ECML PKDD 2020

CNeRG

A DEEP DIVE INTO MULTILINGUAL HATESPEECH CLASSIFICATION




Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, Animesh Mukherjee

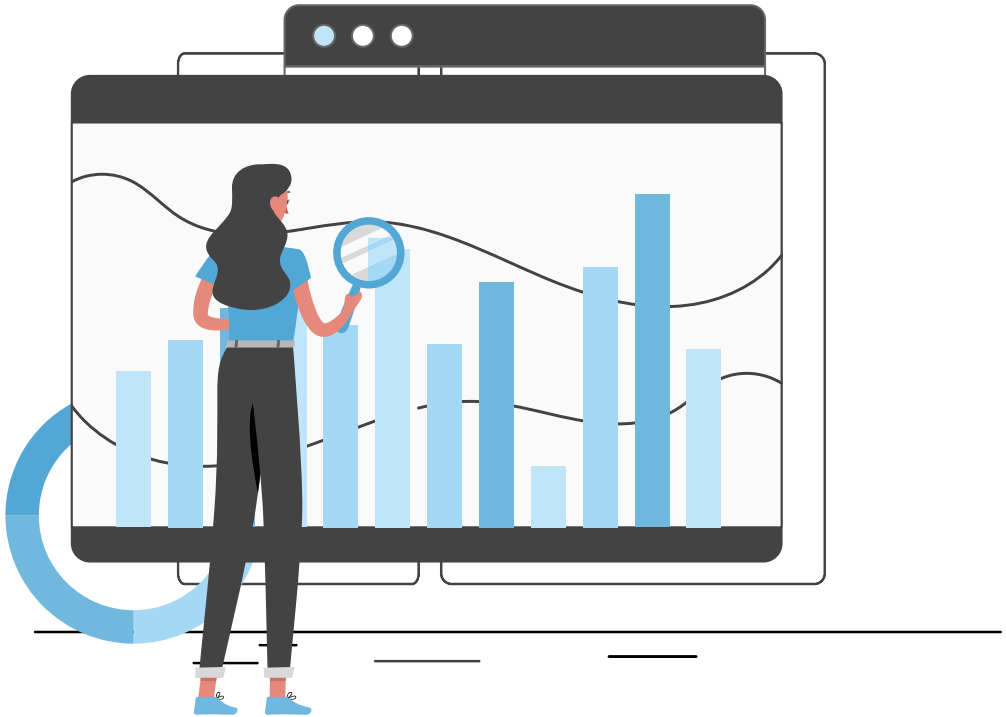
Department of Computer Science and Engineering, IIT Kharagpur, India

WARNING!

The following presentation contains words or phrases that are often considered as offensive and hateful by many.

OVERVIEW

Brief description of our work



HATESPEECH AND ITS HAZARDS

➔ Hate speech is defined as a "*direct and serious attack on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease*"

TEXT	Hatespeech?
I f**king hate ni**ers!	Yes
Jews are the worst people on earth and we should get rid of them.	Yes
Mexicans are f**king great people!	No

➔ Crimes related to hatespeech (e.g., Rohingya genocide, Pittsburgh shooting, etc.) are increasing in number

WORK DESCRIPTION

- First large scale analysis of multilingual hatespeech
- **Languages** - 9 languages: Arabic, English, German, Indonesian, Italian, Polish, Portuguese, Spanish and French
- **Models Used**
  a. MUSE $_{\text{Embeddings}}$ + CNN - GRU
  b. Translation + BERT
  c. LASER $_{\text{Embeddings}}$ + LR
  d. mBERT (Multilingual BERT)
- **Different Scenarios** - Monolingual and Multilingual settings, considered low and high resource cases

OVERALL RESULTS - TOWARD A BENCHMARK

➔ Increasing data ⇒ Better performance.

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

➔ What signals the models pick? LASER + LR → Keyword based | mBERT → Context based

FURTHER DETAILS

Github Link:
https://github.com/punyajoy/DE-LIMIT

HuggingFace:
https://huggingface.co/Hate-speech-CNERG

SCAN ME!

DETAILED DESCRIPTION

EFFECTS OF HATE SPEECH

- ➔ Hate speech is increasingly becoming a concerning issue in several countries.
- ➔ The public expression of hate speech promotes the devaluation of minority members.
- ➔ Frequent and repetitive exposure to hate speech could increase an individual's outgroup prejudice

Pittsburg Shooting

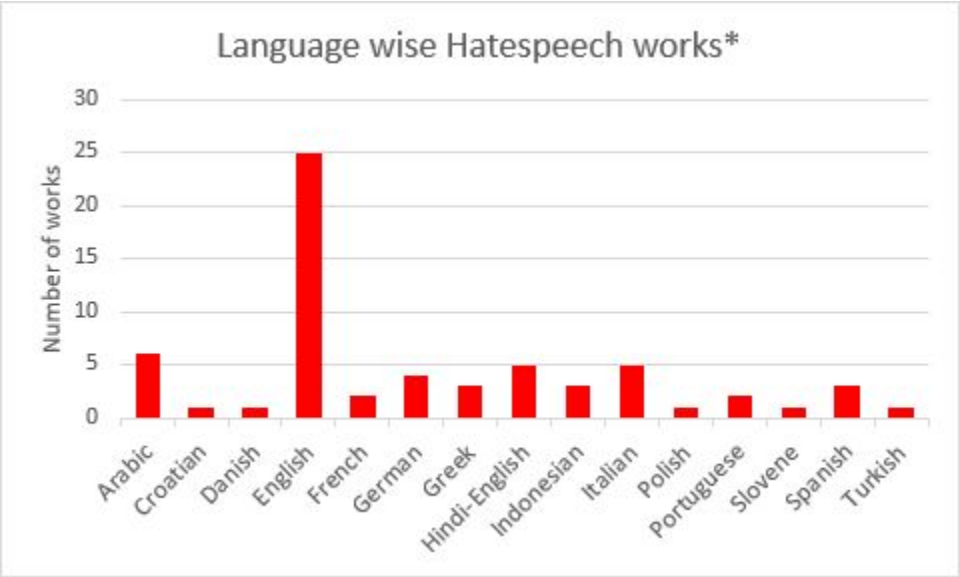
Christchurch Shooting

Rohingya Genocide

Sri Lanka riot

MOTIVATION - WHY IS THIS NECESSARY?

➔ One of the current issues - majority of the hatespeech works are available in English language only.



** - based on data from hatespeechdata.com*

RELATED WORKS

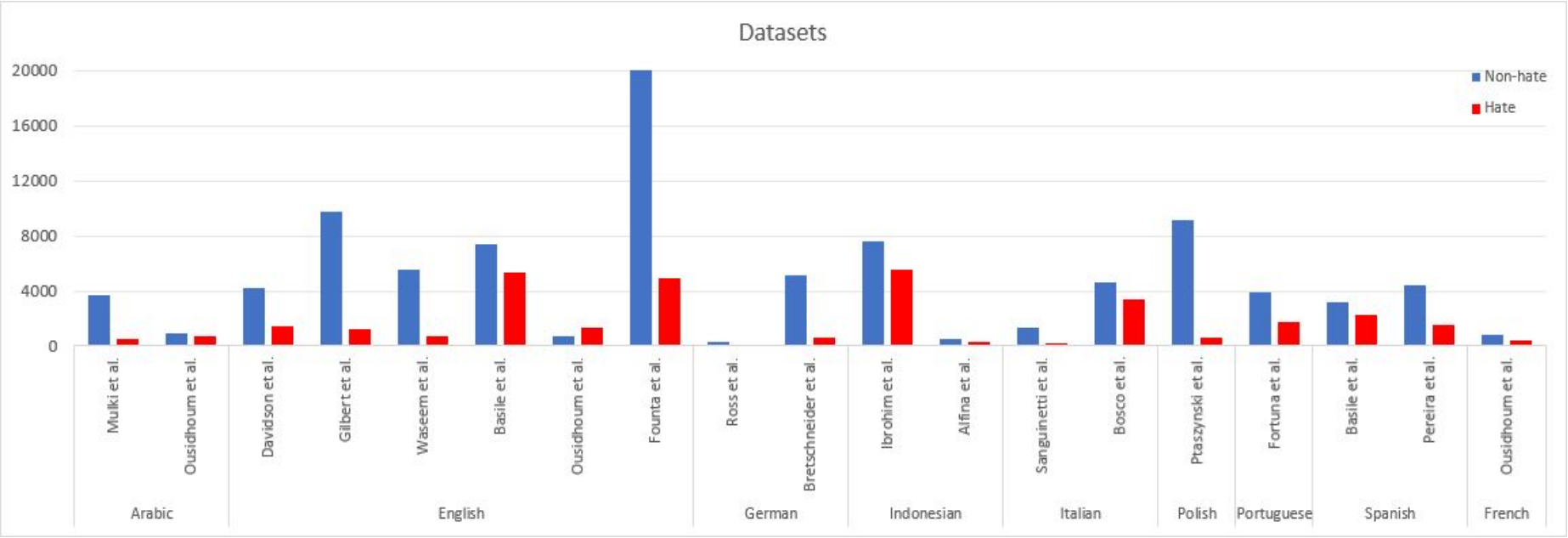
- ➔ The earlier efforts to build hate speech classifiers used simple methods such as dictionary look up, bag-of-words, etc.
- ➔ Recently, complex classification models using deep learning and graph embedding techniques have become popular.
- ➔ Zhang et al.[1] used deep neural network, combining convolutional and gated recurrent networks to improve the results on 6 out of 7 datasets used.
- ➔ Research into the multilingual hate speech is relatively new. Some works such as Corazza et al.[2] have studied hate speech in 3 languages.

[1] Zhang, Ziqi, David Robinson, and Jonathan Tepper. "Detecting hate speech on twitter using a convolution-gru based deep neural network." In *European semantic web conference*, pp. 745-760. Springer, Cham, 2018.
[2] Corazza, Michele, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. "A multilingual evaluation for online hate speech detection." *ACM Transactions on Internet Technology (TOIT)* 20, no. 2 (2020): 1-22.

DATASET DESCRIPTION

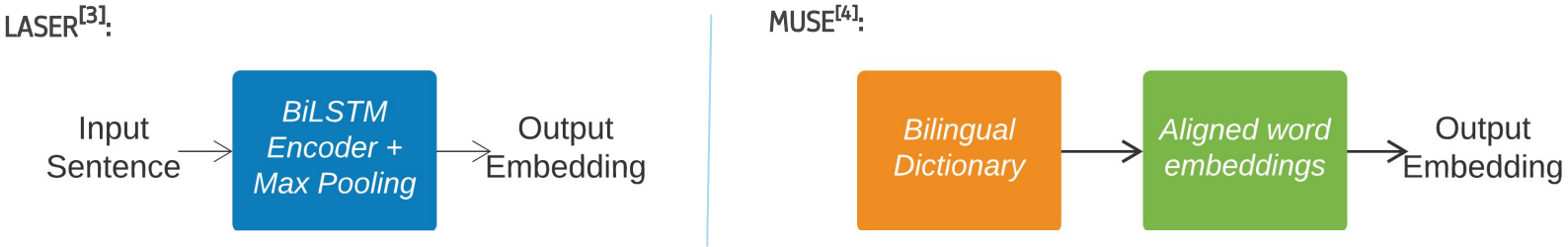
MAJORITY DATASET IN ENGLISH

DATASET IMBALANCE IN MOST CASES



EXPERIMENTAL SETUP

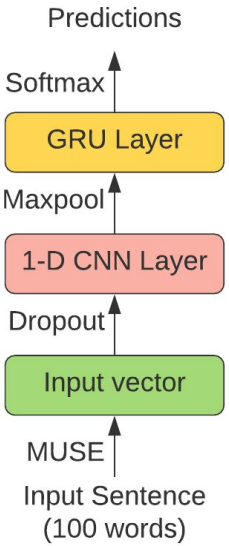
- ➔ For each language, we combine all the datasets and perform stratified train/ validation/ test split in the ratio 70%/10%/20%. We report macro F1-score to measure the performance.
- ➔ For sentences, LASER embeddings were used and for words MUSE embeddings were used to generate the multilingual representation of the corpus.



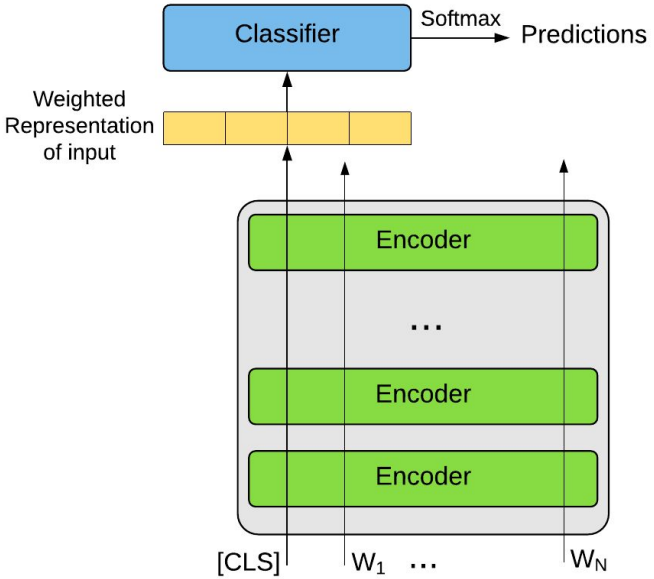
[3] Mikel Artetxe and Holger Schwenk. “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond”. In: Transactions of the Association for Computational Linguistics 7 (2019), pp. 597–610.
[4] Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word translation without parallel data." In *International Conference on Learning Representations*. 2018.

MODEL ARCHITECTURES USED

CNN - GRU:



BERT & mBERT:



EXPERIMENTS

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.

Training

Validation & Testing

RESULTS - MONOLINGUAL

Language L

Same language L

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	0.5551	0.5448	0.7017	0.8115
	LASER + LR	**0.5533**	**0.6755**	**0.7304**	**0.7488**	**0.7698**	0.7920
	mBERT	0.4588	0.4533	0.4408	0.6486	0.7295	**0.8320**
English	MUSE + CNN-GRU	0.4580	0.4594	0.4653	0.4646	0.4813	0.6441
	BERT	0.4071	0.3925	0.4260	0.4720	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	0.7101
German	MUSE + CNN-GRU	0.4708	0.4708	0.4708	0.4708	0.4762	0.5756
	Translation + BERT	0.4812	0.4758	0.4719	0.4729	0.4724	**0.7662**
	LASER + LR	0.4974	**0.5201**	**0.5465**	**0.5925**	**0.6488**	0.6873
	mBERT	**0.5037**	0.4750	0.4708	0.4717	0.5022	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**	0.7872
	mBERT	0.5106	0.5219	0.5414	0.6016	0.6530	**0.8119**
Italian	MUSE + CNN-GRU	0.4055	0.4476	0.4461	0.5206	0.5965	0.7349
	Translation + BERT	0.5006	0.5943	0.6215	0.6678	0.6919	0.7922
	LASER + LR	0.5688	**0.6210**	**0.6843**	**0.7175**	**0.7347**	0.7996
	mBERT	**0.5774**	0.4567	0.5834	0.6664	0.7026	**0.8260**
Polish	MUSE + CNN-GRU	0.4842	0.4842	0.4841	0.4842	0.5180	0.6337
	Translation + BERT	0.4842	0.4853	0.4842	0.4842	0.5066	**0.7161**
	LASER + LR	**0.4889**	**0.4879**	**0.5360**	**0.5739**	**0.6172**	0.6439
	mBERT	0.4829	0.4847	0.4842	0.4842	0.4842	0.7069
Portuguese	MUSE + CNN-GRU	0.4480	0.3807	0.4184	0.4228	0.4562	0.6100
	Translation + BERT	0.4532	0.4893	0.4712	0.5102	0.5994	0.6935
	LASER + LR	**0.5194**	**0.5536**	**0.6070**	**0.6210**	**0.6412**	**0.6941**
	mBERT	0.5154	0.4245	0.4148	0.5493	0.5745	0.6713
Spanish	MUSE + CNN-GRU	0.4382	0.3354	0.3558	0.4203	0.4995	0.6364
	Translation + BERT	0.4598	0.4722	0.5080	0.4576	0.6035	0.7237
	LASER + LR	**0.5168**	**0.5434**	**0.5521**	**0.5938**	**0.6153**	0.6997
	mBERT	0.4395	0.4285	0.4048	0.4861	0.5999	**0.7329**
French	MUSE + CNN-GRU	0.4878	0.4683	0.5008	0.5222	0.5250	0.5619
	Translation + BERT	0.4173	0.4260	0.4429	0.4749	0.6037	**0.6595**
	LASER + LR	**0.5058**	**0.5486**	**0.6136**	**0.6302**	**0.6085**	0.6172
	mBERT	0.4818	0.4139	0.4053	0.4355	0.5701	0.6165

Training

Validation & Testing

Language L

Same language L

RESULTS - MONOLINGUAL

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	0.5551	0.5448	0.7017	0.8115
	LASER + LR	**0.5533**	**0.6755**	**0.7304**	**0.7488**	**0.7698**	0.7920
	mBERT	0.4588	0.4533	0.4408	0.6486	0.7295	**0.8320**
English	MUSE + CNN-GRU	0.4580	0.4594	0.4653	0.4646	0.4813	0.6441
	BERT	0.4071	0.3925	0.4260	0.4720	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	0.7101
German	MUSE + CNN-GRU	0.4708	0.4708	0.4708	0.4708	0.4762	0.5756
	Translation + BERT	0.4812	0.4758	0.4719	0.4729	0.4724	**0.7662**
	LASER + LR	0.4974	**0.5201**	**0.5465**	**0.5925**	**0.6488**	0.6873
	mBERT	**0.5037**	0.4750	0.4708	0.4717	0.5022	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**	0.7872
	mBERT	0.5106	0.5219	0.5414	0.6016	0.6530	**0.8119**
Italian	MUSE + CNN-GRU	0.4055	0.4476	0.4461	0.5206	0.5965	0.7349
	Translation + BERT	0.5006	0.5943	0.6215	0.6678	0.6919	0.7922
	LASER + LR	0.5688	**0.6210**	**0.6843**	**0.7175**	**0.7347**	0.7996
	mBERT	**0.5774**	0.4567	0.5834	0.6664	0.7026	**0.8260**
Polish	MUSE + CNN-GRU	0.4842	0.4842	0.4841	0.4842	0.5180	0.6337
	Translation + BERT	0.4842	0.4853	0.4842	0.4842	0.5066	**0.7161**
	LASER + LR	**0.4889**	**0.4879**	**0.5360**	**0.5739**	**0.6172**	0.6439
	mBERT	0.4829	0.4847	0.4842	0.4842	0.4842	0.7069
Portuguese	MUSE + CNN-GRU	0.4480	0.3807	0.4184	0.4228	0.4562	0.6100
	Translation + BERT	0.4532	0.4893	0.4712	0.5102	0.5994	0.6935
	LASER + LR	**0.5194**	**0.5536**	**0.6070**	**0.6210**	**0.6412**	**0.6941**
	mBERT	0.5154	0.4245	0.4148	0.5493	0.5745	0.6713
Spanish	MUSE + CNN-GRU	0.4382	0.3354	0.3558	0.4203	0.4995	0.6364
	Translation + BERT	0.4598	0.4722	0.5080	0.4576	0.6035	0.7237
	LASER + LR	**0.5168**	**0.5434**	**0.5521**	**0.5938**	**0.6153**	0.6997
	mBERT	0.4395	0.4285	0.4048	0.4861	0.5999	**0.7329**
French	MUSE + CNN-GRU	0.4878	0.4683	0.5008	0.5222	0.5250	0.5619
	Translation + BERT	0.4173	0.4260	0.4429	0.4749	0.6037	**0.6595**
	LASER + LR	**0.5058**	**0.5486**	**0.6136**	**0.6302**	**0.6085**	0.6172
	mBERT	0.4818	0.4139	0.4053	0.4355	0.5701	0.6165

Training

Validation & Testing

RESULTS - MONOLINGUAL

Language L

Same language L

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	0.5551	0.5448	0.7017	0.8115
	LASER + LR	**0.5533**	**0.6755**	**0.7304**	**0.7488**	**0.7698**	0.7920
	mBERT	0.4588	0.4533	0.4408	0.6486	0.7295	**0.8320**
English	MUSE + CNN-GRU	0.4580	0.4594	0.4653	0.4646	0.4813	0.6441
	BERT	0.4071	0.3925	0.4260	0.4720	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	0.7101
German	MUSE + CNN-GRU	0.4708	0.4708	0.4708	0.4708	0.4762	0.5756
	Translation + BERT	0.4812	0.4758	0.4719	0.4729	0.4724	**0.7662**
	LASER + LR	0.4974	**0.5201**	**0.5465**	**0.5925**	**0.6488**	0.6873
	mBERT	**0.5037**	0.4750	0.4708	0.4717	0.5022	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**	0.7872
	mBERT	0.5106	0.5219	0.5414	0.6016	0.6530	**0.8119**
Italian	MUSE + CNN-GRU	0.4055	0.4476	0.4461	0.5206	0.5965	0.7349
	Translation + BERT	0.5006	0.5943	0.6215	0.6678	0.6919	0.7922
	LASER + LR	0.5688	**0.6210**	**0.6843**	**0.7175**	**0.7347**	0.7996
	mBERT	**0.5774**	0.4567	0.5834	0.6664	0.7026	**0.8260**
Polish	MUSE + CNN-GRU	0.4842	0.4842	0.4841	0.4842	0.5180	0.6337
	Translation + BERT	0.4842	0.4853	0.4842	0.4842	0.5066	**0.7161**
	LASER + LR	**0.4889**	**0.4879**	**0.5360**	**0.5739**	**0.6172**	0.6439
	mBERT	0.4829	0.4847	0.4842	0.4842	0.4842	0.7069
Portuguese	MUSE + CNN-GRU	0.4480	0.3807	0.4184	0.4228	0.4562	0.6100
	Translation + BERT	0.4532	0.4893	0.4712	0.5102	0.5994	0.6935
	LASER + LR	**0.5194**	**0.5536**	**0.6070**	**0.6210**	**0.6412**	**0.6941**
	mBERT	0.5154	0.4245	0.4148	0.5493	0.5745	0.6713
Spanish	MUSE + CNN-GRU	0.4382	0.3354	0.3558	0.4203	0.4995	0.6364
	Translation + BERT	0.4598	0.4722	0.5080	0.4576	0.6035	0.7237
	LASER + LR	**0.5168**	**0.5434**	**0.5521**	**0.5938**	**0.6153**	0.6997
	mBERT	0.4395	0.4285	0.4048	0.4861	0.5999	**0.7329**
French	MUSE + CNN-GRU	0.4878	0.4683	0.5008	0.5222	0.5250	0.5619
	Translation + BERT	0.4173	0.4260	0.4429	0.4749	0.6037	**0.6595**
	LASER + LR	**0.5058**	**0.5486**	**0.6136**	**0.6302**	**0.6085**	0.6172
	mBERT	0.4818	0.4139	0.4053	0.4355	0.5701	0.6165

Training

Validation & Testing

RESULTS - MONOLINGUAL

Language L

Same language L

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	0.5551	0.5448	0.7017	0.8115
	LASER + LR	**0.5533**	**0.6755**	**0.7304**	**0.7488**	**0.7698**	0.7920
	mBERT	0.4588	0.4533	0.4408	0.6486	0.7295	**0.8320**
English	MUSE + CNN-GRU	0.4580	0.4594	0.4653	0.4646	0.4813	0.6441
	BERT	0.4071	0.3925	0.4260	0.4720	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	0.7101
German	MUSE + CNN-GRU	0.4708	0.4708	0.4708	0.4708	0.4762	0.5756
	Translation + BERT	0.4812	0.4758	0.4719	0.4729	0.4724	**0.7662**
	LASER + LR	0.4974	**0.5201**	**0.5465**	**0.5925**	**0.6488**	0.6873
	mBERT	**0.5037**	0.4750	0.4708	0.4717	0.5022	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**	0.7872
	mBERT	0.5106	0.5219	0.5414	0.6016	0.6530	**0.8119**
Italian	MUSE + CNN-GRU	0.4055	0.4476	0.4461	0.5206	0.5965	0.7349
	Translation + BERT	0.5006	0.5943	0.6215	0.6678	0.6919	0.7922
	LASER + LR	0.5688	**0.6210**	**0.6843**	**0.7175**	**0.7347**	0.7996
	mBERT	**0.5774**	0.4567	0.5834	0.6664	0.7026	**0.8260**
Polish	MUSE + CNN-GRU	0.4842	0.4842	0.4841	0.4842	0.5180	0.6337
	Translation + BERT	0.4842	0.4853	0.4842	0.4842	0.5066	**0.7161**
	LASER + LR	**0.4889**	**0.4879**	**0.5360**	**0.5739**	**0.6172**	0.6439
	mBERT	0.4829	0.4847	0.4842	0.4842	0.4842	0.7069
Portuguese	MUSE + CNN-GRU	0.4480	0.3807	0.4184	0.4228	0.4562	0.6100
	Translation + BERT	0.4532	0.4893	0.4712	0.5102	0.5994	0.6935
	LASER + LR	**0.5194**	**0.5536**	**0.6070**	**0.6210**	**0.6412**	**0.6941**
	mBERT	0.5154	0.4245	0.4148	0.5493	0.5745	0.6713
Spanish	MUSE + CNN-GRU	0.4382	0.3354	0.3558	0.4203	0.4995	0.6364
	Translation + BERT	0.4598	0.4722	0.5080	0.4576	0.6035	0.7237
	LASER + LR	**0.5168**	**0.5434**	**0.5521**	**0.5938**	**0.6153**	0.6997
	mBERT	0.4395	0.4285	0.4048	0.4861	0.5999	**0.7329**
French	MUSE + CNN-GRU	0.4878	0.4683	0.5008	0.5222	0.5250	0.5619
	Translation + BERT	0.4173	0.4260	0.4429	0.4749	0.6037	**0.6595**
	LASER + LR	**0.5058**	**0.5486**	**0.6136**	**0.6302**	**0.6085**	0.6172
	mBERT	0.4818	0.4139	0.4053	0.4355	0.5701	0.6165

mBERT

Training
Dataset from all but one language

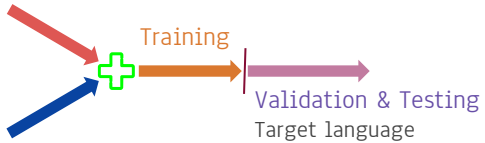
Fine-tuning
Target language dataset (incremental steps)

Validation & Testing
Target language dataset

All but one language datasets

LASER + LR

Target language dataset (incremental steps)



RESULTS - MULTILINGUAL

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**

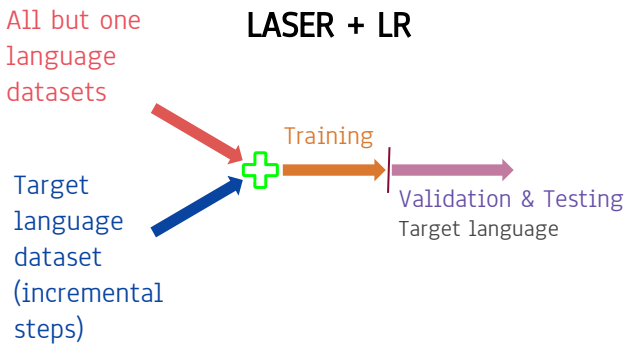
mBERT

Training
Dataset from all but one language

Fine-tuning
Target language dataset (incremental steps)

Validation & Testing
Target language dataset

LASER + LR



RESULTS - MULTILINGUAL

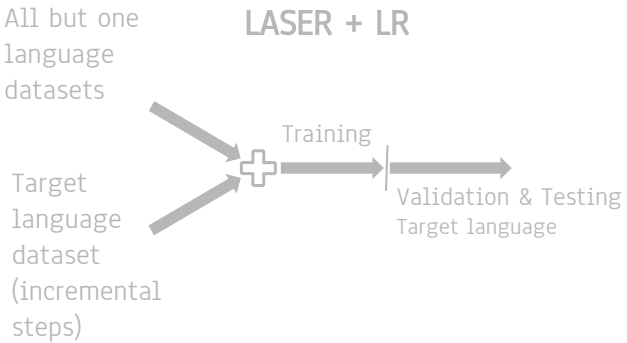
Testing Language	Model	Training Size						
		Zero shot	16	32	64	128	256	Full D
Arabic	LASER + LR	0.4645	**0.4651**	0.4664	0.4704	0.4784	0.4930	0.6751
	mBERT	**0.6442**	0.4535	**0.4738**	**0.5302**	**0.7331**	**0.7707**	**0.8365**
English	LASER + LR	**0.6050**	**0.6051**	**0.6052**	**0.6053**	**0.6054**	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**

mBERT

Training
Dataset from all but one language

Fine-tuning
Target language dataset (incremental steps)

Validation & Testing
Target language dataset



RESULTS - MULTILINGUAL

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**

HATESPEECH BENCHMARKS

Recipes for different languages and resource settings, as obtained in our experiments.

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

INTERPRETABILITY (EXAMPLES)

Interpretability analysis of LASER + LR and mBERT using LIME[5]

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” 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. 2016, pp. 1135–1144.

INTERPRETABILITY (EXAMPLES)

Interpretability analysis of LASER + LR and mBERT using LIME[5]

Yellow - LASER+LR

Green - mBERT

Sentences with hate label
das *pack* muss tag und nacht gejagt werden,ehe sie es mit den deutschen machen !! **Translation** :- the *pack* must be hunted day and night before they do it with the Germans !!

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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. 2016, pp. 1135–1144.

INTERPRETABILITY (EXAMPLES)

Interpretability analysis of LASER + LR and mBERT using LIME[5]

Yellow - LASER+LR

Green - mBERT

Sentences with hate label
das *pack* muss tag und nacht *gejagt* werden,ehe sie es mit den deutschen machen !! **Translation** :- the *pack* must be *hunted* day and night before they do it with the Germans !!

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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. 2016, pp. 1135–1144.

INTERPRETABILITY (EXAMPLES)

Interpretability analysis of LASER + LR and mBERT using LIME[5]

Yellow - LASER+LR

Green - mBERT

Sentences with hate label
das *pack* muss tag und nacht *gejagt* werden,ehe sie es mit den deutschen machen !! **Translation** :- 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 **Translation** :- absolutely! it must be boarded at sea by the navy national arrest all occupants expel migrants... @url

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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. 2016, pp. 1135–1144.

INTERPRETABILITY (EXAMPLES)

Interpretability analysis of LASER + LR and mBERT using LIME[5]

Yellow - LASER+LR

Green - mBERT

Sentences with hate label
das *pack* muss tag und nacht *gejagt* werden,ehe sie es mit den deutschen machen !! **Translation** :- 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 **Translation** :- absolutely! it must be boarded at sea by the navy national arrest all occupants expel migrants... @url

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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. 2016, pp. 1135–1144.

ERROR ANALYSIS - TYPES OF ERRORS

Annotation Dilemma (AD)

Ambiguous instances, where according to us model predicts correctly but annotators have labelled it wrong.

Confounding Factors (CF)

Errors caused when model relies on some irrelevant features like normalized mentions and links.

Hidden Context (HC)

Errors due to model failing to capture context of the post

Abusive Words (AW)

Errors caused due to over dependance of model on abusive words in input.

ERROR ANALYSIS EXAMPLES

Here "parasites" refers to immigrants

mBERT:

Sentence	GT	P	E
"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

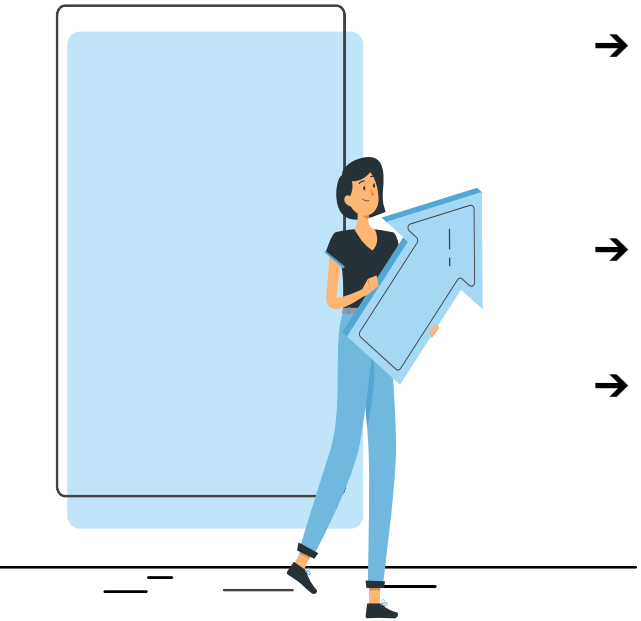
Here "retarded" is used for movie.

LASER + LR:

Sentence	GT	P	E
this movie is actually good cuz its so retarded.	1	0	AW

NOTE:- **For additional examples please check our paper**

CONCLUSION



- In this work, we have analyzed multilingual hate speech using datasets from 16 different sources, comprising of 9 different languages.
- We considered various conditions like low and high resource settings and monolingual or multilingual cases for the different languages.
- As per the observations, for low resource cases, LASER+LR is more effective while for high resource cases, mBERT is usually more effective.

THANK YOU

Github :
https://github.com/punyajoy/DE-LIMIT

HuggingFace :
https://huggingface.co/Hate-speech-CNERG

Contact us:
Sai Saketh Aluru: saisakethaluru@iitkgp.ac.in