SHARE: A LEXICON OF HARMFUL EXPRESSIONS BY SPANISH SPEAKERS

L. Alfonso Ureña López

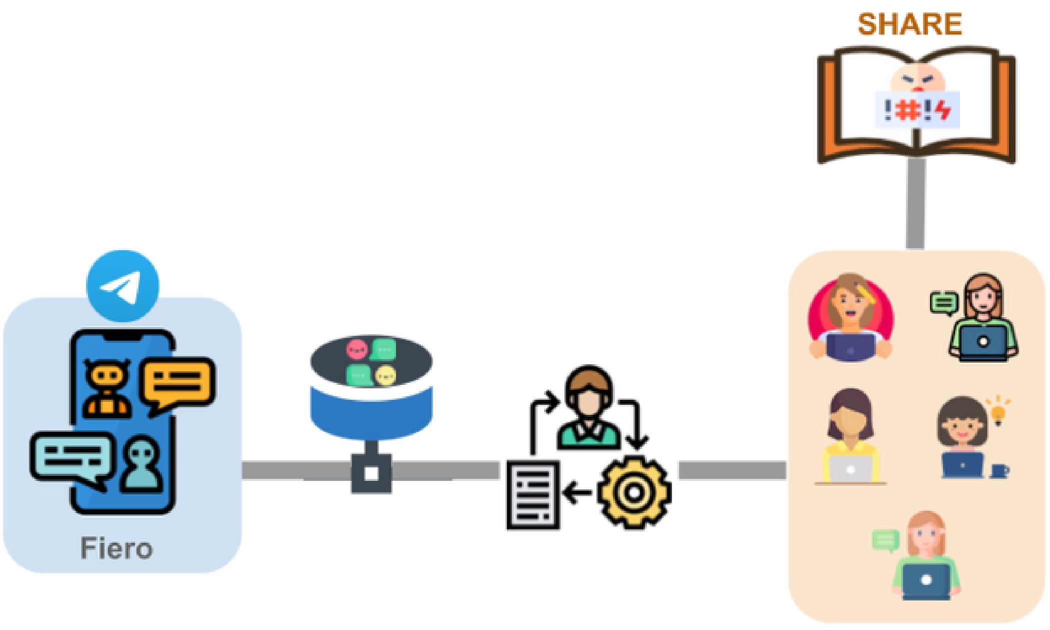
Grupo de Investigación SINAI: Sistemas INteligentes de Acceso a la Información
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén (Spain)

(This work has been presented in LREC2022)

INTRODUCTION

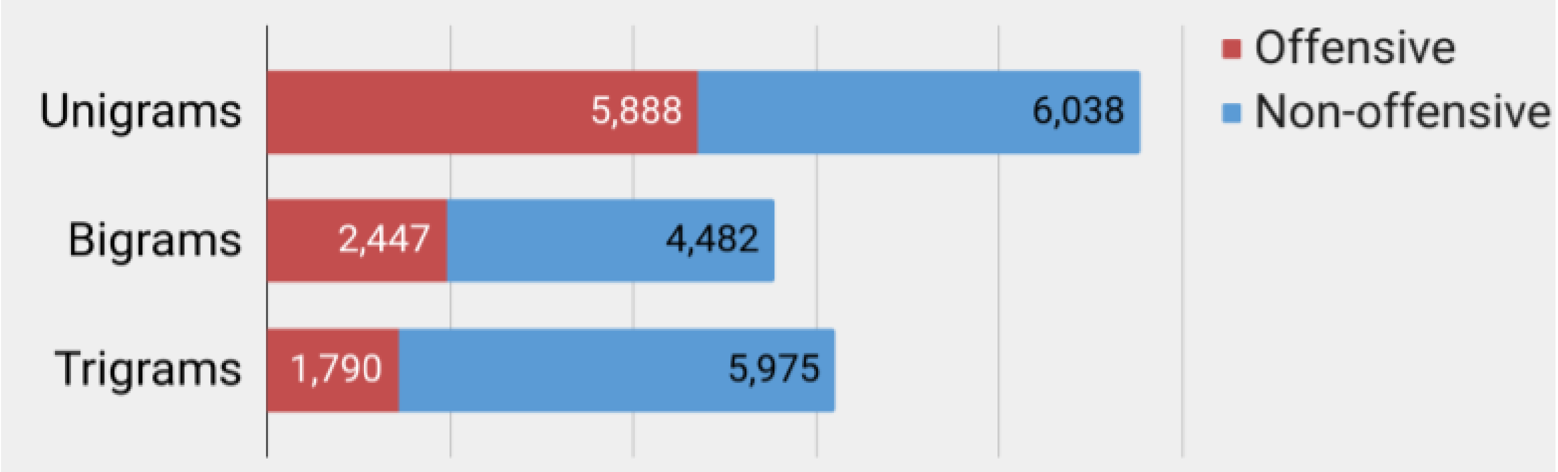
- **SHARE (Spanish HARmful Expressions)** is a new lexical resource composed of **insults and offensive expressions** collected using the **Fiero chatbot** and then **manually labeled** by **5 annotators.**
- We used SHARE to release **OffendES_spans** which is the **OffendES** corpus **automatically annotated** with offensive **entities** relying on **SHARE**.
- We explore the usefulness of **SHARE** for the **interpretability** of offensive comments by comparing it with a BERT-based fine-tuning model.

DATA COLLECTION AND ANNOTATION



STATISTICS

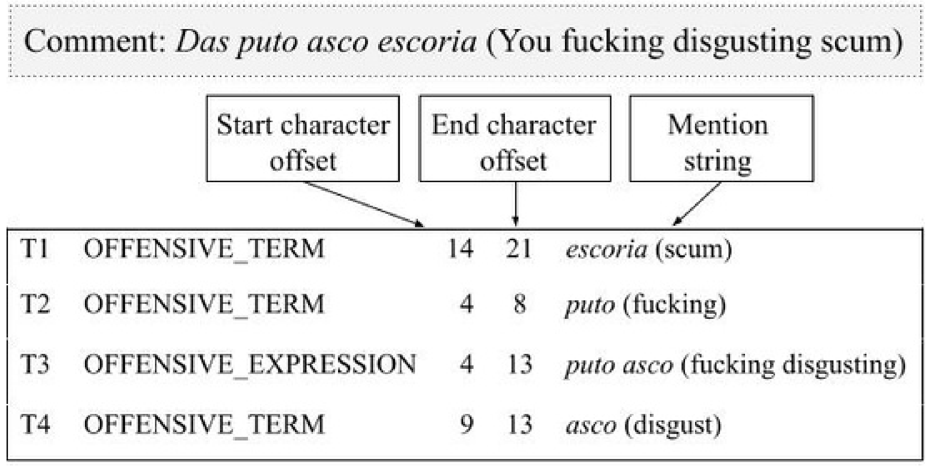
SHARE is composed of **10,125 offensive unigrams and expressions**. The number of offensive **unigrams** represents **58.2%** of the resource followed by **24.2% bigrams** and **17.7% trigrams.**



OFFENSIVE ENTITY RECOGNITION

OffendES_spans

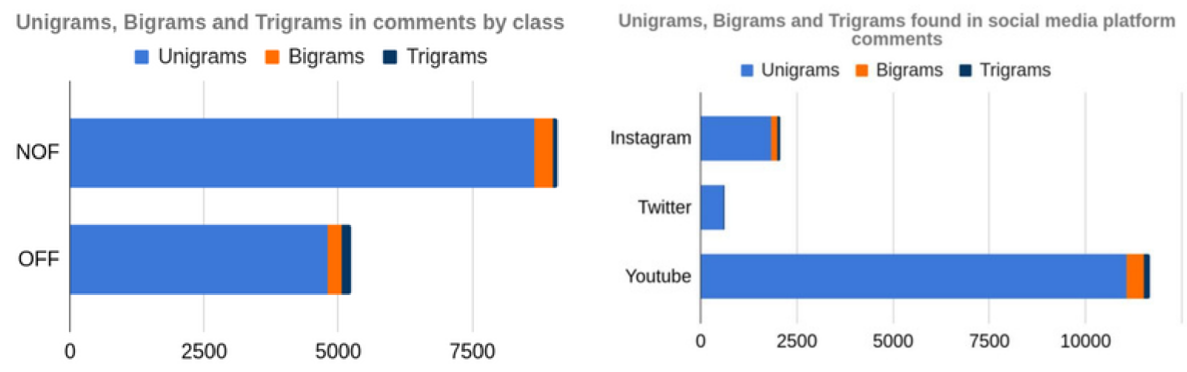
We **automatically annotated** the existing **OffendES** corpus with the terms included in **SHARE**. Two types of entities are found within the ANN files: **OFFENSIVE_EXPRESSION** and **OFFENSIVE_TERM**.



Comment: *Das puto asco escoria* (You fucking disgusting scum)

		Start character offset	End character offset	Mention string
T1	OFFENSIVE_TERM	14	21	*escoria* (scum)
T2	OFFENSIVE_TERM	4	8	*puto* (fucking)
T3	OFFENSIVE_EXPRESSION	4	13	*puto asco* (fucking disgusting)
T4	OFFENSIVE_TERM	9	13	*asco* (disgust)

ANALYSIS

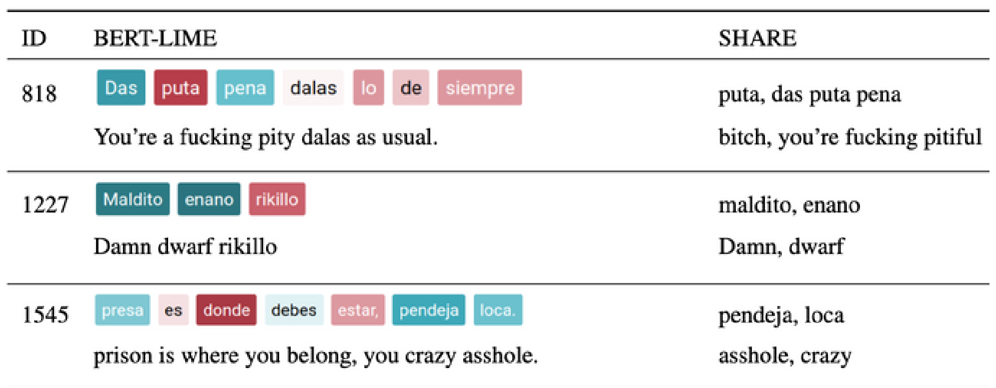


Term	Freq. ↓	Term	Freq. ↓
mierda (shit)	1480	*asco* (disgust)	385
puto (whore)	804	*loca* (crazy)	341
puta (bitch)	706	*gorda* (fat)	336
mala (bad)	510	*coño* (pussy)	331
malo (bad)	442	*basura* (trash)	254
pringada (sucker)	440	*falsa* (false)	239

The 12 most frequent entries in OffendES_spans

TOXIC SPANS DETECTION

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



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

CONCLUSION

We release SHARE, a new lexical resource composed of offensive words and expressions for Spanish. The annotation process by five annotators obtained an agreement of 78.8%. We leverage SHARE to release OffendES_spans by automatically labeled with the terms and expressions found in SHARE. We believe that these new resources will contribute to the offensive language research community, particularly in Spanish, where there is a great scarcity of resources compared to English.

CLARIN CENTRE K
INTELE
DARIAH-EU
Infraestructura de Tecnologías del Lenguaje