

[1] https://bit.ly/3LmX8sJ
[2] https://bit.ly/3MmNXbx
[3] https://hatebase.org/

Unigrams	Bigrams	Trigrams	n-grams (N > 3)
91,005 (55.33%)	16,613 (10.10%)	14,649 (8.92%)	42,200 (25.65%)

Table 1: Total of n-grams in the data collection.

Comments dataset[4] with toxic spans. A recent study introduces MUDES, a multilingual system based on state-of-the-art *Transformers* to detect offensive spans in texts. This system has outperformed the strong baselines of the SemEval Task 5 competition.
Given the importance of this task for offensive language research and its interpretability, in this paper, we leverage SHARE to annotate an existing corpus in Spanish named OffendES.

3. Data Collection and Annotation

In this section, we discuss how the data have been collected and filtered as well as the annotation process.

3.1. Collecting Offensive Terms

In order to collect offensive terms, we used the virtual assistant in Telegram Fiero (Botella-Gil et al., 2021). Fiero was developed for encouraging users to insult in a humorous and sarcastic way with the aim of collecting insults and vulgar expressions from Spanish speakers. This tool was released in July 2019 and in 2020 it became more popular with significantly higher interaction due to the great diffusion and repercussion of Fiero in the radio, press and national television media.
A total of 164,467 comments were collected from 2019 to 2021. In this period, we obtained the number of comments shown in Table 1. 122,267 are composed of one, two, and three words (unigrams, bigrams, and trigrams, respectively). The remaining 42,200 are comments consisting of more than three words. We observed that more than half of the comments are composed of one term (unigrams). Table 2 shows the distribution of comments according to the gender and age of the users who interacted with Fiero. It can be seen that the male population over 18 years interacted the most, collecting a total of 95,513 comments. The younger population (<18) participates to a lesser extent, obtaining a total of 17,037 comments compared to 147,430 comments obtained by users over 18 years old.
After the data collection, we accomplished different pre-processing steps by applying NLP-based automated techniques (both regular expressions and using the Python emoji library[5]):

- Comments have been normalized to lowercase.
- Emojis are removed. For instance, *feo*😊 (ugly😊) have been replaced by *feo* (ugly).

[4] https://bit.ly/3Hnj454
[5] https://pypi.org/project/emoji/

Gender	Age	Comments
Female	>18	51,917
	<18	5,922
Male	>18	95,513
	<18	11,115
Total		164,467

Table 2: Total of n-grams obtained according to gender and age in Fiero.

Unigrams	Bigrams	Trigrams
11,936	6,930	7,765

Table 3: N-grams distributions in comments after preprocessing.

- Comments containing one only character, URL, punctuation marks, numbers, and consonants have been deleted.
- Onomatopoeias such as haha, hehe, jaja, jeje including repeated characters and words that are part of the dialogue but not offensive (e.g., *hola* (hello), *adios* (goodbye), *sí* (yes), *seguro* (sure), *no*, *de acuerdo* (ok), *hola* (hello)) are removed.
- Elongated words and repeated characters are reduced, for example, *toonnnto* (sssiilly) is replaced with *tonto* (silly).
- Comments longer than three words are deleted. We select unigrams, bigrams and trigrams to retrieve insults and expressions. We consider that n-grams containing more than three words are part of comments involving a conversation.
- Duplicate comments have been removed.

After the preprocessing phase, we obtained a total of 26,631 comments. Table 3 shows the distribution of these comments, 11,936 are unigrams, 6,930 bigrams and 7,765 trigrams.

3.2. Annotation Procedure

The final collected terms have been annotated by five annotators. Specifically, we defined the following rules to annotate a term/expression as offensive or non-offensive:

- A comment is considered offensive when it contains some form of unacceptable language (profanity or bad words) or a targeted offense, which may be direct or indirect. This category includes insults, threats, and messages containing profane language or profanity. The message may be directed at an individual, at a group of people who share common characteristics, or at others (organization, situation, event, issue or place) (Plaza-del Arco et al., 2021c).

Agreement			
Unigrams	Bigrams	Trigrams	All
0.6369	0.8183	0.8131	0.7881

Table 4: Kappa coefficient for inter-annotator agreement.

- Comments that contain the verb in front of a negative word, such as *eres idiota* (you are an idiot), are classified as non-offensive because we look only for bad words or expressions.
- Comments consisting of two or more consecutive offensive words are labeled as offensive, e.g. *idiota de mierda* (dumb shit).
- As a general rule, food and animal names are considered not offensive. However, there are some words in these contexts that are commonly used to offend. Therefore, we consider *perro/a, zorra, cerdo/a* (dog, fox, pig) as offensive.

Once the rules have been defined, five annotators labeled a subset of the comments in order to compute the agreement. Specifically, each annotator labeled a total of 4,000 terms (2,000 unigrams, 1,000 bigrams and 1,000 trigrams). After the first annotation, we computed the Cohen's kappa coefficient (Cohen, 1960) to determine the agreement between the annotators. These results can be seen in Table 4. The results obtained with respect to the unigrams is 0.6369, which is considered according to Landis and Koch (1977) a substantial value. In the bigrams and trigrams, we obtain a value of near-perfect agreement, 0.8183 and 0.8131, respectively. With these results, we can observe that comments composed of two or three words are easier to categorize as offensive than those consisting of only one word.

After the first annotation and analyzing that the agreement results obtained were favorable, each annotator labeled 4,927 new comments (2,187 unigrams, 1,286 bigrams and 1,453 trigrams), one of the annotators also labeled a unigram to complete the total of 26,631 labeled comments.

3.3. General Lexical Statistics

In order to perform a statistical analysis of the lexical resource developed, we analyzed the number of offensive and non-offensive terms and the distribution of n-grams labeled as offensive.

Figure 1 shows the distribution of the labeled categories according to the different n-grams taken into account, i.e. the number of offensive and non-offensive unigrams, bigrams, and trigrams. As we can see, the number of offensive (5,888) and non-offensive (6,038) unigrams are similar. When we analyze the bigrams, we can see that the number of non-offensive grows significantly to 4,482, almost double the number of offensive bigrams (2,447). Finally, we found 1,790 trigrams in the resource labeled as offensive and 5,975 in the non-offensive category. In total, SHARE is composed of 10,125 offensive expressions distributed as shown in Figure 2. As we can observe, the number of offensive unigrams represents 58.2% of the resource, which means that more than half of the resource is composed of a single offensive word. The remaining n-grams of the resource are covered by the offensive bigrams and trigrams, 24.2% and 17.7% respectively. These data were obtained taking into account that there was no overlap between unigrams, bigrams, and trigrams. For instance, in the trigram *hijo de puta* (son of a bitch), the word *puta* (bitch) is not considered as an unigram.

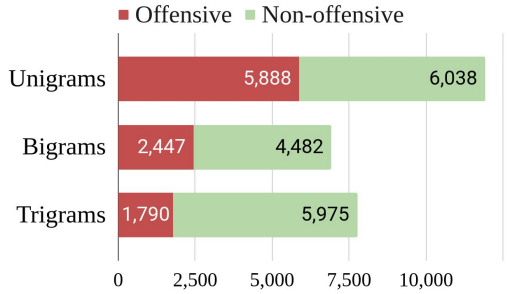

Figure 1: Distribution of the categories annotated according to n-grams selected.

As far as we know, there are available two resources with Spanish offensive terms, the lexicons composed of 502 terms built by (Plaza-Del-Arco et al., 2020) and the HurtLex resource consisting of 2,933 unigrams (Bassignana et al., 2018). We compare them with the SHARE resource in order to observe the difference in terms of size. The SHARE resource exceeds 9,623 insults to the lexicons built by Plaza-Del-Arco et al. (2020) and 7,192 terms to the HurtLex resource. In addition, we checked how many terms match with SHARE, finding that the lexicons built by Plaza-Del-Arco et al. (2020) contain 272 and HurtLex contains 247 matching SHARE terms. In summary, our lexicon offers a large number of offensive terms in the form of insults and expressions commonly used by Spanish speakers.

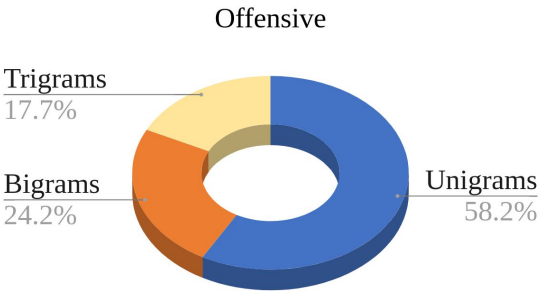

Figure 2: Distribution of n-grams labeled as offensive.

4. Offensiveness Entity Recognition

The development of the SHARE resource allows not only the detection of Spanish offensive texts but also the automatic annotation of offensive entities in corpora. In this section, we leverage an available Spanish corpus labeled with offensive and non-offensive comments to demonstrate the validity of the SHARE resource for NER. The OffendEs resource (Plaza-del Arco et al., 2021c) consists of a dataset based on comments from posts by well-known young Spanish influencers across three social media platforms: Twitter, Instagram, and Youtube. Comments were manually labeled following a fine-grained annotation scheme. The dataset was split by the authors into two sets, one labeled by three annotators (3-Ann) and another labeled by ten annotators (10-Ann). For this research, we consider the 3-Ann subset, consisting of 33,422 instances. The comments in the corpus were annotated into four different categories: offensive, directed to a person (OFP), to a group of people or a collective (OFG); non-offensive, with expletive language (NOE); and non-offensive (NO). Authors group those categories in a binary setup, OFP and OFG are included in the offensive class (OFF) and NO and NOF labels into the non-offensive (NOF) class.

4.1. Corpus annotation

We automatically annotated the OffendES corpus with the terms included in SHARE, we named this new resource OffendES_spans. This strategy involves performing different processing steps to properly match the comments in the corpus with the offensive terms.

The gold standard OffendES_spans corpus has been distributed in CSV format with different fields such as comment, social network, influencer and label, among others. The annotations of offensive terms are included in a separate document (ANN file), with the same name as the ID of the comments.

Two types of entities can be found within the ANN files: OFFENSIVE_TERM, which refers to offensive unigrams, and OFFENSIVE_EXPRESSION, to label entities composed of more than one word (i.e. bigrams and trigrams). Every line of the ANN file contains the mention string of the annotation, its start character offset, and its end character offset, which uniquely locate the mention in the text comment. See Figure 3 for an example of the tab-separated file with the annotation information.

4.2. Exploratory Analysis

After the automatic annotation, we analyzed the offensive terms included in the OffendES comments, i.e, the spans annotated in the OffendES_spans corpus. The 12 most frequent terms annotated are presented in Table 5. As can be seen, the most commonly used offensive terms are *mierda* (shit), *puto* (whore) and *puta* (bitch). Related to bigrams and trigrams, the most frequent ones in the corpus are *puta madre* (fucking mother), *mala persona* (bad person), and *cacho de mierda* (piece of shit). Other terms such as *ignorante de mierda* (ignorant shit), and *necio* (fool) are less frequent but also identified in the corpus.

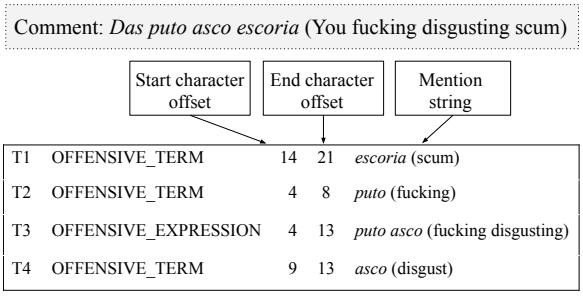

Figure 3: An example of an annotation file in OffendES_spans corpus.

Term	Freq. ↓	Term	Freq. ↓
mierda (shit)	1480	*asco* (disgust)	385
puto (whore)	804	*loca* (crazy)	341
puta (bitch)	706	*gorda* (fat)	336
mala (bad)	510	*coño* (pussy)	331
malo (bad)	442	*basura* (trash)	254
pringada (sucker)	440	*falsa* (false)	239

Table 5: The 12 most frequent entries of offensive terms in OffendES_spans.

Table 6 shows the statistics of the entities found in OffendES_spans using the SHARE resource. Specifically, 11,035 (33.02% of the corpus) comments contain offensive entities from 33,422 comments in OffendES. In the 11,035 comments, 14,311 non-unique entities (repeated) are recognized, where 13,487 (94.24%) are unigrams, 582 (4.12%) are bigrams and 242 (1.64%) are trigrams.

Identification entities/terms	OffendEs_spans
Comments annotated with SHARE	11,035
Unigrams / Uniq. unigrams	13,487 / 636
Bigrams / Uniq. bigrams	582 / 129
Trigrams / Uniq. trigrams	242 / 81

Table 6: Statistics about entities in the OffendEs_spans corpus using SHARE resource. Uniq: unique (not repeated).

In addition, Table 7 shows the total number of NOF and OFF comments that contain at least one offensive entity in OffendES. In the NOF comments (7,293), we found 8,670 unigrams, 329 bigrams, and 83 trigrams. Regarding the OFF comments (3,742), a total of 4,817 unigrams, 253 bigrams, and 159 trigrams are found. It should be noted that the proportion of comments labeled with at least one offensive entity is much higher in the NOF class (21.82%) than in the OFF class

(11.19%) because OffendES is quite unbalanced in the NOF class which include expletive language.

	Comments	Unigrams	Bigrams	Trigrams
NOF	7,293	8,670	329	83
OFF	3,742	4,817	253	159

Table 7: Total number of non-unique unigrams, bigrams and trigrams labeled with SHARE in NOF and OFF.

The OffendES corpus was compiled based on comments from different social networks (Instagram, Twitter, and Youtube). In Table 8 we show the number of entities (unigrams, bigrams, and trigrams) found in the comments categorized by the social media platform. We can observe that the largest number of offensive entities are found on Youtube. Specifically, a total of 11,071 unigrams, 448 bigrams, and 143 trigrams are matched. With a considerable decrease, 1,833 unigrams, 114 bigrams, and 93 offensive trigrams are obtained on Instagram. In the last place, Twitter is the social network with the lowest number of offensive words and expressions including 583 unigrams, 20 bigrams, and 6 trigrams. This result is because of an unbalanced in the number of comments distributed by the social network, 75% of them correspond to Youtube, 18.6% to Instagram, and 6.4% to Twitter.

	Unigrams	Bigrams	Trigrams
Instagram	1,833	114	93
Twitter	583	20	6
Youtube	11,071	448	143

Table 8: Number of non-unique terms labeled in the different social networks.

Finally, as we annotated bigrams and trigrams in the OffendEs corpus, we observed that there are entities that are overlapped (embedded entities). This is considered a challenge for the NLP entity recognition systems. Specifically, we found 589 unigrams which are contained in bigrams. For instance, the entity *puta* (bitch) and *mierda* (shit) are including in the bigram *puta mierda* (fucking piece of shit), or *retrasado* (retarded) into the bigram *retrasado mental* (mentally retarded). A total of 230 unigrams are contained in trigrams such as *violador* (rapist) into *violador de niños* (pedophile) or *puta* (bitch) in *hijo de puta* (son of a bitch) and 26 bigrams are part of trigrams, for instance, *te den* (fuck you) if part of *que te den* (fuck you).

4.3. Toxic Spans Detection

After the OffendES_spans creation, we aimed to develop a system to automatically detect toxic spans in offensive and non-offensive comments and observe its performance. The toxic span detection task attempted to perform the NER task by assigning each token a label.

We used the pre-trained BERT model to detect all possible offensive entities included in a text. To develop the experiments, we fine-tuned the BERT Transformer by using the BETO model (trained on Spanish texts) "*bert-base-spanish-wwm-cased*" according to the Huggingface library (Wolf et al., 2020). Optimization was performed using the Adam optimizer (Kingma and Ba, 2015) with a base learning rate of 1e-5, a batch size of 8 and a maximum sequence of 256.

Table 9 shows the results achieved by the model. As we can see, BERT obtained a 90.01% accuracy, 91.11% recall, and therefore an F_1 score of 91.07%. The results demonstrate the high capability of the transformer-based model in detecting offensive entities by capturing the semantic and syntactic elements of words from a large number of raw text corpora without human intervention. Therefore, we show the utility of SHARE to automatically annotate a corpus with offensive entities and perform the task of automatic offensive span identification.

Model	P (%)	R (%)	F_1 (%)
BERT	91.01	91.11	91.07

Table 9: Evaluation results on toxic spans detection task. P: Precision, R: Recall.

After performing a result and error model analysis, we found that due to the difficulties of the large Spanish vocabulary, BERT was not able to identify offensive terms such as *desequilibrado* (unbalanced), *chismoso* (gossip), *viejuna* (oldie) and *rata de alcantarilla* (sewer rat). In some cases, BERT could not correctly match the start and end of the entities, e.g., the gold standard included *inútil de mierda* (useless shit) and the system only predicted the term *mierda* (shit). However, we observed that the use of transfer learning systems has been crucial in automatically identifying new offensive terms, saving the manual time involved. As a result, BERT recognized offensive terms such as *pendejasito* (little asshole), *aburrida* (boring) and *pederastas* (pedophiles) not included in SHARE.

5. Interpretability for Offensiveness Classification

In order to observe the validity of SHARE as an interpretability tool for offensive language detection in Spanish, we fine-tuned the BERT model on the OffendES_spans corpus and we analyzed a portion of the corpus to compare the attended words with those matched by SHARE.

To perform the fine-tuning of BERT, we replicated the results reported in the OffendES study (Plaza-del Arco et al., 2021c). We obtained a 93.95% F_1 for the NOF class and a 62.82% F_1 in the OFF class, showing a great challenge in the classification of offensive comments. Finally, we achieved a macro-average F_1 of 78.39%.

ID	BERT-LIME	SHARE
818	Das puta pena dalas lo de siempre You're a fucking pity dalas as usual.	puta, das puta pena bitch, you're fucking pitiful
1227	Maldito enano rikillo Damn dwarf rikillo	maldito, enano Damn, dwarf
1545	presa es donde debes estar, pendeja loca. prison is where you belong, you crazy asshole.	pendeja, loca asshole, crazy
1664	No vales nada así te pudras en la cárcel asqueroso de mierda das asco y el MENTIROSO lo eres tú you are worthless hope you rot in jail you filthy piece of shit you are disgusting and you are the liar.	así te pudras, asqueroso, mierda, asqueroso de mierda, asco, das asco, MENTIROSO hope you rot, disgusting, shit, filthy piece of shit, disgusting, you suck, LIER
2071	Eres mala l una psicópata ladrona de perros You are evil and a psychopathic dog thief	mala, psicópata evil, psychopath
3052	TODOS TE ODIAN BRUJA DIFAMADORA, PUDRETE EVERYONE HATES YOU SLANDEROUS WITCH, ROT AWAY	BRUJA, PUDRETE WITCH, ROT AWAY
3173	Tu si que eres patético gañan You really are a pathetic klutz	patético, gañan pathetic, klutz

Table 10: Interpretability comparison between LIME on BERT (BERT-LIME column) and offensive terms matched by the lexicon (SHARE column). Words highlighted in blue are those identified as possibly offensive. These tweets are annotated as *offensive* and classified as *offensive* by BERT.

Regarding the explanation analysis, we used the Local Interpretable Model-agnostic Explanations (LIME) (Tenney et al., 2020) to interpret the individual predictions and to evaluate the confidence of the BERT-based system. LIME is a modular and extensible approach to faithfully explaining the predictions of any model in an interpretable way. Among many other options, LIME provides a score for each feature of the comment. In our study, we have employed LIME on the fine-tuned BERT model for binary classification of offensive and non-offensive comments.

The comparison of the attended words by BERT using LIME and the words matched with SHARE is shown in Table 10. Specifically, seven tweets correctly classified by BERT as offensive are depicted. As can be observed, in most cases, the offensive words identified by BERT match those recognized by our lexicon including *enano* (dwarf), *pendeja* (asshole), *loca* (crazy), *mierda* (shit), *asco* (disgusting), *mala* (evil), *psicópata* (psychopath), *BRUJA* (WITCH), and *patético* (pathetic). Further, there are some instances where SHARE successfully identified offensive terms but BERT failed. For instance, in tweet number 818 the pre-trained language model identifies the word *pena pity* but not the insult *puta* (bitch). Similarly, in tweets number 1664 and 3173 SHARE is able to identify the terms MENTIROSO (LIAR), the offensive expressions *así te pudras* (hope you rot), *asqueroso de mierda* (disgusting piece of shit), *das asco* (you suck) and the swearword *gañan* (klutz). Therefore, we believe that SHARE, in addition to being a helpful tool for explainability, could be incorporated into supervised models to aid classification by developing hybrid methods such as those discussed in Section 2.

6. Conclusion

In this study, we release a new lexical resource composed of offensive words and expressions for Spanish. The vocabulary included in SHARE has been obtained through a previously developed tool (Botella-Gil et al.,

2021). After a thorough cleaning of terms and the generation of an annotation process, these terms have been manually labeled by five annotators with an agreement of 78.8%. Lastly, the resource comprises 5,888 offensive unigrams, 2,447 bigrams, and 1,790 trigrams.

Furthermore, we leverage the SHARE lexicon to automatically label with spans the OffendES corpus which is composed of offensive and non-offensive comments collected from different social networks. With this, we generate the first corpus labeled in Spanish with offensive entities to allow offensive span identification research named OffendES_spans. Using this corpus, we have carried out two different experiments. On the one hand, we have applied the toxic spans detection task using the pre-trained BERT model achieving an F_1 of 91.07%. On the other hand, we also fine-tune a pre-trained model based on BERT for binary classification in OffendES_spans corpus. The output of BERT has been interpreted using an explanation algorithm and compared with the SHARE terms.

In summary, we believe that these generated resources will contribute to the offensive language research community, particularly in Spanish, where there is a great scarcity of resources compared to English. In addition, we believe that these resources will greatly aid in the monitoring of offensive language online, and eventually in the creation of a safer online environment.

7. Acknowledgements

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, LIVINGLANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities of the Spanish Government.

8. Bibliographical References

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (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, June. Association for Computational Linguistics.

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Botella-Gil, B., Plaza-Del-Arco, F. M., Parras Portillo, A. B., and Gutiérrez, Y. (2021). Fiero: Asistente virtual para la captación de insultos. *Procesamiento del Lenguaje Natural*.

Celdrán, P. (2009). *El gran libro de los insultos: tesoro crítico, etimológico e histórico de los insultos españoles*. La Esfera de los Libros.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (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.

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

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

De Mauro, T. (2016). Le parole per ferire. *Internazionale*, 27(9):2016.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July.

Gala, N. and Lafourcade, M. (2010). NLP lexicons: innovative constructions and usages for machines and humans. In *eLEX'2011: Electronic LEXicography in the 21st century: new applications for new users*, page 12.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koufakou, A. and Scott, J. (2020). Lexicon-Enhancement of Embedding-based Approaches Towards the Detection of Abusive Language. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 150–157, Marseille, France, May. European Language Resources Association (ELRA).

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

Lin, B. Y., Lee, D.-H., Shen, M., Moreno, R., Huang, X., Shiralkar, P., and Ren, X. (2020). TriggerNER:

Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July. Association for Computational Linguistics.

Malmasi, S. and Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.*, 30(2):187–202.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Olteanu, A., Castillo, C., Boy, J., and Varshney, K. R. (2018). The Effect of Extremist Violence on Hateful Speech Online. *CoRR*, abs/1804.05704.

Pavlopoulos, J., Sorensen, J., Laugier, L., and Androutsopoulos, I. (2021). SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online, August. Association for Computational Linguistics.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña López, L. A., and Martín-Valdivia, M. T. (2020). Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.*, 20(2), March.

Plaza-del-Arco, F. M., Casavantes, M., Escalante, H. J., Martín Valdivia, M. T., Montejo Ráez, A., Montes y Gómez, M., Jarquín-Vásquez, H., and Villaseñor Pineda, L. (2021a). Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. *Procesamiento del Lenguaje Natural*.

Plaza-del Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021b). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Plaza-del Arco, F. M., Ureña López, L. A., Montejo-Ráez, A., and Martín-Valdivia, M.-T. (2021c). OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, September.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Qian, J., ElSherief, M., Belding, E., and Wang, W. Y. (2019). Learning to Decipher Hate Symbols. 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 3006–3015, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ranasinghe, T. and Zampieri, M. (2020). Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November. Association for Computational Linguistics.

Sarkar, D., Zampieri, M., Ranasinghe, T., and Ororbia, A. (2021). fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., and Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online, October. Association for Computational Linguistics.

Toral, A. and Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Vargas, F., Góes, F., Carvalho, I., Benevenuto, F., and Pardo, T. A. (2021). Contextual-Lexicon Approach for Abusive Language Detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, September.

Wiegand, M. and Siegel, M. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of KONVENS 2018*.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). Inducing a Lexicon of Abusive Words – a Feature-Based Approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Process-*

ing: System Demonstrations, pages 38–45, Online, October. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). 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, June. Association for Computational Linguistics.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (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), December. International Committee for Computational Linguistics.