SHARE: A LEXICON OF HARMFUL EXPRESSIONS BY SPANISH SPEAKERS

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

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.

DATA COLLECTION AND ANNOTATION

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 OffenES_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.

STATISTICS

Offensive
Non-offensive
Unigrams 5,880 6,038
Bigrams 4,447 4,482
Trigrams 5,975 6,500

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 OffenES_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.

OFFENSIVE ENTITY RECOGNITION

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

ANALYSIS

The 12 most frequent entries in OffenES_spans are:

<table>
<thead>
<tr>
<th>Term</th>
<th>UNigrams</th>
<th>Bigrams</th>
<th>Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>maldito</td>
<td>1680</td>
<td>808</td>
<td>520</td>
</tr>
<tr>
<td>enano</td>
<td>1440</td>
<td>600</td>
<td>360</td>
</tr>
<tr>
<td>putas, puta, pena, bicha</td>
<td>1480</td>
<td>680</td>
<td>340</td>
</tr>
<tr>
<td>soco (scumbag)</td>
<td>500</td>
<td>220</td>
<td>110</td>
</tr>
</tbody>
</table>

TOXIC SPANS DETECTION

<table>
<thead>
<tr>
<th>ID</th>
<th>BERT-LIME</th>
<th>SHARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>818</td>
<td>53%</td>
<td>puta, das, puta, pena, bicha, you're fucking pitiful</td>
</tr>
<tr>
<td>1227</td>
<td>maldito, enano</td>
<td>maldito, enano</td>
</tr>
<tr>
<td>1545</td>
<td>prison is where you belong, you crazy asshole.</td>
<td>pendega, loca, asshole, crazy</td>
</tr>
</tbody>
</table>

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 OffenES_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.

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, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by E8OF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities of the Spanish Government.