RESEARCH PROJECT

PROJECT TITLE : MULTILINGUAL HATE SPEECH DETECTION

SUPERVISOR :

DR. ROHIT BENIWAL

SUBMITTED BY :

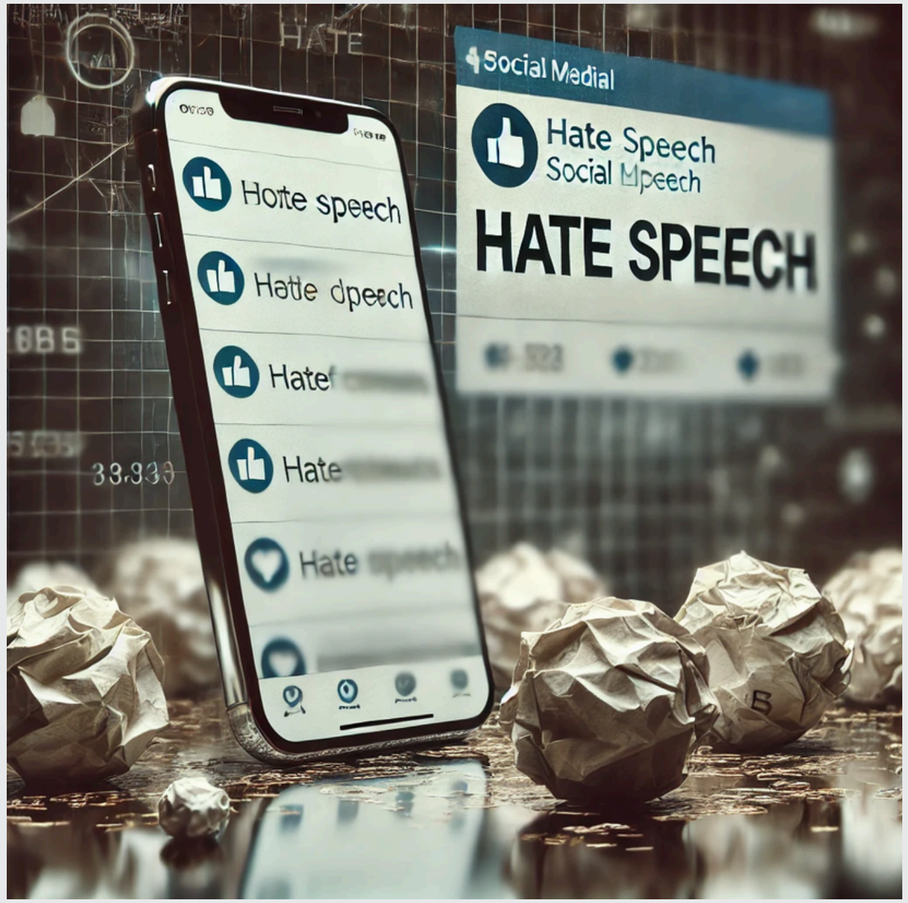
LAKSHYA (2K21/CO/258)

KARTIK NAGPAL (2K21/CO/226)

ROOPAL SHAKYA (2K21/CO/396)

INTRODUCTION

- HATE SPEECH DISRUPTS SOCIETAL HARMONY, FOSTERS DISCRIMINATION, AND INCITES VIOLENCE, ESPECIALLY ONLINE.
- MULTILINGUAL HATE SPEECH DETECTION IS ESSENTIAL FOR PROTECTING LOW-RESOURCE LANGUAGES LIKE HINDI, TAMIL, AND BENGALI.
- KEY CHALLENGES INCLUDE LINGUISTIC DIVERSITY, CODE-MIXED TEXT, AND DEMOGRAPHIC BIASES IN DETECTION MODELS.



PROBLEM STATEMENT

Let $X = \{x1, x2, \ldots, xn\}$ be the set of input texts, and $Y = \{y1, y2, \ldots, yn\}$ be the corresponding labels for each input xi, where $Y = \{Hate, Non - Hate\}$ represents the hate speech or non-hate speech, respectively. The objective of the proposed model is to predict the conditional probability of the label y for a given input x , i.e., $P(y|x)$.

KEY PROBLEMS :

- **DATA SCARCITY :** Limited annotated datasets for regional languages like Hindi, Tamil, and Bengali.
- **CODE - MIXING :** Frequent blending of multiple languages, e.g., Hinglish (Hindi-English).
- **DEMOGRAPHIC GENERALIZATION :** Models trained on specific populations fail to adapt to other cultural contexts.

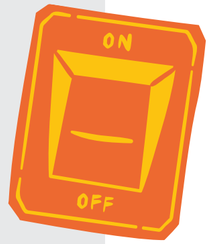
OBJECTIVE

MAIN OBJECTIVE :

Develop a robust multilingual hate speech detection model.

SUB-OBJECTIVE :

- **Development of an Attention-Based Framework for Multilingual Hate Speech Detection-** Introduces a multilingual model using attention mechanisms for hate speechdetection across diverse languages, including Hindi, Tamil, Bengali, Urdu, and Marathi.
- **Improved Handling of Code-Mixed and Low-Resource Languages-** Addresses the challenges of detecting hate speech in code-mixed and low-resource languages by incorporating multilingual embeddings and attention-driven context understanding.
- **Comprehensive Benchmarking Across Multilingual Datasets-** Provides evaluation of the proposed model on multiple language datasets.



LITERATURE REVIEW

S.No	Study	Languages	Model	Dataset	Result
1	Hate speech and offensive language detection in Dravidian languages using deep ensemble framework[8]	Malayalam and Tamil code-mixed	BERT, DNN, and MuRIL (Malayalam) DistilBERT, DNN and xlm-RoBERTa (Tamil)	[19]	F1-Score 0.802 (Malayalam) 0.933 (Tamil)
2	Hate speech detection on Twitter using transfer learning[9]	Urdu	DistilBERT	Not available	F1-score 0.69
3	Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach.[10]	Hinglish	Interpretation(translation and transliteration), mBERT, and Deep neural network	[20]	F1-score Hate (%)-0.56 Nonhate (%)-0.78
4	Investigating Hostile Post Detection in Hindi[11]	Hindi	MuRIL and XLM-RoBERTa	[21],[22]	Coarse Grained F1-sore 0.9716
5	Hate Speech Detection in Hindi [12]	Hindi	MuRIL.	[23]	F1-Score MuRIL: 0.73
6	Combining multiple pre-trained models for hate speech detection in Bengali, Marathi, and Hindi [13]	Bengali, Marathi, Hindi	Ensemble of mBERT and IndicBERT	[24], [25], [26]	F1-Score 0.923 (Bengali) 0.815(Marathi) 0.924(Hindi)
7	An empirical comparison of Hindi-BERT and MuRIL for hate speech detection on social media platforms in Hindi language[14]	Hindi	Hindi-BERT	[27]	Accuracy Hindi-BERT: 82.73 MuRIL: 75.9%
8	Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models[15]	Hindi- English Code mixed	CNN 1D, LSTM, BiLSTM	[28]	Accuracy CNN-1D: 82.62%, LSTM: 80.21%, BiLSTM: 81.48%

KEY FINDINGS :

- Advanced models like BERT, MuRIL, and XLM-RoBERTa are commonly used for hate speech detection.
- Datasets include Hindi-English code-mixed text, Dravidian languages, and Bengali datasets.
- Limitations in prior work:
  - Poor handling of code-mixed languages.
  - Limited datasets for low-resource languages.
  - Challenges with demographic generalization.

CHALLENGED HIGHLIGHTED :

- Annotation biases and cultural variations.
- Overlap of non-hate and mild hate categories leading to misclassification.

DATA COLLECTION

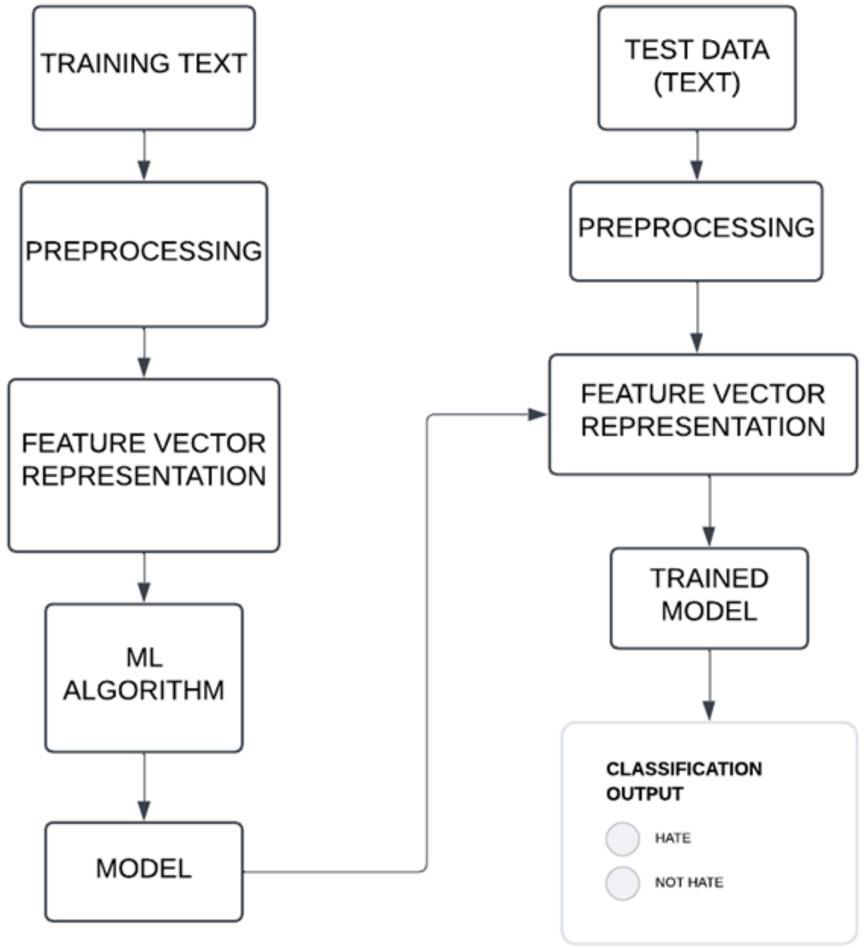
Dataset	Language	Total size	Hate speech size	Non-hate speech size
A Dataset of Hindi–English Code-Mixed Social Media Text for Hate Speech Detection[31]	Hindi-English Code mixed	4575	1661	2914
BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts [32]	Bengali	50,281	24,156	26,125
HASOC-Dravidian-CodeMix [33]	Malayalam–English	5000	2465	2535
HASOC-Dravidian-CodeMix [34]	Tamil–English	5000	2455	2485
HASOC 2019 (Hindi)[34]	Hindi-English Code mixed	4665	2469	2196
L3Cube-MahaHate (Marathi)[35]	Marathi	12500	6250	6250

Datasets from multiple languages were gathered to ensure balanced representation in hate speech detection.

- **Hindi-English :** 4,575 samples (1,661 hate speech, 2,914 non-hate).
- **Bengali :** 50,281 samples (~50% hate speech).
- **Malayalam-English & Tamil-English :** 5,000 samples each with a mix of categories.
- **Marathi :** 12,500 samples (50% hate speech).

PROPOSED MODEL STEPS:

- **Data Collection**
- **Text Preprocessing**
- **Feature Extraction**
    1. mBERT
    2. XLM-RoBERTa
    3. IndicBERT
    4. MuRIL
- **BiLSTM Layer:** Capture contextual dependencies (both forward and backward).
- **Attention Layer:** Enhance focus on contextually significant words.



TEXT PRE-PROCESSING

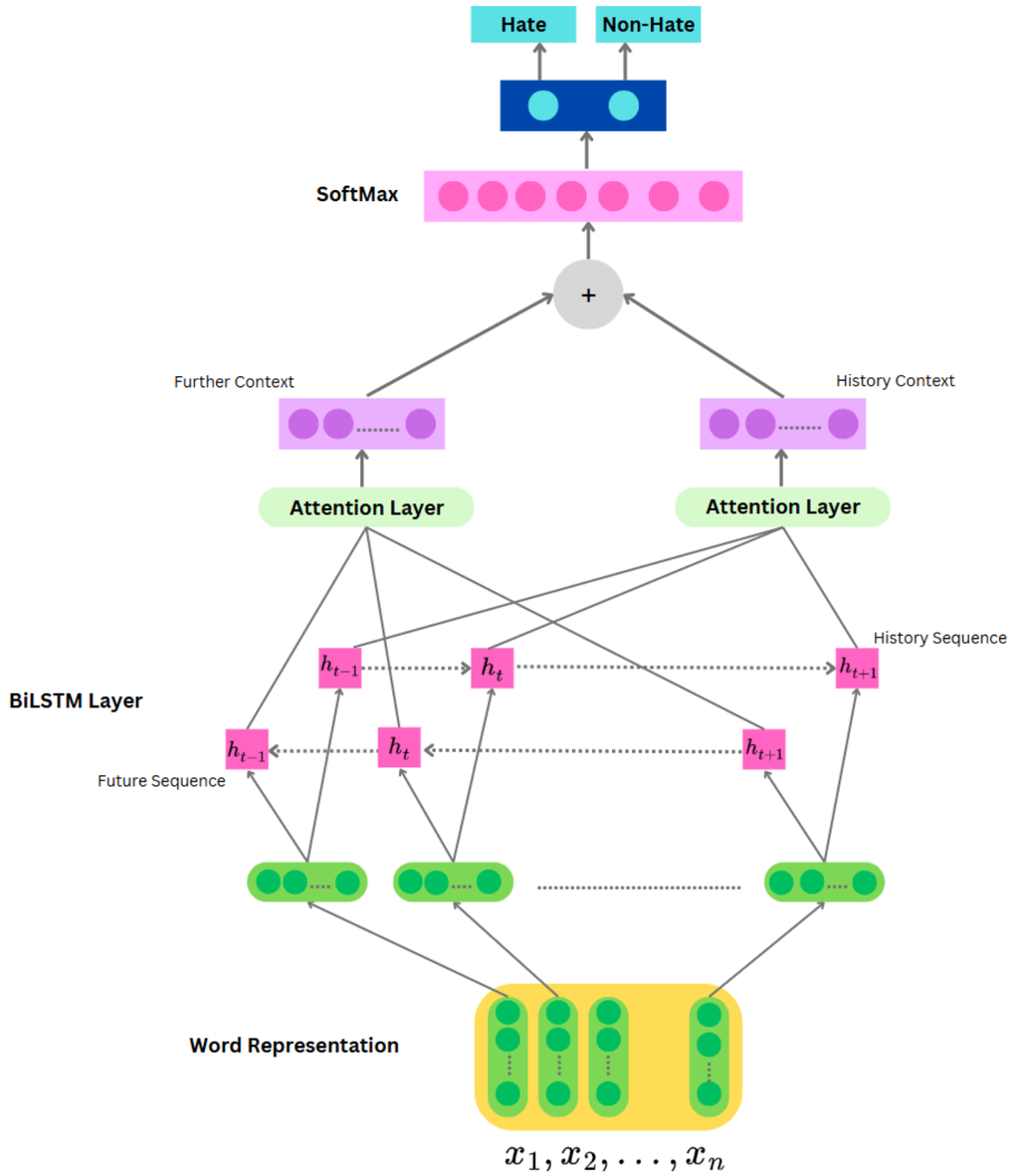
- **Noise Removal :** Eliminate URLs, emojis, hashtags, special characters, and punctuation.
- **Lowercasing**
- **Tokenization :** Using mBERT/XLM-RoBERTa tokenizer.
- **Stopword Removal**
- **Stemming & Lemmatization**

FEATURE EXTRACTION

Following word embeddings models will be explored in the study:

- **mBERT :** Supports 104 languages with 12 layers, 768 hidden dimensions, and ~110M parameters.
- **XLM-RoBERTa :** Optimized for 100 languages. Variants include Base (125M parameters) and Large (355M parameters).
- **IndicBERT :** Lightweight model for 12 Indian languages, based on ALBERT, with 12M parameters.
- **MuRIL :** Pretrained on Indian languages ,12 layers, 768 hidden dimensions, and ~110M parameters.

BLOCK DIAGRAM OF THE PROPOSED MODEL



BiLSTM Layer

Purpose : Captures both forward and backward dependencies in the text sequence.

Input : Feature representation of text ($X = \{x1, x2, \dots, xn\}$), where each $xi \in \mathbb{R}d$ is a word embedding.

Mechanism :

- **Forward LSTM :** Processes the sequence from start to end.
- **Backward LSTM :** Processes the sequence in reverse.
- **Hidden states at time t :**
  - **Forward :** $ht \rightarrow \quad LSTMforward(x_t, \overrightarrow{h_{t-1}})$
  - **Backward :** $ht \leftarrow LSTMbackward(x_t, \overleftarrow{h_{t+1}})$

Final state : Concatenation $ht = [ht \rightarrow, ht \leftarrow]$.

Output : Sequence of context-aware embeddings ($H \in \mathbb{R}n \times 2d$) capturing past and future context for each word.

ATTENATION LAYER

Purpose : Highlights important words (e.g., hate speech indicators) by dynamically weighting their relevance.

Process :

- **Hidden Representation:**
  - **Forward :** $U_f^{\rightarrow} = \tanh(\mathrm{W}h_f^{\rightarrow} + \mathrm{b})$
  - **Backward :** Similar computation for $hb \leftarrow$.
- **Attention Weights :** Calculate importance using a softmax over the similarity between u and a context vector v.
  - **Forward :** $$a_f^{\rightarrow} = \frac{exp\left(u_f^{\rightarrow} \cdot v_f^{\rightarrow}\right)}{\sum_{i=1}^{M} exp(u_f^{\rightarrow} \cdot v_f^{\rightarrow})}$$
  - **Backward :** Similar computation for $$a_b^{\leftarrow} = \frac{exp\left(u_f^{\leftarrow} \cdot v_f^{\leftarrow}\right)}{\sum_{i=1}^{M} exp(u_f^{\leftarrow} \cdot v_f^{\leftarrow})}$$
- **Context Representation :**
  - **Forward :** Weighted sum $Fc = \Sigma af \rightarrow i * hf \rightarrow i$.
  - **Backward :** Weighted sum $Hc = \Sigma ab \leftarrow i * hb \leftarrow i$.
- **Final Representation :** Concatenate forward and backward contexts: S = [Fc, Hc].

Output : Attention-enhanced representations are passed through a dropout layer and classified using a softmax layer as ‘Hate’ or ‘Non-Hate’.

RESULT ANALYSIS

Model Name	Accuracy	Precision	Recall	F1-Score
mBert	0.7678	0.5900	0.5804	0.5851
XLM-RoBERTa	0.7700	0.5804	0.5934	0.5868
mBert +LSTM	0.7560	0.6040	0.6120	0.6080
mBert+BiLSTM	0.7760	0.6190	0.5950	0.6070
XLM-RoBERTa+LSTM	0.7890	0.6200	0.6230	0.6215
Proposed Method	0.8000	0.6296	0.6296	0.6296

EVALUATION METRICS

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

Experimental Setup

- **Dataset :** Evaluated on the Hindi dataset.
- **Preprocessing :** Tokenized and padded to 128 tokens.
- **Model Configuration :**
  - Word Embeddings size: 768
  - BiLSTM with 128 hidden units and dropout.
  - **Learning rate :** 0.001 .
  - **Batch size :** 16, epochs: 50.
  - **Optimizer :** Adam.
  - **Loss**: Binary Cross-Entropy Loss

REFERENCES

[8] Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Comput. Speech Lang. 75, C (Sep 2022). https://doi.org/10.1016/j.csl.2022.101386

[9] Raza Ali, Umar Farooq, Umair Arshad, Waseem Shahzad, Mirza Omer Beg, Hate speech detection on Twitter using transfer learning, Computer Speech & Language, Volume 74, 2022, 101365, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2022.101365.

[10] Biradar, S., Saumya, S. & chauhan, A. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach.. Soc. Netw. Anal. Min. 12, 87 (2022). https://doi.org/10.1007/s13278-022-00920-w

[11] Varad Bhatnagar, Prince Kumar, Pushpak Bhattacharyya, Investigating Hostile Post Detection in Hindi, Neurocomputing, Volume 474, 2022, Pages 60-81, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2021.11.096.

[12] Bansod, Pranjali Prakash, "Hate Speech Detection in Hindi" (2023). Master's Projects. 1265.DOI:https://doi.org/10.31979/etd.yc74-7qas https://scholarworks.sjsu.edu/etd_projects/1265

[13] Nandi, A., Sarkar, K., Mallick, A. et al. Combining multiple pre-trained models for hate speech detection in Bengali, Marathi, and Hindi. Multimed Tools Appl 83, 77733–77757 (2024). https://doi.org/10.1007/s11042-023-17934-x

[14] Rakshit, Shayoni and Dhawan, Himani and Gupta, Tanya and Narula, Rachna, An empirical comparison of Hindi-BERT and MuRIL for hate speech detection on social media platforms in Hindi language (July 29, 2024). Available at SSRN: https://ssrn.com/abstract=4909420 or http://dx.doi.org/10.2139/ssrn.4909420

[15] Kamble, S., & Joshi, A. (2018). Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. ArXiv, abs/1811.05145.

[16] T. Y.S.S. Santosh and K. V.S. Aravind. 2019. Hate Speech Detection in Hindi-English Code-Mixed Social Media Text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CODS-COMAD '19). Association for Computing Machinery, New York, NY, USA, 310–313. https://doi.org/10.1145/3297001.3297048

[17] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting Offensive Tweets in Hindi-English Code-Switched Language. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

[18] K Sreelakshmi, B Premjith, K.P. Soman, Detection of Hate Speech Text in Hindi-English Code-mixed Data, Procedia Computer Science, Volume 171, 2020, Pages 737-744, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.080. (https://www.sciencedirect.com/science/article/pii/S1877050920310498)

[19] https://dravidian-codemix.github.io/HASOC-2021/index.html

[20] https://github.com/deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text

[21] M. Bhardwaj, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Hostility detection dataset in hindi (2020). arXiv:2011.03588.

[22] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. PYKL, A. Das, A. Ekbal, S. Akhtar, T. Chakraborty, Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, in: Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT), Springer, 2021.

[23] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, et al., ''Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages,'' arXiv preprint arXiv:2112.09301, 2021.

[24] https://www.kaggle.com/datasets/naurosromim/bdshs

[25] https://hasocfire.github.io/hasoc/2019/dataset.html

[26] https://github.com/l3cube-pune/MarathiNLP

[27] https://constraint-shared-task-2021.github.io/

[28] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

[29] https://hasocfire.github.io/hasoc/

[30] Mathur, Puneet, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. (2018) "Did you offend me? classification of offensive tweets in hinglish language." Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) :138-148

[31] https://github.com/punyajoy/HateSpeech-Hindi-Eng

[32] https://github.com/naurosromim/hate-speech-dataset-for-Bengali-social-media

[33] https://sites.google.com/view/dravidian-codemix-fire2020/overview

[34] https://hasocfire.github.io/hasoc/2019/dataset.html

[35] https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaHate

THANK YOU