PROJECT REPORT

Topic: Profanity, Hate and Offensive Language Detection

Team no.10

Group members:

Somana Sarath Kumar - 2022204011

Gaurang Patil - 2023701002

Ishank Kapania - 2023701012

Problem Statement

Social media is often a hotbed for spreading hate, and controlling hateful, offensive, and profane content on social media is a challenging task for social media platforms owing to their huge userbase. Recent developments in NLP have played a key role in detecting content which can cause disturbance to individuals and society. While there has been a significant amount of research on detecting hate speech in English, there has been little research on detecting hate speech in Indic languages, especially low resource languages like Marathi and Telugu. Additionally, most research work in hate speech detection in Indic languages has been limited to building models trained on relatively small datasets sourced from a single social media platform. However, the socio-economic background of users across these social media platforms often varies quite significantly. Consequently, the kind of content generated by users also varies across these platforms. Therefore, by combining multiple datasets (of a particular language) and training models on these richer datasets, we can possibly build models which can generalize better when deployed in production. We choose to work on 3 Indic languages, namely- Hindi, Marathi, and Telugu, keeping in mind the language proficiency of the team members. The dataset details have been specified on the next page.

Datasets

Lang.	*Dataset*	*Link*	*Dataset Details*		
			Approx Size	*Classes*	*Additional description*
Hindi	MACD (Multilingual Abusive Comment Detection at Scale)	https://github.com/ShareChatAI/MACD/tree/main/dataset	33.4 k	Abusive Non abusive comments	Comprises of comments from ShareChat.
	Constraint@AAAI2021 - Hostile Post Detection in Hindi	https://competitions.codalab.org/competitions/26654	8.2k	Hostile • Fake news • Hate • Offense • Defame Non hostile	Comprises of data from Twitter, Facebook and WhatsApp.
	HateCheckHIn	https://github.com/hate-alert/HateCheckHIn/blob/main/monolingual_functionalities.csv	4.7k	Hateful Non hateful	Comprises of data from Twitter.
	HASOC 2019	https://hasocfire.github.io/hasoc/2019/dataset.html	6k	A: Hate, Offensive or None B: Hate, Offensive, or Profane C: Targeted or Untargeted	Posts from Twitter and Facebook
Marathi	DeepOffense	https://github.com/TharinduDR/DeepOffen	3.1k	Offensive Non offensive	

		se/tree/master/examples/marathi/data			
	L3Cube-MahaHate	https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaHate	4-class 25000 samples: labels- hate, offensive, pofane, and none. 2-class 37500 samples: labels- hate and none. Comprises of data from Twitter.		
Telugu	MACD (Multilingual Abusive Comment Detection at Scale)	https://github.com/ShareChatAI/MACD/tree/main/dataset	30k	Abusive Non abusive	Comprises of comments from ShareChat
	Dream-T!	https://github.com/Cha14ran/DREAM-T	35k	Hate Non-hate	Comprises of data mainly from Telugu websites

Note: HASOC 2022 Marathi Task 3A dataset was to be used in addition to the ones mentioned, however when we requested for the access to the password to the CSV file, we did not receive any reply.

Baseline methods

Model	Conference	Description and justification as baseline
AbuseXLMR	Published in NeurIPS 2022.	AbuseXLMR has been trained on a well-balanced dataset sourced from ShareChat with diverse set of 70k users. Additionally, it is trained on multilingual data from 15+languages. The model is particularly pretrained over social media data unlike other models which are usually trained on large scale multilingual datasets. It is more likely to adapt well to social media nuances like spelling and grammatical mistakes. Domain adaptability is the most important reason why it has been selected as an appropriate baseline model.
QutNocturnal@ HASOC'19: CNN	HASOC19 Hindi -Winner's solution - published in CEUR Workshop Proceedings	Instead of using word embeddings trained over larger corpus from a general domain, they used a small collection of relevant tweets –random, sarcastic tweet for pretraining. The model won HASOC19 Hindi hate speech detection task and hence serves as an appropriate baseline.
MahaHateBERT	Published in ACL 2022	It is the MahaBERT model (mBERT fine tuned on various Marathi datasets) fine tuned further on MahaHate dataset, which is the largest Marathi hate speech detection dataset available so far. This makes it an appropriate baseline.
Random Forest regressor	HASOC'22 Task 3A winner's solution - published in CEUR Workshop Proceedings	The random forest regressor predicts a real value as its output and then rounds up the value to make the final prediction. It is the winner's solution for HASOC 2022 task 3A (Marathi). Hence it is an appropriate baseline for Marathi.
BERT-Te (LTRCTelugu)	Published in ACM Transactions 2022	BERT model fine-tuned on large Telugu corpus. Considering lack of resources and models for Telugu this is an appropriate choice as a baseline model.

Advanced methods

We intend to train/ fine tune the baseline models over datasets sourced from different platforms or additional datasets from same platform. Given below is the justification for the same and why we feel that it has the potential beat the results produced by the baseline models:

- **Training on larger dataset**: some of the baselines for Indic languages have been trained on relatively smaller datasets which limits the model's ability to generalize well on unseen data. Our hypothesis is that by training the models on a larger data set, we would be able to generalize better.
- **Training on data sources from multiple social media platforms:** the user base across social media platforms has a varied socio-economic background. Consequently, the comments and posts are also likely to vary to a certain extent. Therefore, by training on data sourced from different platforms, we expect that the resultant model is likely to generalize better than baseline models trained on data sourced from a single platform.
- As mentioned in the earlier documents, the following advanced method would be considered only if we have sufficient time. Handling emojis in Marathi L3Cube MahaHate dataset: Considering that the original paper claims that the emojis have been removed from the text, which may have had an impact on the accuracy. There are some universally accepted symbols of profanity, which means that including emojis as a part of the classification problem is really important. This is especially true when we require a fine-grained classification. So, one of the research papers that we had surveyed talked about how emojis could be replaced by their text equivalent. There are some libraries which can detect emojis in a string, and there are libraries which can also convert those emojis into their text equivalents. The only problem with them is that the text is in English and needs to be translated into the Indic language. Now that that text is in Marathi, we would require the textual equivalent to be translated into Marathi.

Results

Lang.	*Model*	*Results*			
		Metrics authors obtained on corresponding test set	***Metrics obtained by us on corresponding test set***	***Metrics obtained on the test data from collated dataset***	***Metrics obtained on original test set after training on collated data***
Hindi	AbuseXLMR	Accuracy: 87.96%			
	QutNocturnal CNN	Accuracy: 82.00%	Accuracy: 80.35%	Accuracy: 84.83%	Accuracy: 71.85%
Marathi	MahaHate BERT	Accuracy: 90.90%	Accuracy: 90.30%	Accuracy: 93.3%	Accuracy: 91.1%
	Random Forest	Macro F1: 0.97	*Unable to test due to unavailability of data-* Accuracy on L3cube dataset: 84.83%	Accuracy: 87.92%	Accuracy: 83.95% (L3cube dataset)
Telugu	AbuseXLMR	Accuracy: 91.40%	Accuracy 91.64% F1 Macro : 0.916	Accuracy: 94.44%, F1_macro: 0.91, F1_micro: 0.94	Accuracy: 90.45%
	BERT-Te	F1-score: 0.64	F1 Macro Score : 0.498 Accuracy : 0.98	Accuracy: 79.23%, f1_macro: 0.696 f1_micro: 0.79	Accuracy: 95.22%, F1_macro: 0.48, F1_micro: 0.95

Summary of results :

- QutNocturnal CNN, when trained over collated data, gives lesser accuracy when tested on the original test dataset.
- MahaHate BERT when trained over collated data gives higher accuracy when tested on the original dataset, however, it is comparable to the accuracy obtained by the authors in their paper.

- AbuseXLMR greatly outperforms BERT-TE on collated dataset. AbuseXLMR although performs worse on original test set, performs better on collated dataset.

Discussion and Conclusions

Challenges	How we resolved
Differences in annotation guidelines across datasets.	We decided to have a binary classification setting wherein we have two classes. For example, 'hate' and 'offensive' speech can be considered as one category. The problem with having fine grained classification was that the annotation guidelines would have come into the picture and would have decreased the overall accuracy.
Older versions of tensorflow were used by baseline code available- which meant that many functions would not run as expected.	Modified the code to suit the latest version of tensorflow.
Training models over large datasets on local machines.	Transferred our workflow to Google Colaboratory where we used GPUs to speed up the training of the models.
Hindi lemmatizer code for QutNocturnal CNN available on Github was not working as expected, possibly due to lack of maintenance in the library code.	Searched for an alternative library for Hindi lemmatization and used it as an alternative.
Marathi dataset had issues with pre-processing text.	Used MahaNLP library to remove stopwords and URLs. Wrote function to remove non-Devanagari script.

AbuseXLMR outperforms QutNocturnal CNN on collated dataset, likely due to the fact that it has been trained on a large multilingual dataset, takes context into consideration, unlike Word2Vec embeddings used in the CNN model.

Similarly, MahaHateBERT outperforms Random forest because it considers the context and also the fact that it has been fine-tuned on similar data makes it more likely to perform well.

There is some evidence suggesting that training on collated datasets gives better performance, however this is applicable mostly to contextual models and not traditional methods.

In future, AbuseXLMR can be used for training and testing on Marathi dataset considering that it has already shown good performance in few-shot setting. Additionally, there are some universally known symbols of profanity which means that the text equivalent of emojis can be used as substitute for the emoji to improve the classification accuracy. AbuseXLMR has been trained without substituting or removing emojis and it might prove effective without substitution as well.

References

- Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT Models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- *Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP Tasks in Telugu Language. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 1, Article 18 (January 2023), 34 pages. https://doi.org/10.1145/3531535*
- https://proceedings.neurips.cc/paper_files/paper/2022/hash/a7c4163b33286261b24c72fd3d1707c9-Abstract-Datasets_and_Benchmarks.html
- https://github.com/mdabashar/QutNocturnal-Hasoc2019
- *Sayani Ghosal, Amita Jain, Devendra Kumar Tayal, Varun G. Menon, and Akshi Kumar. 2023. Inculcating Context for Emoji Powered Bengali Hate Speech Detection using Extended Fuzzy SVM and Text Embedding Models. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (March 2023). https://doi.org/10.1145/3589001*
- Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.
- *Sayani Ghosal, Amita Jain, Devendra Kumar Tayal, Varun G. Menon, and Akshi Kumar. 2023. Inculcating Context for Emoji Powered Bengali Hate Speech Detection using Extended Fuzzy SVM and Text Embedding Models. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (March 2023). https://doi.org/10.1145/3589001*