

Multilingual Resources for Offensive Language Detection

Aymé Arango^{1,3} Jorge Pérez^{1,3} Bárbara Poblete^{1,3} Valentina Proust^{2,3} Magdalena Saldaña^{2,3}

¹Departament of Computer Science, University of Chile

²Communication Faculty, Pontifical Catholic University of Chile

³Millennium Institute of Data Foundation, Santiago, Chile

Abstract

Most of the published approaches and resources for offensive language and hate speech detection are tailored for the English language. In consequence, cross-lingual and cross-cultural perspectives lack some essential resources. The lack of diversity of the datasets in Spanish is notable. Variations throughout Spanish-speaking countries make existing datasets not enough to encompass the task in the different Spanish variants. We manually annotated 9834 tweets from Chile to enrich the existing Spanish resources with different words and new targets of hate that have not been considered in previous studies. We conducted several cross-dataset evaluation experiments of the models published in the literature using our Chilean dataset and two others in English and Spanish. We propose a comparative framework for quickly conducting comparative experiments using different previously published models. In addition, we set up a Codalab competition for further comparison of new models in a standard scenario, that is, data partitions and evaluation metrics. All resources can be accessed through a centralized repository for researchers to get a complete picture of the progress on the multilingual hate speech and offensive language detection task.

1 Introduction

Offensive language frequently appears on social network interactions ¹. According to Sigurbergsson and Derczynski (2020) *offensive language* encompass a range of expressions from profanities to much more severe types of language among which is *hate speech*. Hate speech is usually defined as communications of animosity or disparagement of an individual or a group on account of a group characteristic². Offensive language and hate speech

bring along the risk of encouraging real hate crimes. Due to the large amount of content generated in social media, automatic moderation is necessary to perform offensive content detection.

Machine learning models are used in most of the published approach for this purpose. The necessary resources are available almost exclusively for the English language (Anzovino et al., 2018; Hosseinmardi et al., 2015; Davidson et al., 2017). On the other hand, the cross-lingual and cross-cultural perspectives have been under addressed in the related literature. The lack of adequately annotated datasets is one of the limiting factors for developing these subtasks (Yin and Zubiaga, 2021; Fortuna and Nunes, 2018). In addition, the publicly available resources are accessible through the correspondent description papers. These resources have insufficient lack of centralized repositories for datasets and classification models. This situation makes it difficult for researchers to get a complete picture of the progress on the task.

Most of these existing datasets contain English examples, though we have gathered some datasets in Portuguese (Fortuna et al., 2019), Arabic (Mulki et al., 2019), Italian (Sanguinetti et al., 2018) and Spanish (Pereira-Kohatsu et al., 2019). In the particular case of the Spanish language, only a few datasets can be found. The geographical origin of them is limited to Spain (Pereira-Kohatsu et al., 2019), México (Álvarez-Carmona et al., 2018), or unknown (Basile et al., 2019). Since the hate speech phenomenon depends on the socio-cultural context (Sap et al., 2019), the targets of hate could change depending on the origin of the messages. The Spanish language specific features spoken in different countries, makes models poorly generalizable when training with these existing resources. We propose a manually annotated dataset for offensive language detection. The dataset is composed of 9834 tweets from Chile and is meant to enrich the existing Spanish resources with different words

¹<https://www.channel4.com/news/george-floyd-death-has-led-to-increasing-online-hate-speech-report-claims>

²<https://www.encyclopedia.com/international/encyclopedias-almanacs-transcripts-and-maps/hate-speech>

and new targets of hate that have not been considered in previous studies.

We conducted several evaluation experiments of the models published in the literature using our Chilean dataset and two others in English and Spanish. We propose a comparative framework for quickly conducting comparative experiments. This framework facilitates the application of existing models by including each original implementation as sub-models. In addition, we set up a Codalab competition for further comparison of new models in a standard scenario, that is, data partitions and evaluation metrics.

In summary, we developed the following resources for multilingual hate speech detection:

1. Chilean dataset for offensive language detection: We annotated a Spanish Twitter dataset in several categories related to the phenomenon of offensive language, including a hate speech category. This dataset is composed of 9834 Spanish tweets and is, as far as we know, the first one where the data was originated in South America.
2. Comparative framework: We constructed a library of models using published cross-lingual offensiveness detectors. The library facilitates the use of models by providing a common interface. Moreover, we set up a Codalab competition for further comparison of emergent models.
3. Resource repository: We organized the existing datasets into a structured repository to facilitate authors finding existing resources in several languages. The repository contains annotations of the main characteristics of the existing datasets and direct links for downloading them. In addition to datasets, it contains tools for using existing multilingual hate speech detection models.

In Section 2, we describe the existing datasets for offensive language detection as well as we comment on the diversity of existing Spanish resources. Next, in Section 3, we describe the Chilean dataset we constructed for offensive language detection, including a hate speech category. Finally, in Section 4, we describe the tools we created for helping the authors to replicate and compare new approaches with the existing ones in a cross-lingual environ-

ment. All resources described in the paper will be integrated in our centralized code repository³.

Ethical Considerations: The annotators inferred only female and male genders of the authors and targets of tweets. The genders were inferred from names and pronouns. Due to the non-binary nature of gender, this label should be used carefully to avoid unfair models.

OFFENSIVE CONTENT WARNING. Because of the topic of our research, certain examples are potentially offensive. We minimized as much as possible the number of examples and obfuscated offensive words.

2 Related Work

One of the essential steps for the research in *offensive language* detection using machine learning is dataset acquisition. Even when several social media platforms exist to get data from them, constructing a balanced labeled dataset is a costly task in time and effort. There is not a dataset considered as standard for this task. Therefore researchers have to search in the related literature for the adequate one for their experiment.

Most of the existing datasets have been annotated for the English language (Dinakar et al., 2011; Hosseinmardi et al., 2015; Waseem and Hovy, 2016; Founta et al., 2018) though there exist a few in other languages such as Spanish (Basile et al., 2019), Italian (Sanguinetti et al., 2018) and Arabic (Mubarak et al., 2017). It is important to mention that even for English, the task is far from being solved (Arango et al., 2020).

In most cases, the datasets only contain texts messages and not other information regarding authors, locallocation, or the conversation to which the tweet belongs. The lack of information makes the datasets out of context and limits the use of different features. Regarding the data sources, most of the datasets have been recovered from the Twitter platform, though a few are composed of Facebook messages (Bosco et al., 2018) or Youtube comments (Dinakar et al., 2011). As far as we know, there exists one data repository⁴ for organizing offensive language datasets.

³<https://github.com/aymeam/Datasets-for-Hate-Speech-Detection>

⁴<https://hatespeechdata.com/>

2.1 Spanish Datasets and the Multicultural Problem:

To the best of our knowledge, there are four different datasets (Basile et al., 2019; Pereira-Kohatsu et al., 2019; Álvarez-Carmona et al., 2018; Fersini et al., 2018) in the Spanish language, related to the task of offensive language detection, with a total of 26 000 messages labeled for hate speech or aggressive content. One of these datasets contained messages that originated in Spain (Pereira-Kohatsu et al., 2019) (6000 tweets). Two of them from unknown origin: IberEval 2018 (Fersini et al., 2018) (4138 tweets) and SemEval 2019 (Basile et al., 2019) (5365 tweets). The remaining dataset was constructed with messages from Mexico: MEX-A3T (11 000 tweets) being the only resource related to the hate speech phenomenon built for Latin-American Spanish.

Being the hate speech phenomenon a cultural problem, we consider that a model trained on these datasets would not be able to generalize over different Spanish data from different cultures.

3 Chilean Dataset for Offensive Language Detection

The research in the Spanish language has been limited, in part, due to the lack of resources. As we described in Section 2, the few Spanish available datasets are composed of examples of the variant of Spanish spoken in specific regions of the world with the cultural background associated with it.

We consider it necessary to leverage the first dataset representative of the Spanish spoken in South America, particularly Chile. The examples in this dataset would enrich the understanding of offensive language and hate speech by introducing terms mainly used in this region and targets of hate unconsidered in previous studies. Next, we describe the process of annotation and general features of our datasets.

3.1 Data Recovering

For recovering an initial corpus, we followed a strategy commonly used in the related literature (Basile et al., 2019; Waseem and Hovy, 2016) which is identifying words that serve as seeds for querying online platforms. The use of seeds would guarantee a higher probability for hateful content to appear in the crawled data.

Seeds The seeds were gathered by surveying a group of seven Chilean students. The list includes

terms (or phrases) used in Chile. Some of these terms are offensive, but others are neutral terms related to polemic subjects such as sexual nature, immigration, and others (e.g. *haitianos*, *indígenas*, *lesbianas*). We recovered a total of 132 seeds that can be read in our code repository.

Search Parameters Using the pre-defined seeds and with the help of the Twitter API⁵, we downloaded approximately 61 000 tweets. The tweets' language was restricted to Spanish, and the geo-location was prefixed for the Chile area. Along with each tweet, we recovered the conversation (sequence of tweets) that originated them in case of existing. These conversations serve as context for each tweet (Qian et al., 2019).

Sample for Annotation From the 61 000 tweets recovered, we selected 10 000 (one-sixth), taking a proportional amount of tweets originating from each seed. In this way, we maintained a representative sample of all sources.

3.2 Annotation

Three external annotators under contract conducted the process of labeling the dataset, all three were native Chileans. First, they went through a training process, where the three of them labeled the same set of tweets to make sure they annotated the content as similarly as possible. They repeated this process with different sets of tweets until achieving an inter-annotator agreement higher than 90% agreement and a Krippendorff's alpha higher than 0.7 in all the pre-defined labels (Neuendorf, 2002). After the training process, they proceeded to label the final dataset, a portion each. Table 1 contains a summary of this measure obtained during the training process.

3.3 Chilean Dataset Description

The final dataset contains 9834 tweets annotated with several labels, some of them related to offensive content based on Chen's categorization of uncivil speech (Chen, 2017). In addition, it includes annotations that contextualize the messages, such as the target of offensive speech and the use of irony. As described above, the dataset also contains the conversation that originated each of them. These conversations serve as context for the annotated tweets. Next, we explore the main characteristics of the resulting dataset.

⁵<https://developer.twitter.com/en/docs/twitter-api>

3.3.1 Offensive Content Labels



Some of the labels in the final dataset encompass different types of offensive content. These labels are *hate speech*, *unintended profanity/vulgarity*, *insult/appellation*, *intentional profanity/vulgarity*. The other labels are not directly related to the offensive phenomenon, but help contextualize the messages and generalize the dataset.

hate speech The tweet contains hate speech if it includes stereotypical language to offend minority groups such as women, immigrants, sexual or racial minorities.

For example, the tweet: “*La mapuche es un asqueroso trapo y los mapuches; cero aporte, son gasto, daño y destrucción, tampoco originarios.*” (“*The Mapuche woman is a disgusting rag and the Mapuche people; zero contribution, they are a waste of money, damage, and destruction, not natives either.*”) is labeled as hateful because the author is attributing detrimental characteristics to the *mapuche* people which are a minority group of indigenous people in Chile and Argentina.

Hate against this particular minority is also an example of the dependence of the hate speech phenomenon on socio-cultural factors.



insult/appellation A tweet is labeled as positive for insults or name calling if the tweet includes nicknames, phrases, or words that are not profane but are offensive (such as “s***id” or “j**k”).

For example: “ *está “mujer” me da vergüenza ajena.*” (“ *This “woman” embarrasses me*”), is labeled as containing insulting language because the intention is to offend a person (this woman) without using profane words. On the other hand, the tweet: “*Ma***to flaite hediondo a marihuana.*” (“*D**n marijuana-smelly chav.*”) also belongs to this class because of the use of “*flaite*” a pejorative word used in Chile for referring to marginal or uneducated people (Rojas, 2015).

unintended profanity/vulgar language Some tweets may contain profane words without the intention of offending anyone, like in: “*Que manera de echar de menos ese estadio por la grandísima co***a de su madre*” (“*I really miss that mother f*** stadium*”). This kind of tweet is labeled as containing unintended profanity. In this case, *mother f**** is an expression used for making emphasis on how much the author misses the stadium.

| Label | Positives (%) | K |
|----------------------------------------------------------------|---------------|------|
| intentional profanity/vulgarity grosería c/intención | 2668 (27,13) | 0,72 |
| unintended profanity/vulgarity grosería s/intención | 1358 (13,80) | 0,75 |
| insult/appellation insulto/sobrenombre | 4036 (41,04) | 0,86 |
| hate speech discurso de odio | 633 (6,43) | 0,74 |
| migration migración | 405 (4,11) | 0,84 |
| Venezuela Venezuela | 199 (20,2) | 0,73 |
| domestic politics política nacional | 3438 (34,96) | 0,81 |
| marginalized gropus grupos marginalizados | 886 (9,0) | 0,74 |
| “others” “otros” | 5220 (53,08) | 0,73 |
| sarcasm/irony/mockery sarcasmo/ironía/burla | 2125 (21,60) | 0,7 |
| legitimate question pregunta legítima | 89 (0,9) | 1 |
| evidence evidencia | 427 (4,34) | 0,71 |
| female figure figura femenina | 1436 (14,60) | 0,72 |
| male figure figura masculina | 2872 (29,20) | 0,75 |
| anonymous author autor anonimo | 6391 (6498) | 0,92 |
| female author author femenino | 2102 (21,37) | |
| male author author masculino | 4695 (47,74) | 0,81 |
| unk-gender author género desconocido | 3037 (30,88) | |

Table 1: The column “Label” shows each label of the dataset in both, English and in Spanish languages. The column “Positives (%)” shows the number and percent of tweets labeled as positive for each label. Finally, the column “K” shows the Krippendorff’s measure obtained during the training stage for each of the labels.

intentional profanity/vulgar language: A different type of profanity can be found in the tweet: “*Les dije que el árbitro era un CO***A DE SU MADRE*  ” (“*I told you the referee was a MOTHER F****  ”). Even when we have the same words as in previous example, in this case, the annotators marked this tweet as containing *intentional profanity*, as the author has the intention to insult a person using profane words (*the referee*).

3.3.2 Tweets Content

Other labels are meant to enrich the dataset by spotting linguistic and semantic information of the tweets. In this sense, we can find annotations regarding the content of the tweet.

male figure: The tweet labels containing male or female figures are the ones, offensive or not, directed to a particular person identified by annotators as male, for example: “*Tremendo hijo de p**a eres Marcos.*” (“*You are a tremendous son of a b***, Marcos.*”) is labeled as *male or female figure* since the message is directed to *Antonio*, a male.

female figure: Similar to the *male figure* label, the tweet: “*Y q dice la autodenominada candidata feminista al respecto*” (“*And what does the self-appointed feminist candidate have to say about it?*”) is labeled as *female figure* since the author poses a question to a female (*feminist candidate*).

mention to [topic] There are five labels used to mark when a tweet makes reference to different topics such as *immigration*, *domestic politics*, *marginalized groups* and *others*. As an example of *domestic politics* is the tweet “*Vamos a botar a la feminazi*, 🇺🇸 #VOTACIONES2021” (“*We are going to kick out the feminazi*, 🇺🇸 #ELECTIONS2021”).

sarcasm/irony/mockery The use of humor or sarcasm was also identified in this label. This label could be helpful to disambiguate the message’s intention, that is, the intention of hurting. (e.g. “*Aquí llenando la piscina con las lágrimas de los fachos*” (“*Here filling the pool with fascists’ tears*”).

evidence This category is based on Chen’s (Chen, 2017) definition of deliberative speech, a condition set to foster healthy conversations on social media. The tweets are labeled positive for evidence if they provide statistical evidence, citations, or links with extra information instead of a mere opinion. For example: “*Expulsión de migrantes efectuada este domingo en la RM https://t.co/**** vía @*****” (“*Expulsion of migrants carried out this Sunday in the Metropolitan Region https://t.co/**** via @*****”) is labeled as *evidence* because it includes a link to a news source.

legitimate question Also based on Chen’s work (2017), a tweet contains a *legitimate question* if it poses a non-rhetorical question, for example asking for more information about a particular event, like in the tweet: “*¿A los venezolanos le están solicitando visa para entrar a Peru?*” (“*Are Venezuelans requested to have a visa to enter Peru?*”).

| | | insult | prof/vulg | hate | off |
|---------|-----|--------|-----------|------|------|
| dummy | F1 | 48.7 | 45.9 | 49.3 | 48.6 |
| seed | F1 | 58.8 | 51.8 | 47.9 | 51.6 |
| EMB +RF | F1 | 66.3 | 69.8 | 55.5 | 66.0 |
| | ROC | 77.3 | 76.0 | 79.8 | 71.8 |

Table 2: The Table shows the F-score obtained using different baselines in different classification tasks over our dataset. Baselines: dummy = random predictions; seed = all messages containing one of the offensive seeds used for recovering the dataset is predicted as positive; EMB+RF = Spanish Glove Embeddings and Radom Forest Classifier; Tasks: insult, profanity/vulgarity (prof/vulg); hate and offensive (off) detection.

3.3.3 Tweets’ Author Information

All the tweets contain a label of the authors’ gender: 2102 tweets were sent by a *female author*, 4694 by a *male author*. The rest of the authors were identified as *undetermined-gender* since the user name does not suggest either a male or female gender (e.g., “DVM”; “Patria y Libertad”). The annotators also labeled information about the anonymity of the authors. The tweet is labeled as anonymous if the username is a nickname (e.g. “DVM”) or a name without last name (e.g. “patricia”). There are 5371 unique Twitter users in the dataset.

The 50,67% of the tweets in any offensive categories were sent by users labeled as males, 20,22% by females and the rest from undetermined-gender users.

Table 1 contains a summary of the dataset columns. A sample of the dataset can be found in our repository ⁶ and will be completely published soon.

3.4 Offensive Content Detection Baselines

We implemented some baselines for offensive language detection over our dataset. We defined different classification tasks: *insult*, *profanity/vulgarity* (intentional or not) and *hate speech* detection. In addition, we tested baselines to identify if a tweet belongs to any of the offensive classes. Therefore, we set the target *offensive* if the tweet is labeled as any of the offensive labels (*insult* or *profanity/vulgarity* or *hate speech*). The results were obtained with a 5-Fold cross validation .

⁶https://anonymous.4open.science/r/Datasets-for-Hate-Speech-Detection-0D50/Chilean%20dataset/Dataset_sample_500.csv

dummy classifier We predict the values of the classes randomly, making use of the Sklearn⁷ dummy classifier.

seed classifier: To verify that there is no seed bias, we conducted a baseline classification method consisting of labeling as positive those tweets containing one of the offensive seeds previously used to recover the dataset (See Section 3.1). Our results show the best performance on the *insult* detection task showing a higher bias in this category. The list of offensive seeds can be found in our code repository. This result was expected since this category is positive depending on the existence of certain words. On the other hand, the rest of the tasks showed nearly random results.

EMB + RF We tested a third baseline using Spanish FastText embeddings⁸ and Random Forest classifier. The word embeddings of 100 dimensions were first averaged into one single vector and used as input for a Random Forest Classifier with default parameters. We show the results for 5-fold cross-validation. The results with this approach, compare to dummy and seed classifiers, showed the best results.

The F-Score obtained with the different methods in the four tasks can be shown in Table 2.

4 Comparative Framework

In the related literature of offensive language detection, there is a lack of comparative studies. This situation is more noticeable in cross-lingual approaches as a relatively new sub-area. There is no consensus about the best approaches for solving the cross-lingual detection task.

With the purpose of alleviating this situation, we propose two tools for cross-lingual approaches comparison:

1. A python library that contains published cross-lingual hate speech detection models as methods: The library has five published models. Each model consists of the original implementation code as a sub-module, plus a class interface that standardizes all models' input to simplify their use. In addition, the library contains the main class whose attributes are the previously mentioned models and auxiliary tools for evaluation and data management. A

⁷<https://scikit-learn.org/stable/>

⁸<https://github.com/dccuchile/spanish-word-embeddings>

| | | ACL19 | EMNLP20 | ECML20 |
|---------------------|-----|-------|---------|--------|
| $EN \rightarrow ES$ | F1 | 48.42 | 53.26 | 64.56 |
| | ROC | 50.83 | 63.42 | 73.14 |
| $ES \rightarrow EN$ | F1 | 45.54 | 60.22 | 60.09 |
| | ROC | 49.20 | 69.53 | 63.91 |
| $EN \rightarrow CL$ | F1 | 49.17 | 38.19 | 48.83 |
| | ROC | 50.12 | 48.16 | 60.85 |
| $CL \rightarrow EN$ | F1 | 44.58 | 47.33 | 51.6 |
| | ROC | 51.25 | 47.83 | 54.33 |

Table 3: Cross-lingual experiments using there different datasets: English (Basile et al., 2019) (EN), Spanish (Basile et al., 2019) (ES), and our Spanish dataset recovered from Chile (CL). Models: ACL19 (Pamungkas and Patti, 2019); EMNLP20 (Ranasinghe and Zampieri, 2020); ECML20 (Aluru et al., 2020); WEBSci21 (Vitiugin et al., 2021).

brief description of the models can be found in Section 4.1

2. An open competition in Codalab⁹ for further comparison. We set up an open competition in Codalab to promote fair comparison among cross-lingual approaches. Different leaderboards can be found for the different configurations.

4.1 Cross-lingual Models

We found a few papers describing cross-lingual approaches. We included them in our library using the original companion code.

ACL19 As a preparation stage for the model proposed by Pamungkas et al. (2021), it is necessary to translate the data into the target language. The model consists of training two different LSTM architectures. The first one is trained with the original training data, and the other is trained using the data translated into the testing language. Finally, the two outputs are concatenated and used as input for a final linear output layer.

ECML20 In this paper, Aluru et al. (2020) described different approaches for cross-lingual hate speech detection with different architectures. Those are the multilingual Bert model, the GRU model, and a combination of LASER embeddings and Logistic Regression (LR) classifier. The model

⁹https://codalab.lisn.upsaclay.fr/competitions/1221?secret_key=c1de3893-de48-4ca1-8071-89e82f189039

that combines LASER embeddings and LR classifier turned to be the best approach. Our library includes three types of models, though in Table 3 we only report the best results.

EMNLP20 [Ranasinghe and Zampieri \(2020\)](#) proposed a transfer learning strategy. First, an XLM-R classification model is trained using data from one language, and the weights are saved. Then, these weights are used to initialize the model and predict labels in a different language.

We used our library for reproducing the previously mentioned models in a cross-lingual way using three different languages English, Spanish ([Basile et al., 2019](#)), and our Chilean dataset. In Table 3, we show the results we obtained in different cross-lingual experiments.

4.2 Evaluation Datasets

For evaluation, we used the Spanish (*ES*) and English (*EN*) datasets constructed for the SemEval 2019 competition ([Basile et al., 2019](#)). As we mentioned in Section 2, the authors of these datasets did not specify any location for recovering the data. Examining the tweets objects of the Spanish dataset, we noticed only a few with geo-location information, some belonging to Spain, México, though most of them were unknown. We compare the cross-datasets performance with the performance across different variants of Spanish: general Spanish (*ES*) and the variant of Spanish spoken in Chile (*CL*). To this end, we add experiments using our previously described Chilean (*CL*) dataset. We show precision, recall, and F-score metrics, the commonly used metrics, and the ROC metric.

4.2.1 Cross-lingual Results

In general, the cross-lingual setup, including the Spanish (*ES*) dataset, performed better than Chilean (*CL*). One of the reasons for this could be the data used for pre-trained models; for example, *ECML20* model is based on LASER representations. These are multilingual sentence embeddings constructed from parallel data. The data used may not encompass some of the words used in South America, though a more profound analysis is needed. Despite presenting a simple structure (LASER + LR), *ECML20* model showed the overall best results.

4.2.2 Cross-cultural Results

We tested the models in monolingual Spanish setups but using datasets from different socio-cultural

| | | ACL19 | EMNLP20 | ECML20 |
|---------------------|-----|-------|---------|--------|
| $CL \rightarrow ES$ | F1 | 50.0 | 53.1 | 56.7 |
| | ROC | 51.2 | 57.0 | 64.2 |
| $ES \rightarrow CL$ | F1 | 46.1 | 41.3 | 46.7 |
| | ROC | 49.9 | 46.6 | 53.0 |

Table 4: Cross-cultural experiments using two different datasets: Spanish ([Basile et al., 2019](#)) (*ES*) and our Spanish dataset recovered from Chile (*CL*). Models: ACL19 ([Pamungkas and Patti, 2019](#)); EMNLP20 ([Ranasinghe and Zampieri, 2020](#)); ECML20 ([Aluru et al., 2020](#)); WEBSci21 ([Vitiugin et al., 2021](#)).

contexts.

One of the datasets is the SemEval Spanish dataset ([Basile et al., 2019](#)) with examples originated in Spain. The other is our dataset, also in Spanish, but originated in Chile. The results in terms of F1 and ROC are shown in Table 4.

The best overall results were obtained using the *ECML20* model in the $CL \rightarrow ES$ configuration. Despite being datasets from the same language, the knowledge transfer was, in general, poor. All the results were lower than the ones obtained in an inside-dataset experiment shown in Table 2. These results evidence of the differences between the two Spanish variants, the different hate targets of the two geographical regions, though much more inside in this regard is needed.

4.3 Repository Description

To facilitate finding an appropriate dataset, we organized them in a centralized repository. So far, we have listed 39 datasets, 20 of which are in the English language and 19 others in different languages such as Arabic (5), Spanish (4), Italian (3), Portuguese (1), among others.

In our repository the datasets are separated by languages and have the following structure:

- **Datasets (Link to paper):** Abbreviated name of the dataset with a link for downloading the paper description.
- **Objects:** Which are the type of objects (e.g. *tweets, images, sentences*).
- **Size:** The number of objects in the dataset.
- **Available:** A direct link for downloading the dataset is provided if the dataset is publicly available.

- Labels: The labels in which the objects are categorized (e.g. (*hateful, non-hateful*), (*racist, sexist, either*))

Approximately, 64% are composed of tweets, but other objects can be found, such as Facebook comments or Twitter users. Although some of the below-listed datasets are not explicitly available, they could be obtained from the authors if requested. Our comparative framework (Section 4) facilitates the use of previously published models for cross lingual hate speech detection.

5 Conclusions

We described three resources for the multilingual offensive language detection task. These resources would be helpful in the development of the multilingual sub-area, which have been under-addressed.

We constructed the first Chilean dataset for hate speech and offensive language to alleviate this situation. The dataset contains 9834 tweets in the Spanish language that originated in Chile. The tweets are labeled in several categories related to offensive content. Furthermore, it includes annotations associated with the content of the tweets.

Finally, we created a comparative framework (library + competition) to facilitate researchers to compare new models with the existing ones. The library is implemented in python and contains, as submodels, previously published cross-lingual approaches for hate speech detection. The competition is hosted in Codalab and offers a scenario for comparing new models with the existing ones.

The resource repository would facilitate researchers to find, in one place, the datasets that better meet their needs as well as tools for easily comparing their work with previously existing models. From our repository, it is noticeable the lack of available Spanish examples. Moreover, there is a low representation of different types of Spanish spoken worldwide.

6 References

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. [Deep learning models for multilingual hate speech detection](#). *CoRR*, abs/2004.06465.

Miguel Á Álvarez-Carmona, Estefania Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-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 *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain*, volume 6.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2020. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *IS*, page 101584.

Valerio Basile, Cristina Bosco, Viviana Patti, Manuela Sanguinetti, Elisabetta Fersini, Debora Nozza, Francisco Rangel, and Paolo Rosso. 2019. Shared task on multilingual detection of hate. *SemEval 2019*, Task 5.

Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Gina Masullo Chen. 2017. *Online incivility and public debate: Nasty talk*. Springer.

Thomas Davidson, Dana Warmesley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. pages 512–515.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 Workshop (ICWSM)*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. [Large scale crowdsourcing and characterization of twitter abusive behavior](#). In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

- Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International Conference on Social Informatics*, pages 49–66. Springer.
- Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.
- Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. [L-HSAB: A Levantine Twitter dataset for hate speech and abusive language](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.
- Kimberly A Neuendorf. 2002. Defining content analysis. *Content analysis guidebook*. Thousand Oaks, CA: Sage.
- Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.
- Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proc. 57th ACL*, pages 363–370.
- 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.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.
- Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*.
- Darío Rojas. 2015. Flaite: algunos apuntes etimológicos. *Alpha (Osorno)*, (40):193–200.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. pages 2798–2895.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1668–1678.
- Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508.
- Fedor Vitiugin, Yaras Senarath, and Hemant Purohit. 2021. Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In *13th ACM Web Science Conference 2021*.
- Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. SRW@HLT-NAACL*, pages 88–93.
- Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.