A Deep Dive into Multilingual Hate Speech Classification

Sai Saketh Aluru, Binny Mathew[(✉)], Punyajoy Saha, and Animesh Mukherjee

Indian Institute of Technology Kharagpur, Kharagpur, India
{saisakethaluru,binnymathew,punyajoys}@iitkgp.ac.in,
animeshm@cse.iitkgp.ac.in

Abstract. Hate speech is a serious issue that is currently plaguing the society and has been responsible for severe incidents such as the genocide of the Rohingya community in Myanmar. Social media has allowed people to spread such hateful content even faster. This is especially concerning for countries which lack hate speech detection systems. In this paper, using hate speech dataset in 9 languages from 16 different sources, we perform the first extensive evaluation of multilingual hate speech detection. We analyze the performance of different deep learning models in various scenarios. We observe that in low resource scenario LASER embedding with Logistic regression perform the best, whereas in high resource scenario, BERT based models perform much better. We also observe that simple techniques such as translating to English and using BERT, achieves competitive results in several languages. For cross-lingual classification, we observe that data from other languages seem to improve the performance, especially in the low resource settings. Further, in case of zero-shot classification, evaluation on Italian and Portuguese dataset achieve good results. Our proposed framework could be used as an efficient solution for low-resource languages. These models could also act as good baselines for future multilingual hate speech detection tasks. Our code (Code: https://github.com/punyajoy/DE-LIMIT) and models (Models: https://huggingface.co/Hate-speech-CNERG) are available online.
Warning: contains material that many will find offensive or hateful.

Keywords: Hate speech · Multilingual · Classification · BERT · Embeddings

1 Introduction

Online social media has allowed dissemination of information at a faster rate than ever [22,23]. This has allowed bad actors to use this for their nefarious purposes such as propaganda spreading, fake news, and *hate speech*. Hate speech is defined as a "direct and serious attack on any protected category of people based on their

---

Table 1. Examples of hate speech.

Text	Hate speech?
I f**king hate ni**ers!	Yes
Jews are the worst people on earth and we should get rid of them	Yes
"6 million was not enough. next time ovens will be the least of your concerns # sixmillionmore"	Yes
Mexicans are f**king great people!	No

race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease" [11]. Representative examples of hate speech are provided in Table 1.

Hate speech is increasingly becoming a concerning issue in several countries. Crimes related to hate speech have been increasing in the recent times with some of them leading to severe incidents such as the genocide of the Rohingya community in Myanmar, the anti-Muslim mob violence in Sri Lanka, and the Pittsburg shooting. Frequent and repetitive exposure to hate speech has been shown to desensitize the individual to this form of speech and subsequently to lower evaluations of the victims and greater distancing, thus increasing outgroup prejudice [35]. The public expressions of hate speech has also been shown to affect the devaluation of minority members [18], the exclusion of minorities from the society [26], and the discriminatory distribution of public resources [12].

While the research in hate speech detection has been growing rapidly, one of the current issues is that majority of the datasets are available in English language only. Thus, hate speech in other languages are not detected properly and this could be detrimental. While there are few datasets [3,27] in other language available, as we observe, they are relatively small in size.

In this paper, we perform the first large scale analysis of multilingual hate speech by analyzing the performance of deep learning models on 16 datasets from 9 different languages. We consider two different scenarios and discuss the classifier performance. In the first scenario (monolingual setting), we only consider the training and testing from the same language. We observe that in low resource scenario models using LASER embedding with Logistic regression perform the best, whereas in high resource scenario, BERT based models perform much better. We also observe that simple techniques such as translating to English and using BERT, achieves competitive results in several languages. In the second scenario (multilingual setting), we consider training data from all the other languages and test on one target language. Here, we observe that including data from other languages is quite effective especially when there is almost no training data available for the target language (aka zero shot). Finally, from the summary of the results that we obtain, we construct a catalogue indicating which model is effective for a particular language depending on the extent of the data available. We believe that this catalogue is one of the most important contributions of our work which can be readily referred to by future researchers working to advance the state-of-the-art in multilingual hate speech detection.

The rest of the paper is structured as follows. Section 2 presents the related literature for hate speech classification. In Sect. 3, we present the datasets used for the analysis. Section 4 provides details about the models and experimental settings. In Sect. 5, we note the key results of our experiments. In Sect. 6 we discuss the results and provide error analysis.

2   Related Works

Hate speech lies in a *complex nexus with freedom of expression, individual, group and minority rights, as well as concepts of dignity, liberty and equality* [16]. Computational approaches to tackle hate speech has recently gained a lot of interest. The earlier efforts to build hate speech classifiers used simple methods such as dictionary look up [19], bag-of-words [6]. Fortuna *et al.* [13] conducted a comprehensive survey on this subject.

With the availability of larger datasets, researchers started using complex models to improve the classifier performance. These include deep learning [37] and graph embedding techniques [30] to detect hate speech in social media posts. Zhang *et al.* [37] used deep neural network, combining convolutional and gated recurrent networks to improve the results on 6 out of 7 datasets used. In this paper, we have used the same CNN-GRU model for one of our experimental settings (monolingual scenario).

Research into the multilingual aspect of hate speech is relatively new. Datasets for languages such as Arabic and French [27], Indonesian [21], Italian [33], Polish [29], Portuguese [14], and Spanish [3] have been made available for research. To the best of our knowledge, very few works have tried to utilize these datasets to build multilingual classifiers. Huang *et al.* [20] used Twitter hate speech corpus from five languages and annotated them with demographic information. Using this new dataset they study the demographic bias in hate speech classification. Corazza *et al.* [8] used three datasets from three languages (English, Italian, and German) to study the multilingual hate speech. The authors used models such as SVM, and Bi-LSTM to build hate speech detection models. Our work is different from these existing works as we perform the experiment on a much larger set of languages (9) using more datasets (16). Our work tries to utilize the existing hate speech resources to develop models that could be generalized for hate speech detection in other languages.

3   Dataset Description

We looked into the datasets available for hate speech and found 16 publicly[1] available sources in 9 different languages[2]. One of the immediate issues, we observed was the mixing of several types of categories (offensive, profanity, abusive, insult

---

[1] Note that although Table 2 contains 19 entries, there are three occurrences of Ousidhoum *et al.* [27] and two occurrences of Basile *et al.* [3] for different languages.

[2] We relied on http://hatespeechdata.com for most of the datasets.

etc.). Although these categories are related to hate speech, they should not be considered as the same [9]. For this reason, we only use two labels: *hate speech* and *normal*, and discard other labels. Next, we explain the datasets in different languages. The overall dataset statistics are noted in Table 2.

Arabic: We found two arabic datasets that were built for hate speech detection.

– Mulki *et al.* [25] : A Twitter dataset[3] for hate speech and abusive language. For our task, we ignored the abusive class and only considered the hate and normal class.
– Ousidhoum *et al.* [27]: A Twitter dataset[4] with multi-label annotations. We have only considered those datapoints which have either hate speech or normal in the annotation label.

Table 2. Dataset details

Language	Dataset	Source	Hate	Non-hate	Total
Arabic	Mulki *et al.* [25]	Twitter	468	3,652	4,120
	Ousidhoum *et al.* [27]	Twitter	755	915	1,670
English	Davidson *et al.* [9]	Twitter	1,430	4,163	5,593
	Gibert *et al.* [17]	Stormfront	1,196	9,748	10,944
	Waseem *et al.* [36]	Twitter	759	5,545	6,304
	Basile *et al.* [3]	Twitter	5,390	7,415	12,805
	Ousidhoum *et al.* [27]	Twitter	1,278	661	1,939
	Founta *et al.* [15]	Twitter	4,948	53,790	58,738
German	Ross *et al.* [32]	Twitter	54	315	369
	Bretschneider *et al.* [5]	Facebook	625	5,161	5,786
Indonesian	Ibrohim *et al.* [21]	Twitter	5,561	7,608	13,169
	Alfina *et al.* [1]	Twitter	260	453	713
Italian	Sanguinetti *et al.* [33]	Twitter	231	1,329	1,560
	Bosco *et al.* [4]	Facebook & Twitter	3,355	4,645	8,000
Polish	Ptaszynski *et al.* [29]	Twitter	598	9,190	9,788
Portuguese	Fortuna *et al.* [14]	Twitter	1,788	3,882	5,670
Spanish	Basile *et al.* [3]	Twitter	2,228	3,137	5,365
	Pereira *et al.* [28]	Twitter	1,567	4,433	6,000
French	Ousidhoum *et al.* [27]	Twitter	399	821	1,220
Total			32,890	126,863	159,753

---

[3] https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset.
[4] https://github.com/HKUST-KnowComp/MLMA_hate_speech.

English: Majority of the hate speech datasets are available in English language. We select six such publicly available datasets.

– Davidson *et al.* [9] provided a three class Twitter dataset[5], the classes being hate speech, abusive speech, and normal. We have only considered the hate speech and normal class for our task.
– Gibert *et al.* [17] provided a hate speech dataset[6] consisting sentences from Stormfront[7], a white supremacist forum. Each sentence is tagged as either hate or normal.
– Waseem *et al.* [36] provided a Twitter dataset[8] annotated into classes: sexism, racism, and neither. We considered the tweets tagged as sexism or racism as hate speech and neither class as normal.
– Basile *et al.* [3] provided multilingual Twitter dataset[9] for hate speech against immigrants and women. Each post is tagged as either hate speech or normal.
– Ousidhoum *et al.* [27] provided Twitter dataset (See Footnote 6) with multi-label annotations. We have only considered those datapoints which have either hate speech or normal in the annotation label.
– Founta *et al.* [15] provided a large dataset[10] of 100K annotations divided in four classes: hate speech, abusive, spam, and normal. For our task, we have only considered the datapoints marked as either hate or normal, and ignored the other classes.

German: We select two datasets available in German language.

– Ross *et al.* [32] provided a German hate speech dataset[11] for the refugee crisis. Each tweet is tagged as hate speech or normal.
– Bretschneider *et al.* [5] provided a Facebook hate speech dataset[12] against foreigners and refugees.

Indonesian. We found two datasets for the Indonesian language.

– Ibrohim *et al.* [21] provided an Indonesian multi-label hate speech and abusive dataset[13]. We only consider the hate speech label for our task and other labels are ignored.
– Alfina *et al.* [1] provided an Indonesian hate speech dataset[14]. Each post is tagged as hateful or normal.

---

[5] https://github.com/t-davidson/hate-speech-and-offensive-language.
[6] https://github.com/aitor-garcia-p/hate-speech-dataset.
[7] www.stormfront.org.
[8] https://github.com/zeerakw/hatespeech.
[9] https://github.com/msang/hateval.
[10] https://github.com/ENCASEH2020/hatespeech-twitter.
[11] https://github.com/UCSM-DUE/IWG_hatespeech_public.
[12] http://www.ub-web.de/research/.
[13] https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection.
[14] https://github.com/ialfina/id-hatespeech-detection.

Italian. We found two datasets for the Italian language.

– Sanguinetti *et al.* [33] provided an Italian hate speech dataset[15] against the minorities in Italy.
– Bosco *et al.* [4] provided hate speech dataset[16] collected from Twitter and Facebook.

Polish. We found only one dataset for the Polish language

– Ptaszynski *et al.* [29] provided a cyberbullying dataset[17] for the Polish language. We have only considered hate speech and normal class for our task.

Portuguese. We found one dataset for the Portuguese language

– Fortuna *et al.* [14] developed a hierarchical hate speech dataset[18] for the Portuguese language. For our task, we have used the binary class of hate speech or normal.

Spanish. We found two dataset for the Spanish language.

– Basile *et al.* [3] provided multilingual hate speech dataset (See Footnote 11) against immigrants and women.
– Pereira *et al.* [28] provided hate speech dataset[19] for the Spanish language.

French

– Ousidhoum *et al.* [27] provided Twitter dataset (See Footnote 6) with multi-label annotations. We have only considered those data points which have either hate speech or normal in the annotation label.

4   Experiments

For each language, we combine all the datasets and perform stratified train/ validation/ test split in the ratio 70%/10%/20%. For all the experiments, we use the same splits of train/val/test. Thus, the results are comparable across different models and settings. We report macro F1-score to measure the classifier performance. In case we select a subset of the dataset for the experiment, we repeated the subset selection with 5 different random sets and report the average performance. This would help to reduce the performance variation across different sets. In our experiments, the subsets are stratified samples of size $16, 32, 64, 128, 256$.

---

[15] https://github.com/msang/hate-speech-corpus.
[16] https://github.com/msang/haspeede2018.
[17] http://poleval.pl/tasks/task6.
[18] https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset.
[19] https://zenodo.org/record/2592149.

4.1 Embeddings

In order to train models in multilingual setting, we need multilingual word/sentence embeddings. For sentences, LASER embeddings were used and for words MUSE embeddings were used.

Laser embeddings: LASER[20] denotes Language-Agnostic SEntence Representations [2]. Given an input sentence, LASER provides sentence embeddindgs which are obtained by applying max-pooling operation over the output of a BiLSTM encoder. The system uses a single BiLSTM encoder with a shared BPE vocabulary for all languages.

Muse embeddings: MUSE[21] denotes Multilingual Unsupervised and Supervised Embeddings. Given an input word, MUSE gives as output the corresponding word embedding [7]. MUSE builds a bilingual dictionary between two languages without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way.

4.2 Models

CNN-GRU (Zhang *et al.* [37]): This model initially maps each of the word in a sentence into a 300 dimensional vector using the pretrained Google News Corpus embeddings [24]. It also pads/clips the sentences to a maximum of 100 words. Then this 300×100 vector is passed through drop layer and finally to a 1-D convolution layer with 100 filters. Further, a maxpool layer reduces the dimension to 25×100 feature matrix. Now this is passed through a GRU layer and it outputs a 100×100 dimension matrix which is globally max-pooled to provide a 1×100 vector. This is further passed through a softmax layer to give us the final prediction.

BERT: BERT [10] stands for Bidirectional Encoder Representations from Transformers pretrained on data from english language. It is a stack of transformer encoder layers with multiple "heads", i.e. fully connected neural networks augmented with a self attention mechanism. For every input token in a sequence, each head computes key value and query vectors which are further used to create a weighted representation. The outputs of each head in the same layer are combined and run through a fully connected layer. Each layer is wrapped with a skip connection and a layer normalization is applied after it. In our model we set the token length to 128 for faster processing of the query[22].

mBERT: Multilingual BERT (mBERT[23]) is a version of BERT that was trained on Wikipedia in 104 languages. Languages with a lot of data were sub-sampled and others were super sampled and the model was pretrained using the same

---

[20] https://github.com/facebookresearch/LASER.

[21] https://github.com/facebookresearch/MUSE.

[22] In the total data 0.17% datapoints have more than 128 tokens when tokenized, thus justifying our choice.

[23] https://tinyurl.com/yxh57v3a.

method as BERT. mBERT generalizes across some scripts and can retrieve parallel sentences. mBERT is simply trained on a multilingual corpus with no language IDs, but it encodes language identities. We used mBERT to train hate speech detection model in different languages once again limiting to a maximum of 128 tokens for sentence representation.

Translation: One simple way to utilize datasets in different languages is to rely on translation. Simple techniques of translation has shown to give good results in tasks such as sentiment analysis [34]. We use Google Translate[24] to convert all the datasets in different languages to English since translation to English from other languages typically have less errors in comparison to the other way round.

For our experiments we use the following four models:

1. **MUSE + CNN-GRU:** For the given input sentence, we first obtain the corresponding MUSE embeddings which are then passed as input to the CNN-GRU model.
2. **Translation + BERT:** The input sentence is first translated to the English language which are then provided as input to the BERT model.
3. **LASER + LR:** For the given input sentence, we first obtain the corresponding LASER embeddings which are then passed as input to a Logistic Regression (LR) model.
4. **mBert:** The input sentence is directly fed to the mBert model.

4.3   Hyperparameter Optimization

We use the validation set performance to select the best set of hyperparameters for the test set. The hyperparameters used in our experiments are as follows: batch size: 16, learning rate: $2e^{-5}, 3e^{-5}, 5e^{-5}$ and epochs: $1, 2, 3, 4, 5$.

5   Results

5.1   Monolingual Scenario

In this setting, we use the data from the same language for training, validation and testing. This scenario commonly occurs in the real world where monolingual dataset is used to build classifiers for a specific language.

Observations: Table 3 reports the results of the monolingual scenario. As expected, we observe that with increasing training data, the classifier performance increases as well. However, the relative performance seem to vary depending on the language and the model. We make several observations. First, **LASER + LR** performs the best in low-resource settings (16, 32, 64, 128, 256) for all the languages. Second, we observe that **MUSE + CNN-GRU** performs the worst in almost all the cases. Third, **Translation + BERT** seems to achieve

---

competitive performance for some of the languages such as German, Polish, Portuguese, and Spanish. Overall we observe that there is no 'one single recipe' for all languages; however, **Translation + BERT** seems to be an excellent compromise. We believe that improved translations in some languages can further improve the performance of this model.

Although **LASER + LR** seems to be doing good in low resource setting, if enough data is available, we observe that BERT based models: **Translation + BERT** (English, German, Polish, and French) and **mBERT** (Arabic, Indonesian, Italian, and Spanish) are doing much better. However, what is more interesting is that although BERT based models are known to be successful when a larger number of datapoints are available, even with 256 datapoints some of these models seem to come very close to **LASER + LR**; for instance, **Translation + BERT** (Spanish, French) and **mBERT** (Arabic, Indonesian, Italian).

5.2   Multilingual Scenario

In this setting, we will use the dataset from all the languages expect one $(N-1)$, and use the validation and test set of the remaining language. This scenario represents when one wishes to employ the existing hate speech dataset to build a classifier for a new language. We have considered **LASER + LR** and **mBERT** that are most relevant for this analysis. In the **LASER + LR** model, we take the LASER embeddings from the $(N-1)$ languages and add to this the target language data points in incremental steps of $16, 32, 64, 128$ and 256. The logistic regression model is trained on the combined data, and we test it on the held out test set of the target language.

For using the multilingual setting in **mBERT** we adopt a two-step fine-tuning method. For a language L, we use the dataset for $N-1$ languages (except the Lth language) to train the **mBERT** model. On this trained **mBERT** model, we perform a second stage of fine-tuning using the training data of the target language in incremental steps of $16, 32, 64, 128, 256$. The model was then evaluated on the test set of the Lth language.

We also test the models for zero shot performance. In this case, the model is not provided any data of the target language. So, the model is trained on the $(N-1)$ languages and directly tested on the Nth language test set. This would be the case in which we would like to directly deploy a hate speech classifier for a language which does not have any training data.

Observations: Table 4 reports the results of the multilingual scenario. Similar to the monolingual scenario, we observe that with increasing training data, the classifier performance increases in general.

This is especially true in low resource settings of the target languages such as English, Indonesian, Italian, Polish, Portuguese.

In case of zero shot evaluation, we observe that **mBERT** performs better than **LASER + LR** in three languages (Arabic, German, and French). **LASER + LR** perform better on the remaining six languages with the results in Italian

Table 3. Monolingual scenario: the training, validation and testing data is used from the same language. Here, Full D represents the full training data. The **bold** figures represent the best scores and <u>underline</u> represents the second best.

Language	Model	Training size					
		16	32	64	128	256	Full D
Arabic	MUSE + CNN-GRU	0.4412	0.4438	0.4486	0.4664	0.5818	0.7368
	Translation + BERT	0.4555	0.4495	<u>0.5551</u>	0.5448	0.7017	<u>0.8115</u>
	LASER + LR	**0.5533**	**0.6755**	**0.7304**	**0.7488**	**0.7698**	0.7920
	mBert	<u>0.4588</u>	<u>0.4533</u>	0.4408	<u>0.6486</u>	<u>0.7295</u>	**0.8320**
English	MUSE + CNN-GRU	<u>0.4580</u>	<u>0.4594</u>	<u>0.4653</u>	0.4646	<u>0.4813</u>	0.6441
	BERT	0.4071	0.3925	0.4260	<u>0.4720</u>	0.4578	**0.7143**
	LASER + LR	**0.4617**	**0.4899**	**0.5376**	**0.5624**	**0.5885**	0.6526
	mBert	0.1773	0.3251	0.4488	0.4578	0.4578	<u>0.7101</u>
German	MUSE + CNN-GRU	0.4708	0.4708	0.4708	0.4708	0.4762	0.5756
	Translation + BERT	0.4812	<u>0.4758</u>	<u>0.4719</u>	<u>0.4729</u>	0.4724	**0.7662**
	LASER + LR	<u>0.4974</u>	**0.5201**	**0.5465**	**0.5925**	**0.6488**	<u>0.6873</u>
	mBert	**0.5037**	0.4750	0.4708	0.4717	<u>0.5022</u>	0.6517
Indonesian	MUSE + CNN-GRU	0.4250	0.4823	0.5263	0.5354	0.5890	0.7110
	Translation + BERT	0.4957	0.5003	0.5179	0.5682	0.6341	0.7670
	LASER + LR	**0.5226**	**0.5376**	**0.5882**	**0.6259**	**0.6890**	<u>0.7872</u>
	mBert	<u>0.5106</u>	<u>0.5219</u>	<u>0.5414</u>	<u>0.6016</u>	<u>0.6530</u>	**0.8119**
Italian	MUSE + CNN-GRU	0.4055	0.4476	0.4461	0.5206	0.5965	0.7349
	Translation + BERT	0.5006	<u>0.5943</u>	<u>0.6215</u>	<u>0.6678</u>	0.6919	0.7922
	LASER + LR	<u>0.5688</u>	**0.6210**	**0.6843**	**0.7175**	**0.7347**	<u>0.7996</u>
	mBert	**0.5774**	0.4567	0.5834	0.6664	<u>0.7026</u>	**0.8260**
Polish	MUSE + CNN-GRU	<u>0.4842</u>	0.4842	0.4841	<u>0.4842</u>	<u>0.5180</u>	0.6337
	Translation + BERT	<u>0.4842</u>	<u>0.4853</u>	<u>0.4842</u>	<u>0.4842</u>	0.5066	**0.7161**
	LASER + LR	**0.4889**	**0.4879**	**0.5360**	**0.5739**	**0.6172**	0.6439
	mBert	0.4829	0.4847	<u>0.4842</u>	<u>0.4842</u>	0.4842	<u>0.7069</u>
Portuguese	MUSE + CNN-GRU	0.4480	0.3807	0.4184	0.4228	0.4562	0.6100
	Translation + BERT	0.4532	<u>0.4893</u>	<u>0.4712</u>	0.5102	<u>0.5994</u>	<u>0.6935</u>
	LASER + LR	**0.5194**	**0.5536**	**0.6070**	**0.6210**	**0.6412**	**0.6941**
	mBert	<u>0.5154</u>	0.4245	0.4148	<u>0.5493</u>	0.5745	0.6713
Spanish	MUSE + CNN-GRU	0.4382	0.3354	0.3558	0.4203	0.4995	0.6364
	Translation + BERT	<u>0.4598</u>	<u>0.4722</u>	<u>0.5080</u>	0.4576	<u>0.6035</u>	<u>0.7237</u>
	LASER + LR	**0.5168**	**0.5434**	**0.5521**	**0.5938**	**0.6153**	0.6997
	mBert	0.4395	0.4285	0.4048	<u>0.4861</u>	0.5999	**0.7329**
French	MUSE + CNN-GRU	<u>0.4878</u>	<u>0.4683</u>	<u>0.5008</u>	<u>0.5222</u>	0.5250	0.5619
	Translation + BERT	0.4173	0.4260	0.4429	0.4749	<u>0.6037</u>	**0.6595**
	LASER + LR	**0.5058**	**0.5486**	**0.6136**	**0.6302**	**0.6085**	<u>0.6172</u>
	mBert	0.4818	0.4139	0.4053	0.4355	0.5701	0.6165

Table 4. Multilingual scenario: the training data is from all the languages except one and the validation and testing data is from the remaining language. The **bold** figures represent the best scores.

Testing language	Model	Training size						
		Zero shot	16	32	64	128	256	Full D
Arabic	LASER + LR	0.4645	**0.4651**	0.4664	0.4704	0.4784	0.4930	0.6751
	mBert	**0.6442**	0.4535	**0.4738**	**0.5302**	**0.7331**	**0.7707**	**0.8365**
English	LASER + LR	**0.6050**	**0.6051**	**0.6052**	**0.6053**	**0.6054**	0.6060	0.6808
	mBert	0.4971	0.4750	0.4670	0.5044	0.5242	**0.6091**	**0.7374**
German	LASER + LR	0.4695	0.4661	0.4727	0.4729	**0.4740**	0.4784	0.5622
	mBert	**0.5437**	**0.5146**	**0.4927**	**0.4733**	0.4718	**0.4786**	**0.6651**
Indonesian	LASER + LR	**0.6263**	**0.6251**	**0.6252**	**0.6241**	**0.6182**	0.6151	0.5977
	mBert	0.5113	0.5186	0.5049	0.4871	0.5864	**0.6318**	**0.8044**
Italian	LASER + LR	**0.6861**	**0.6857**	**0.6855**	**0.6855**	**0.6860**	0.6867	0.7071
	mBert	0.5335	0.5318	0.5444	0.6696	0.6704	**0.7189**	**0.8147**
Polish	LASER + LR	**0.5912**	**0.5926**	**0.5931**	**0.5935**	**0.5901**	**0.5829**	0.5672
	mBert	0.0725	0.4961	0.5049	0.4841	0.4842	0.4842	**0.6670**
Portuguese	LASER + LR	**0.6567**	**0.6565**	**0.6566**	**0.6563**	**0.6565**	**0.6573**	**0.6755**
	mBert	0.5995	0.5526	0.5694	0.5961	0.6148	0.6294	0.6660
Spanish	LASER + LR	**0.5408**	**0.5415**	**0.5417**	**0.5406**	0.5434	0.5437	0.5708
	mBert	0.2677	0.4464	0.4751	0.5126	**0.6080**	**0.6302**	**0.7383**
French	LASER + LR	0.4228	0.4180	0.4171	0.4180	0.4181	0.4198	0.4684
	mBert	**0.5487**	**0.5310**	**0.5138**	**0.5698**	**0.5849**	**0.5948**	**0.5968**

and Portuguese being pretty good. In case of Portuguese, zero shot **Laser + LR** (without any Portuguese training data) obtains an F-score of 0.6567, close to the best result of 0.6941 (using full Portuguese training data).

For the languages such as Arabic, German, and French, **mBERT** seems to be performing better than **LASER + LR** is almost all the cases (low resource and Full D). **LASER + LR**, on the other hand, is able to perform well for Portuguese language in all the cases. For the rest of the five languages, we observe that **LASER + LR** is performing better in low resource settings, but on using the full training data of the target language, **mBERT** performs better.

5.3   Possible Recipes Across Languages

As we have used the same test set for both the scenarios, we can easily compare the results to access which is better. Using the results from monolingual and multilingual scenario, we can decide the best kind of models to use based on the availability of the data. The possible recipes are presented as a catalogue in Table 5. Overall we observe that **LASER + LR** model works better for low resource settings while BERT based models work well for high resource settings. This possibly indicates that BERT based models, in general can work well when there is larger data available thus allowing for a more accurate fine-tuning. We believe that this catalogue is one of the most important contributions of our work which can be readily referred to by future researchers working to advance the state-of-the-art in multilingual hate speech detection.

Table 5. The table describes the best model to use in low and high resource scenario. In general, LASER + LR performs well in low resource setting and BERT based models are better in high resource settings

Language	Low resource	High resource
Arabic	Monolingual, LASER + LR	Multilingual, mBERT
English	Multilingual, LASER + LR	Multilingual, mBERT
German	Monolingual, LASER + LR	Translation + BERT
Indonesian	Multilingual, LASER + LR	Monolingual, mBERT
Italian	Multilingual, LASER + LR	Monolingual, mBERT
Polish	Multilingual, LASER + LR	Translation + BERT
Portuguese	Multilingual, LASER + LR	Monolingual, LASER+LR
Spanish	Monolingual, LASER + LR	Multilingual, mBERT
French	Monolingual, LASER + LR	Translation + BERT

6   Discussion and Error Analysis

6.1   Interpretability

In order to compare the interpretability of **mBERT** and **LASER + LR**, we use LIME [31] to calculate the average importance given to words by a particular model. We compute the top 5 most predictive words and their attention for each sentence in the test set. The total score for each word is calculated by summing up all the attentions for each of the sentences where the word occurs in the top 5 LIME features. The average predictive score for each word is calculated by dividing this total score by the occurrence count of each word. In Table 6 we note the top 5 words having the highest attention scores and compare them qualitatively across models.

While comparing the models' interpretability in Table 6, we see that **LASER + LR** focuses more on the hateful keywords compared to **mBERT**, i.e., words like 'pigs' etc. **mBERT** seems to search for some context of the hate keywords as shown in Table 7. Models dependent on the keywords can be useful when we are in a highly toxic environment such as GAB[25] since most of the derogatory keywords typically occur very close or at least simultaneously along with the hate target, for e.g., the first case in Table 1. In sites which are less toxic like Twitter, complex methods giving attention to the context like **mBERT** might be more helpful,for e.g., the third case in Table 1.

6.2   Error Analysis

In order to delve further into the models, we conduct an error analysis[26] on both the **mBERT** and **LASER + LR** models using a sample of posts where the

---

[25] https://en.wikipedia.org/wiki/Gab_(social_network).

[26] Note that we rely on translation for interpretations of the errors and the translation itself might also have some error.

output was wrongly classified from the test set.We analyze the common errors and categorize them into the following four types:

1. **Wrong classification due to annotation's dilemma (AD):** These error cases occur due to ambiguous instances where according to us the model predicts correctly but the annotators have labelled it wrong.
2. **Wrong classification due to confounding factors (CF):** These error cases are caused when the model predictions rely on some irrelevant features like normalized form of mentions (*@user*) and links (*URL*) in the text.
3. **Wrong classification due to hidden context (HC):** These error cases are caused when the model fails to capture the context of the post.
4. **Wrong classification due to abusive words (AW):** These error cases are caused by over-dependence of the model on the abusive words.

Table 8 shows the errors of the **mBERT** and **LASER + LR** models. For **mBERT**, the first example has no specific indication of being a hate speech and is considered an error on the part of annotators. In the second example the author of the post actually wants the reader to not use the abusive terms, i.e., sl*t

Table 6. Interpretations of the model outcomes.

German		Indonesian	
mBERT	LASER + LR	mBERT	LASER + LR
spendieren (spend)	fotzen (pu**ies)	loo (loo)	NAJIS (unclean)
drogen (drugs)	Trottel (fool)	rusak (broken)	bajingan (son of a bi**h)
schœn (beautiful)	abschaum (scum)	makhluk (creature)	MAMPUS (dead)
kastrieren (castrate)	WICHSER (w**ker)	pengkhianatan (betrayal)	Idiot (idiot)
einsetzen (deploy)	Scheissen (shit)	celeng (wild boar)	F**kYou (f**k you)
Italian		Polish	
mBERT	LASER + LR	mBERT	LASER + LR
innervosirmi (get nervous)	Schifo (schifo)	stanowisk (posts)	pieprzysz (f**k)
vomitata (vomited)	demoliscile (demoliscile)	pomysł (idea)	gówno (shit)
cascarci (fall for)	disonesti (dishonest)	powiedzieli (they said)	idiota (idiot)
italioti (italioti)	massacrale (massacrale)	cwelica (cwelica)	Idiotów (idiots)
annegano (drown)	schifoso (lousy)	obrazka (picture)	świry (suck)
Portuguese		Spanish	
mBERT	LASER + LR	mBERT	LASER + LR
fuder (f**k)	FOFURA (cuteness)	Hxrry_again (hxrry_again)	piratas (pirates)
heterofobicos (heterophobic)	tretas (fights)	majisimos (majestic)	MARICA (sissy)
vagabunda (slut)	porcaria (filth)	mate (mate)	perseguidos (persecuted)
cracuda (crunchy)	foda (f**k)	publicidad (advertising)	pegaso6038 (pegasus6038)
femimimismo (feminism)	heterofobicos (heterophobic)	sevilla (seville)	Putas (wh**es)
French		English	
mBERT	LASER + LR	mBERT	LASER + LR
mongol (mongolian)	jérusalem (jerusalem)	n**guh	Rapist
medelin (medelin)	ptdrrrrrrrrrr (ptdrrrrrrrrrr)	fa**otnwe	BlackB4Illegal
arabe (arab)	negrophobe (ne*rophobe)	n**let	Misttgg
barges (barges)	juifs (jews)	nig	sexualevak
marocains (moroccons)	bf (bf)	h*ed	prostitute
Arabic			
mBERT	LASER + LR		
تّساعده (help him)	خنازير (pigs)		
ورد (flower)	رووووح (rooooh)		
يامطي (my run)	ياكلبة (eat it)		
خارجية (external)	خنزير (pig)		
لهجة (dialect)	قذره (dirty)		

Table 7. Examples showing word with the highest predictive word for both **mBERT** and *LASER + LR*.

Sentences with hate label
das *pack* muss tag und nacht **gejagt** werden,ehe sie es mit den deutschen machen !! (**Translated :-** the *pack* must be **hunted** day and night before they do it with the Germans !!)
absolument ! il faut l'arraisonner en mer par la marin nationale arrêter tous les occupants **expulser** les *migrant*... @url (**Translated :-** absolutely! it must be boarded at sea by the navy national arrest all occupants **expel** *migrants*... @url)

Table 8. Various types of errors **(E)** for the models **(M)** : mBERT and LASER + LR. The ground truth **(GT)** and prediction **(P)** consist of 0 (Non-Hate)/1 (Hate) label.

M	Sentences	GT	P	E
mBERT	Arabic Translation: He and his father, and Abu Alto and Abu Israel, are doomed to go to Israel to blind, insolent Syrian opponents, and to betray that I have not seen and my eyes have seen	1	0	AD
	"If you have tries to get w/a girl you are not allowed to call her demeaning names like "slut whore etc." sorry bout yall"	0	1	AW
	"Könnten wir Schmarotzer und Kriminelle loswerden würde die Asylanten-Schwemme auf beherrschbare Zahlen runtergehen" **Translation:** If we could get rid of parasites and criminals, the asylum seeker flood would drop to manageable numbers	1	0	HC
LASER + LR	"Die hat jede Art von Realität verloren und braucht dringend Hilfe am besten ne Einweisung in die Geschlossene für immer und Ewig und ihr Gefolge gleich mit" **Translation:** She has lost all kind of reality and urgently needs help, best a briefing in the closed forever and ever and her followers at the same time	1	0	AD
	"USER USER Gw mah tetep anti cina... gara gara gw ngga bisa sipit dan putih kayak mereka...wkwkwk" **Translation:** USER USER I am still anti-Chinese ... because I can't be narrow and white like them ... hahaha	0	1	CF
	"RT @mundodrogado: Antes o homossexualismo era proibido.Depois passou a ser tolerado.Hoje é normal. Eu vou embora antes que vire obrigatór" **Translation:** RT @mundodrogado: Before homosexuality was forbidden. Then it became tolerated. Today it's normal. I'm leaving before it becomes mandatory...	1	0	HC
	this movie is actually good cuz its so retarded	0	1	AW

and wh*re (*found using LIME*) but the model picks them as indicators of hate speech. The third example has mentioned the term "parasite" as a derogatory remark to refugees and the model did not understand it.

For the **LASER + LR** model, the first example is an error on the part of the annotators. In the second case the model captures the word "USER" (*found using LIME*), a confounding factor which affects the models' prediction. For the third case, the author says (s)he will leave before homosexuality gets normalized which shows his/her hatred toward the LGBT community but the model is unable to capture this. In the last case the model predicts hate speech based on the word "retarded" (*found using LIME)* which should not be the case.

7   Conclusion

In this paper, we perform the first large scale analysis of multilingual hate speech. Using 16 datasets from 9 languages, we use deep learning models to develop classifiers for multilingual hate speech classification. We perform many experiments under various conditions – low and high resource, monolingual and multilingual settings – for a variety of languages. Overall we see that for low resource, LASER + LR is more effective while for high resource BERT models are more effective. We finally suggest a catalogue which we believe will be beneficial for future research in multilingual hate speech detection. Our Code (Code: https://github. com/punyajoy/DE-LIMIT) and Models (Models: https://huggingface.co/Hate-speech-CNERG) are available online for other researchers to use.

References

1. Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the Indonesian language: a dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 233–238. IEEE (2017)
2. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguist. **7**, 597–610 (2019)
3. Basile, V., et al.: 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, pp. 54–63 (2019)
4. Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T.: Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. vol. 2263, pp. 1–9. CEUR (2018)
5. Bretschneider, U., Peters, R.: Detecting offensive statements towards foreigners in social media. In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017)
6. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Sci. **5**(1), 11 (2016)
7. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)

8. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: A multilingual evaluation for online hate speech detection. ACM Trans. Internet Technol. (TOIT) **20**(2), 1–22 (2020)

9. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media (2017)

10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018)

11. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding, E.: Hate lingo: a target-based linguistic analysis of hate speech in social media. In: Twelfth International AAAI Conference on Web and Social Media (2018)

12. Fasoli, F., Maass, A., Carnaghi, A.: Labelling and discrimination: do homophobic epithets undermine fair distribution of resources? Br. J. Soc. Psychol. **54**(2), 383–393 (2015)

13. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 85 (2018)

14. Fortuna, P., da Silva, J.R., Wanner, L., Nunes, S., et al.: A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 94–104 (2019)

15. Founta, A.M., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Twelfth International AAAI Conference on Web and Social Media (2018)

16. Gagliardone, I., Gal, D., Alves, T., Martinez, G.: Countering Online Hate Speech. Unesco Publishing (2015)

17. de Gibert, O., Perez, N., Pablos, A.G., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 11–20 (2018)

18. Greenberg, J., Pyszczynski, T.: The effect of an overheard ethnic slur on evaluations of the target: how to spread a social disease. J. Exper. Soc. Psychol. **21**(1), 61–72 (1985)

19. Guermazi, R., Hammami, M., Hamadou, A.B.: Using a semi-automatic keyword dictionary for improving violent web site filtering. In: 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, pp. 337–344. IEEE (2007)

20. Huang, X., Xing, L., Dernoncourt, F., Paul, M.J.: Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. arXiv preprint arXiv:2002.10361 (2020)

21. Ibrohim, M.O., Budi, I.: Multi-label hate speech and abusive language detection in indonesian twitter. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57 (2019)

22. Mathew, B., Dutt, R., Goyal, P., Mukherjee, A.: Spread of hate speech in online social media. In: Proceedings of the 10th ACM Conference on Web Science, pp. 173–182 (2019)

23. Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., Mukherjee, A.: Temporal effects of unmoderated hate speech in gab. arXiv preprint arXiv:1909.10966 (2019)

24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

25. Mulki, H., Haddad, H., Ali, C.B., Alshabani, H.: L-hsab: A levantine twitter dataset for hate speech and abusive language. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 111–118 (2019)

26. Mullen, B., Rice, D.R.: Ethnophaulisms and exclusion: the behavioral consequences of cognitive representation of ethnic immigrant groups. Personal. Soc. Psychol. Bull. **29**(8), 1056–1067 (2003)

27. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y.: Multilingual and multi-aspect hate speech analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4667–4676 (2019)

28. Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. Sensors (Basel, Switzerland), **19**(21) (2019)

29. Ptaszynski, M., Pieciukiewicz, A., Dybała, P.: Results of the poleval 2019 shared task 6: first dataset and open shared task for automatic cyberbullying detection in polish twitter. In: Proceedings of the PolEval2019Workshop, p. 89 (2019)

30. Ribeiro, M.H., Calais, P.H., Santos, Y.A., Almeida, V.A., Meira Jr, W.: Characterizing and detecting hateful users on twitter. In: Twelfth International AAAI Conference on Web and Social Media (2018)

31. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

32. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv preprint arXiv:1701.08118 (2017)

33. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M.: An Italian twitter corpus of hate speech against immigrants. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)

34. Singhal, P., Bhattacharyya, P.: Borrow a little from your rich cousin: using embeddings and polarities of English words for multilingual sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3053–3062 (2016)

35. Soral, W., Bilewicz, M., Winiewski, M.: Exposure to hate speech increases prejudice through desensitization. Aggressive Behav. **44**(2), 136–146 (2018)

36. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93 (2016)

37. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-GRU based deep neural network. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 745–760. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_48