WebSci'22
BARCELONA

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Hostility Detection in Online Hindi-English Code-Mixed Conversations

14th ACM Web Science Conference (ACM WebSci'22)

26-29, June, 2022

Barcelona, Spain

Authors

Aditi Bagora , Kamal Shrestha, Kaushal Maurya Maunendra Sankar Desarkar

Indian Institute of Technology Hyderabad (IITH),
Hyderabad, India

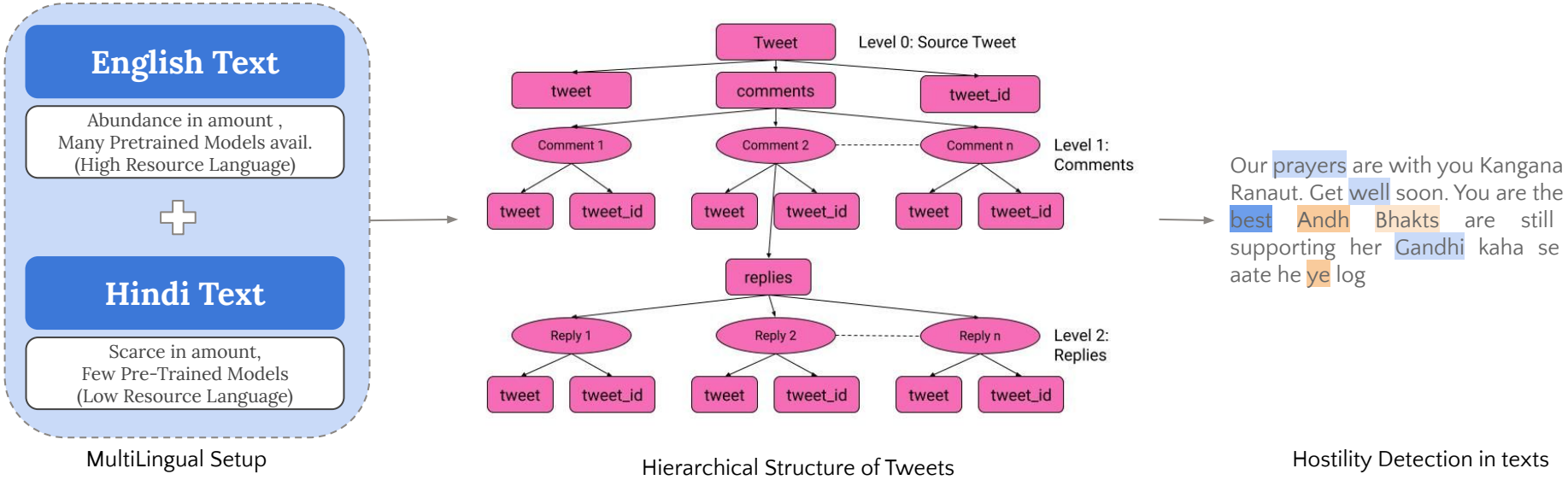
Outline

- Introduction
- Problem Statement
- Proposed Model
- Experimental Setup
- Results and Analysis
- Conclusions

Outline

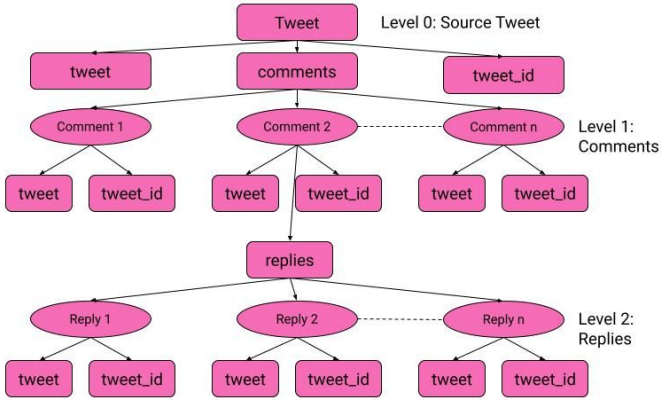
- Introduction
- Problem Statement
- Proposed Model
- Experimental Setup
- Results and Analysis
- Conclusions

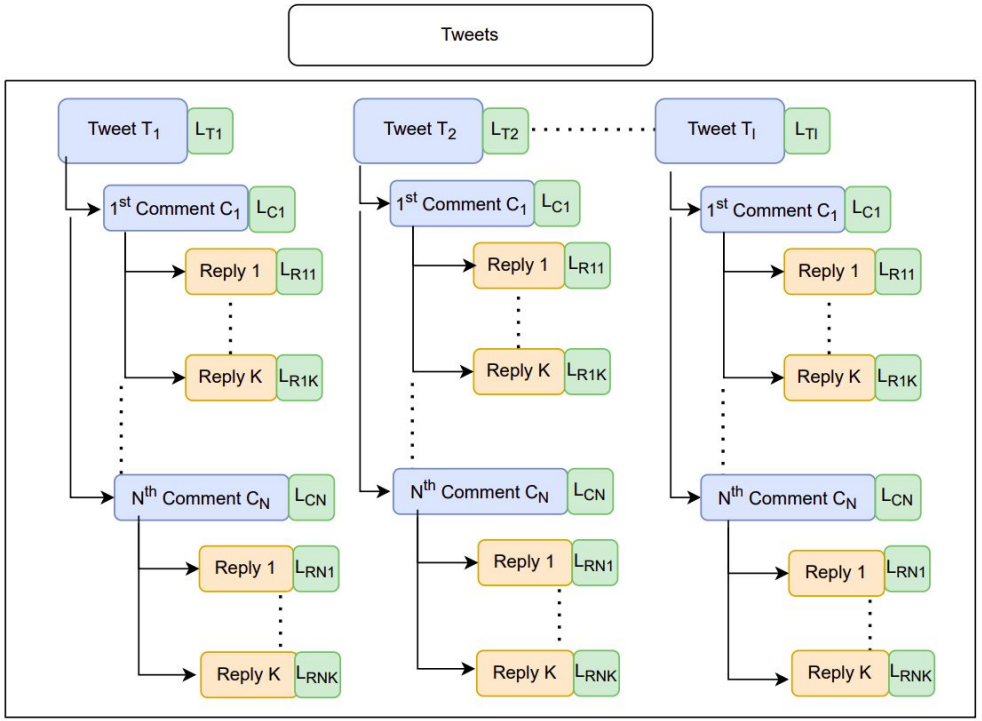
Hostility detection in hierarchical structure of tweets written in multiple languages.



General structure of tweets and challenging multilingual conversational datasets

Structure of Conversational Code-Mixed Tweets





Outline

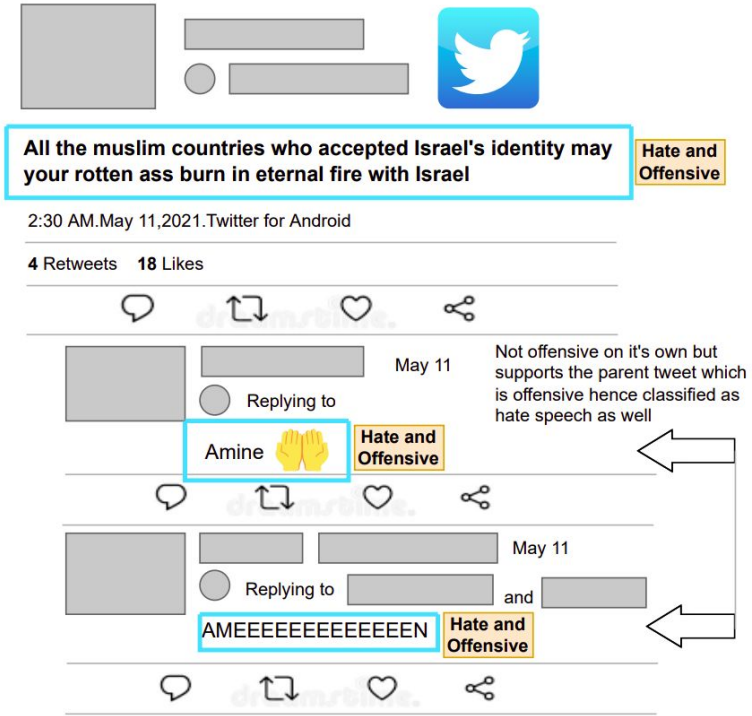
- Introduction
- **Problem Statement**
- Proposed Model
- Experimental Setup
- Results and Analysis
- Conclusions

Problem Statement

Determine whether a given Code-Mixed text (i.e., post/comment/reply) is Non-Hate-Offensive (NONE) or Hate and Offensive (HOF).

(NONE) Non-Hate-Offensive

(HOF) Hate and Offensive



Literature Review

Battling Hateful Content in Indic Languages HASOC '21

Aditya Kadam[a], Anmol Goel[a], Jivitesh Jain[a], Jushaan Singh Kalra[b], Mallika Subramanian[a], Manvith Reddy[a], Prashant Kodali[a], T.H. Arjun[a], Manish Shrivastava[a] and Ponnurangam Kumaraguru[a]

[a]*International Institute of Information Technology, Hyderabad, India*
[b]*Delhi Technological University, Delhi, India*

Exploring Transformer Based Models to Identify Hate Speech and Offensive Content in English and Indo-Aryan Languages

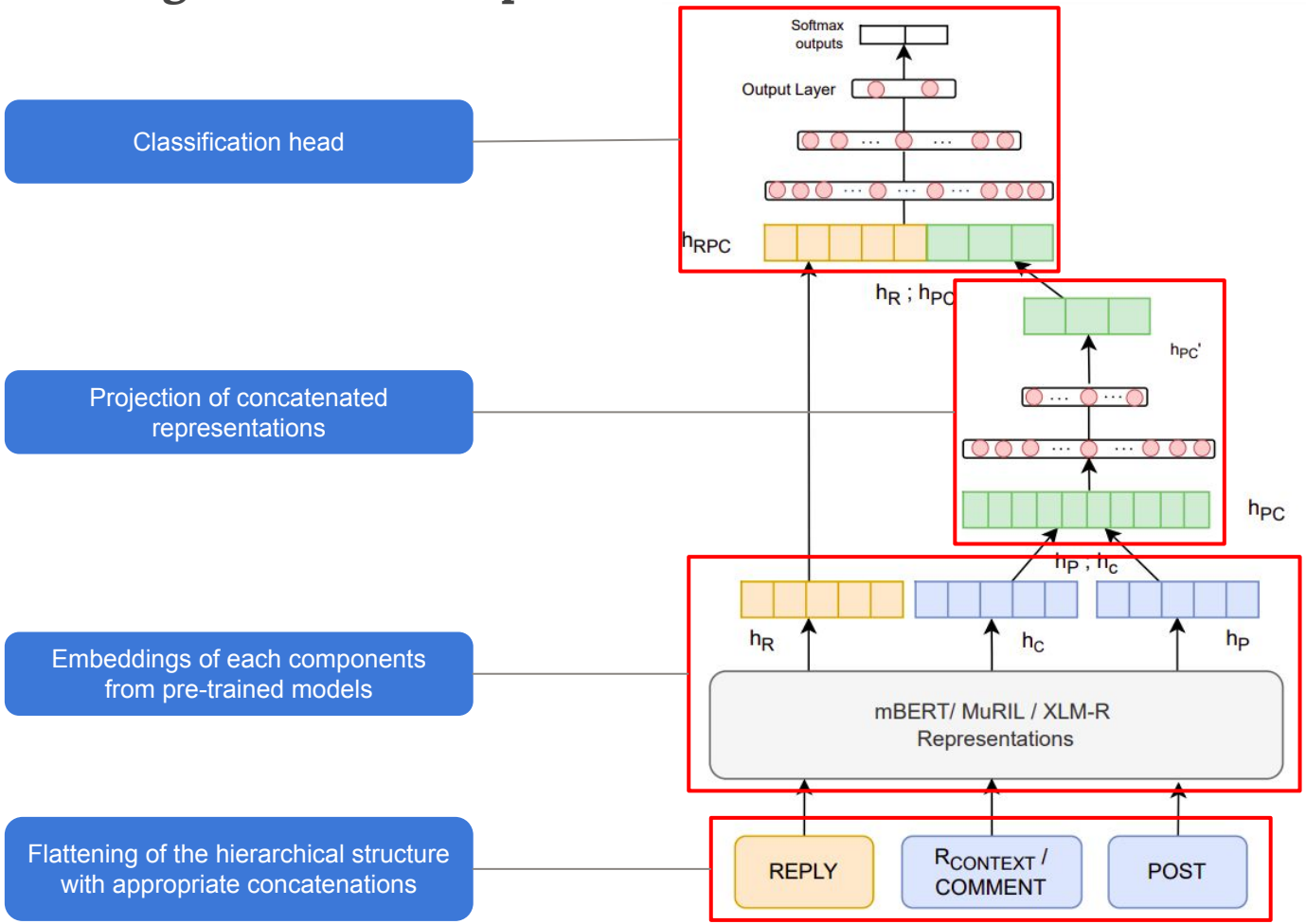
Somnath Banerjee[a], Maulindu Sarkar[b], Nancy Agrawal[b], Punyajoy Saha[a] and Mithun Das[a]

[a]*Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India*
[b]*Department of Electrical Engineering , Indian Institute of Technology, Kharagpur, West Bengal, India*

Outline

- Introduction
- Problem Statement
- **Proposed Model**
- Experimental Setup
- Results and Analysis
- Conclusions

Architectural Diagram of the Proposed Model



Formulation of the Proposed Model

Processing for POST/COMMENT/$R_{CONTEXT}$/REPLY

POST(P)

- Step 1: h_p = PT(P)
- Step 2: h_L = MLP(h_p)
- Step 3: h_S = SL(h_L)

$R_{CONTEXT}$/COMMENT(C)

- Step 1: h_p= PT(P) , h_c = PT(C)
- Step 2: $h_{pc} = h_p;h_c$
- Step 3: h_L = MLP(h_{p_c})
- Step 4: h_S = SL(h_L)

REPLY(R)

- Step 1: h_p = PT(P), h_c = PT(C), h_r = PT(R)
- Step 2: $h_{pc} = h_p;h_c$
- Step 3: h_{L1} = MLP(h_{p_c})
- Step 4: $h_{rpc} = h_r; h_{L_1}$
- Step 5: h_{L2} = MLP'($h_{r_{p_c}}$)
- Step 6: h_S = SL(h_{L_2})

Creation of Rcontext for replies

- If post has only one reply then the Rcontext for the reply is parent comment only.
- If post has k replies then the Rcontext for t th reply is concatenation of comment and 1 to (t – 1) th replies.

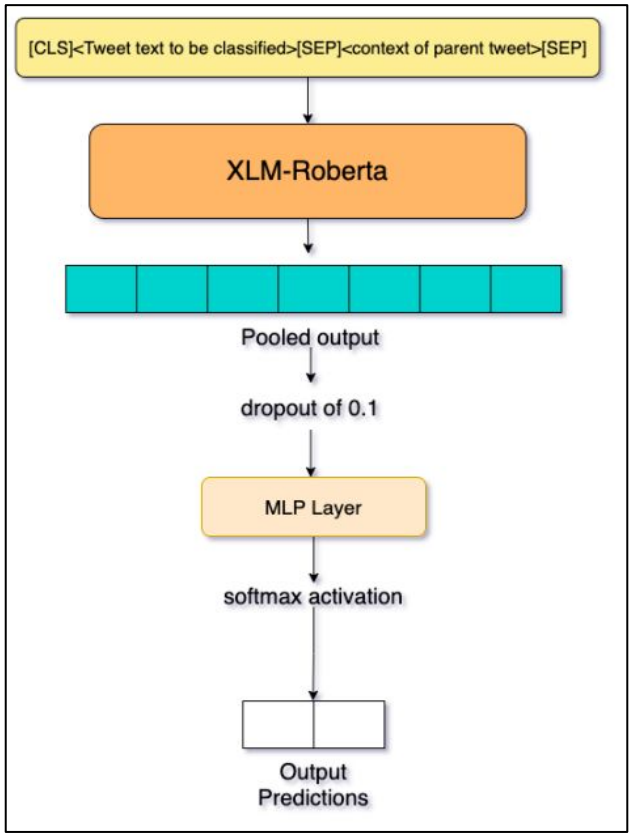
Process

- If the input is only post, a two-layer Multi-layer Perceptron (MLP) is used to obtain (h_L).
- For the comment, ([h_c ;h_p]) are concatenated and hL is obtained.
- For reply, ([hc ;hp]) are concatenated where hc is representation of rcontext. This is passed through MLP layer to obtain ($hL1$). ([hr;$hL1$]) is concatenated and passed through another MLP layer to obtain ($hL2$).
- Logits are obtained (hs) by passing the representations through a softmax layer (SL)

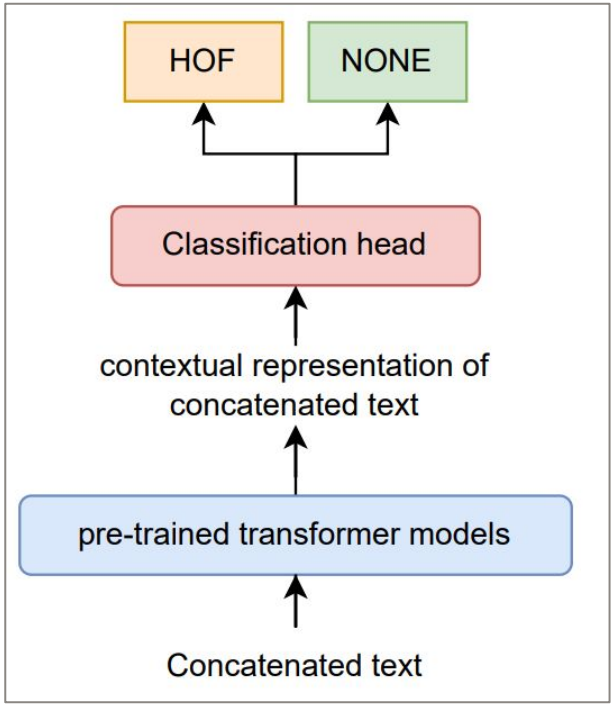
Outline

- Introduction
- Problem Statement
- Proposed Model
- **Experimental Setup**
- Results and Analysis
- Conclusions

Baselines

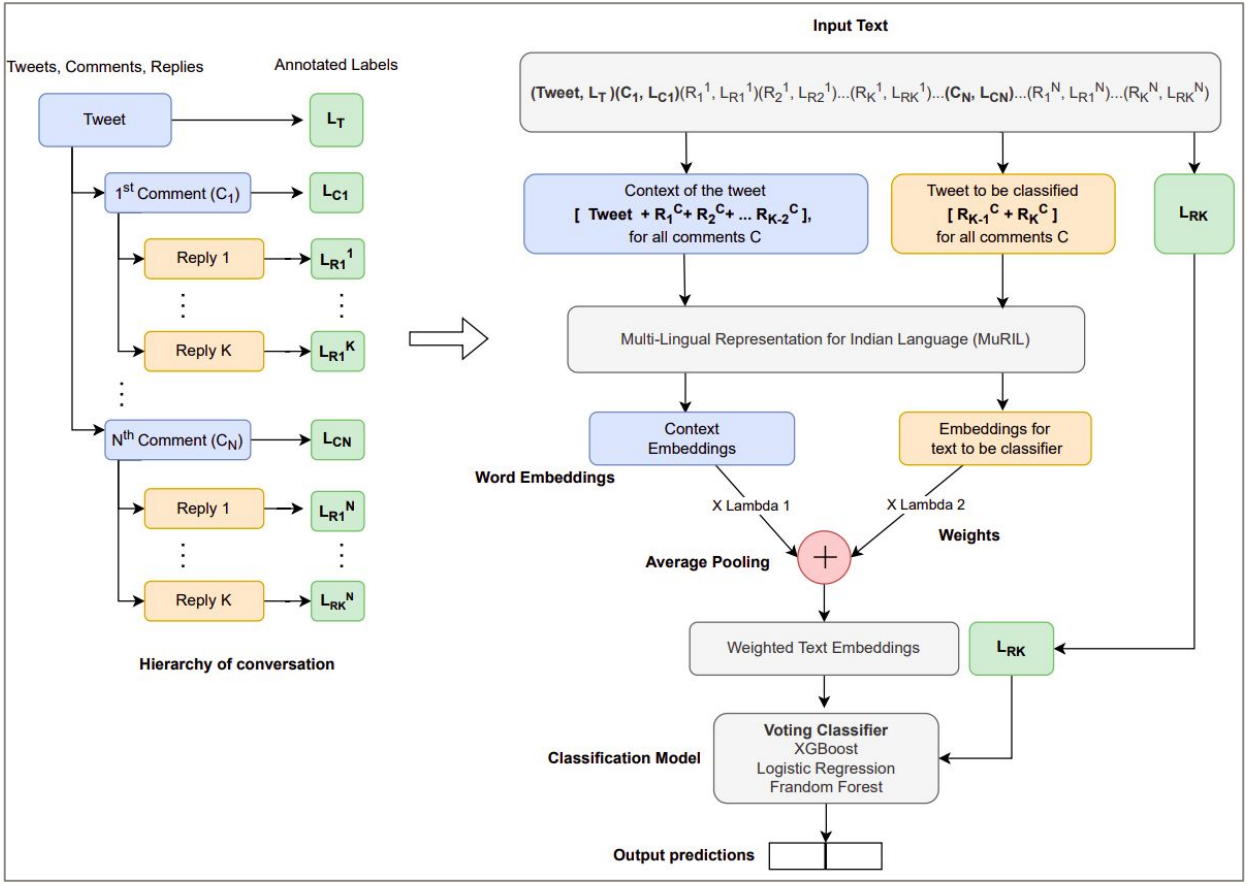


Code-Mixed XLMR



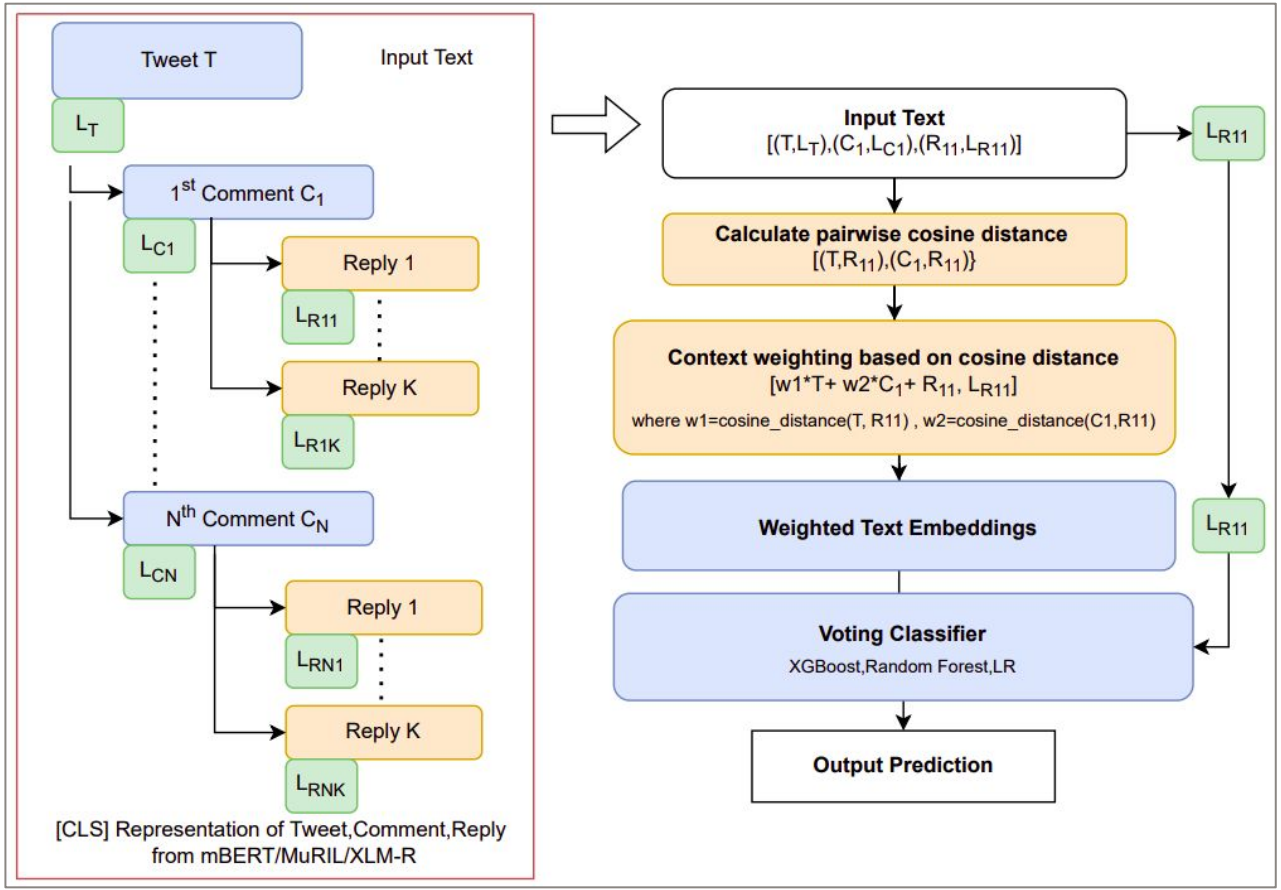
Simple concatenation baseline (SCB)

Baselines



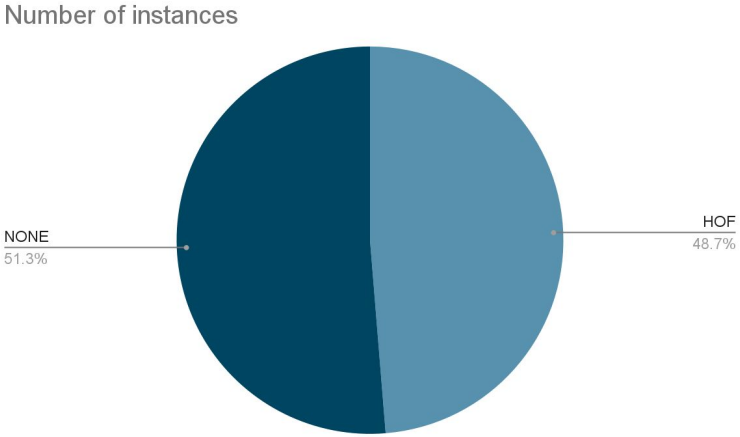
Weighted context baseline (WCB)

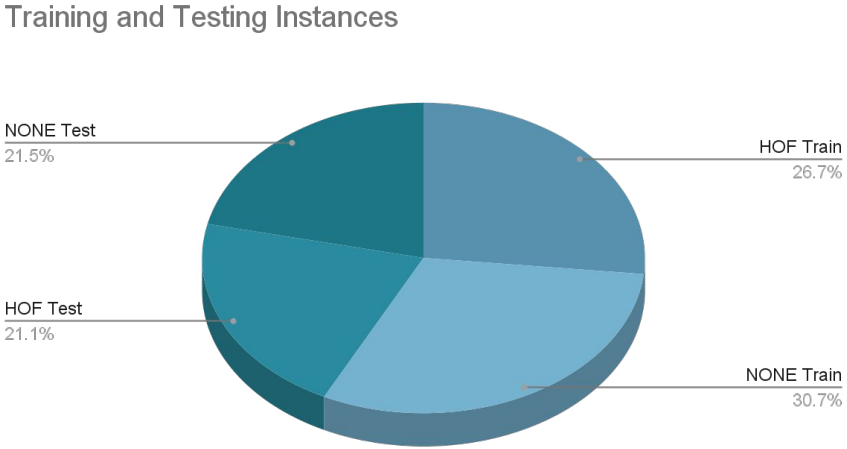
Baselines



Cosine Attention Baseline(CAB)

Dataset and Evaluation Metrics





Dataset	#C	#E	HOF	NONE
Train set	37	2819	1309 (46%)	1510 (54%)
Test set	25	2092	1037 (50%)	1055 (50%)
Total	62	4911	2346 (48%)	2565 (52%)

Accuracy

The number of correct predictions in all the predictions.

$$Accuracy = \frac{\text{Number of correct Prediction}}{\text{Total number of predictions}}$$

F1 Score

The harmonic mean of precision and recall.

$$F_1 Score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Outline

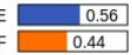
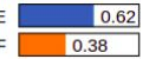
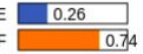
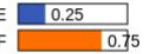
- Introduction
- Problem Statement
- Proposed Model
- Experimental Setup
- **Results and Analysis**
- Conclusions

Proposed model outperforms all baselines

Model	Method	Accuracy					F1 Score				
		RF	LR	XGB	VC	Direct FT	RF	LR	XGB	VC	Direct FT
CM-XLMR	XLM-R + Norm	-	-	-	-	0.61	-	-	-	-	0.46
SCB	mBERT	0.55	0.61	0.49	0.57	0.56	0.55	0.60	0.57	0.49	0.50
	MuRIL	0.50	0.40	0.45	0.46	0.57	0.50	0.29	0.45	0.45	0.51
	XLM-R	0.55	0.58	0.52	0.58	0.40	0.54	0.49	0.50	0.53	0.27
WBC	mBERT	0.62	0.59	0.61	0.62	0.66	0.61	0.57	0.60	0.61	0.64
	MuRIL	0.59	0.41	0.54	0.53	0.40	0.55	0.29	0.52	0.53	0.29
	XLM-R	0.64	0.64	0.59	0.64	0.66	0.60	0.62	0.57	0.61	0.65
SLCB	mBERT	0.64	0.55	0.60	0.62	0.66	0.58	0.57	0.54	0.58	0.61
	MuRIL	0.64	0.60	0.55	0.62	0.62	0.57	0.56	0.54	0.57	0.55
	XLM-R	0.64	0.62	0.61	0.65	0.40	0.62	0.60	0.59	0.63	0.27
CAB	mBERT	0.57	0.58	0.55	0.58	0.58	0.57	0.58	0.55	0.58	0.53
	MuRIL	0.60	0.59	0.61	0.65	0.58	0.60	0.58	0.61	0.64	0.54
	XLM-R	0.62	0.64	0.59	0.64	0.63	0.61	0.64	0.59	0.64	0.60
Hierarchial	mBERT	0.54	0.58	0.60	0.62	0.60	0.52	0.54	0.56	0.62	0.65
	MuRIL	0.59	0.63	0.62	0.64	0.63	0.55	0.61	0.60	0.64	0.67
	XLM-R	0.63	0.61	0.64	0.66	**0.68**	0.62	0.60	0.62	0.63	**0.72**

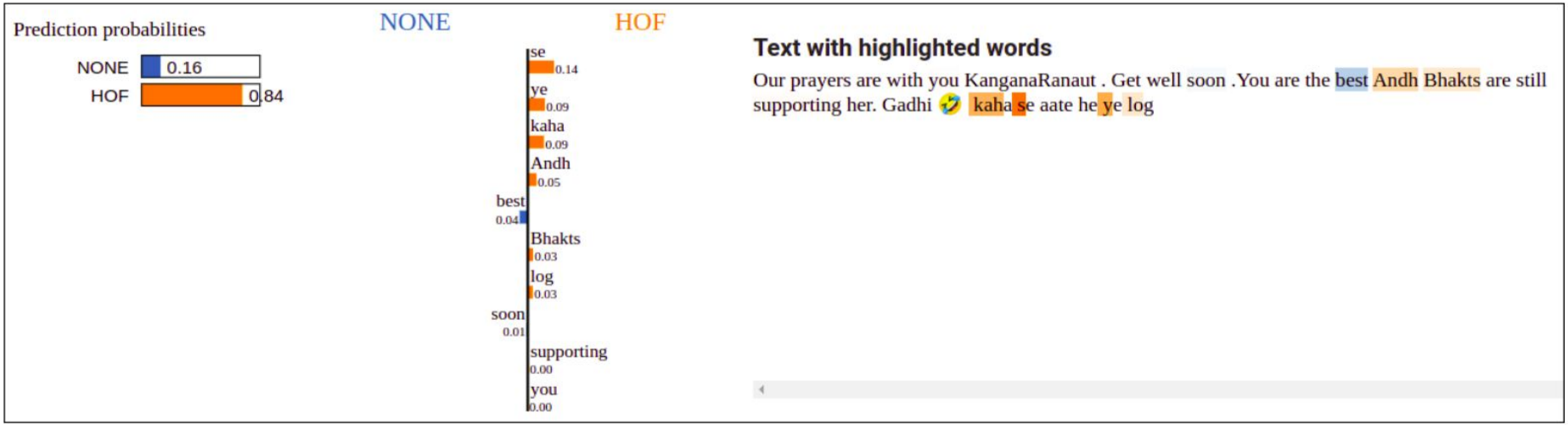
Accuracy and F1 scores for baselines and proposed model. Symbol '-' indicates that the results are not available. CM-XLMR = Code-Mixed XLM-R, RF = Random Forest, LR = Logistic Regression, XGB = XG-Boost, VC = Voting Classifier, Direct FT = Direct Fine-Tuning

Visualization of proposed model's attention while predicting labels using LIME Analysis

Exp.	GT	Predictions	Type	Texts with highlighted words
1	NONE		P	Ok but liberals are not 'anti-Hindu' ..
2	NONE		P+C	Ok but liberals are not 'anti-Hindu' .. You're anti-Hindu and non-liberal. Evidence maang lo bas, flood hon jaayega.
3	NONE		P+C+R	Ok but liberals are not 'anti-Hindu' .. You're anti-Hindu and non-liberal. Evidence maang lo bas, flood hon jaayega. Flood it.
4	NONE		P	Our prayers are with you KanganaRanaut . Get well soon .You are the best
5	HOF		P+C	Our prayers are with you KanganaRanaut . Get well soon .You are the best Andh Bhakts are still supporting her. Gadhi 🤣
6	HOF		P+C+R	Our prayers are with you KanganaRanaut . Get well soon .You are the best Andh Bhakts are still supporting her. Gadhi 🤣 kaha se aate he ye log
7	NONE		P	Religious conversion has become the biggest national challenge in India after terrorism. आपदा_में_धर्मपरिवर्तन_का_खेल
8	NONE		P+C	Religious conversion has become the biggest national challenge in India after terrorism. आपदा_में_धर्मपरिवर्तन_का_खेल If someone change his religion by his choose then what is your problem?
9	NONE		P+C+R	Religious conversion has become the biggest national challenge in India after terrorism. आपदा_में_धर्मपरिवर्तन_का_खेल If someone change his religion by his choose then what is your problem? Appne dekhona bhai, kyu khujli horahi he?

LIME Analysis, *GT = Ground Truth, P = POST, C = COMMENT, R = REPLY

Attention distribution in a instance



All the components and layers in the proposed architecture is important for classification

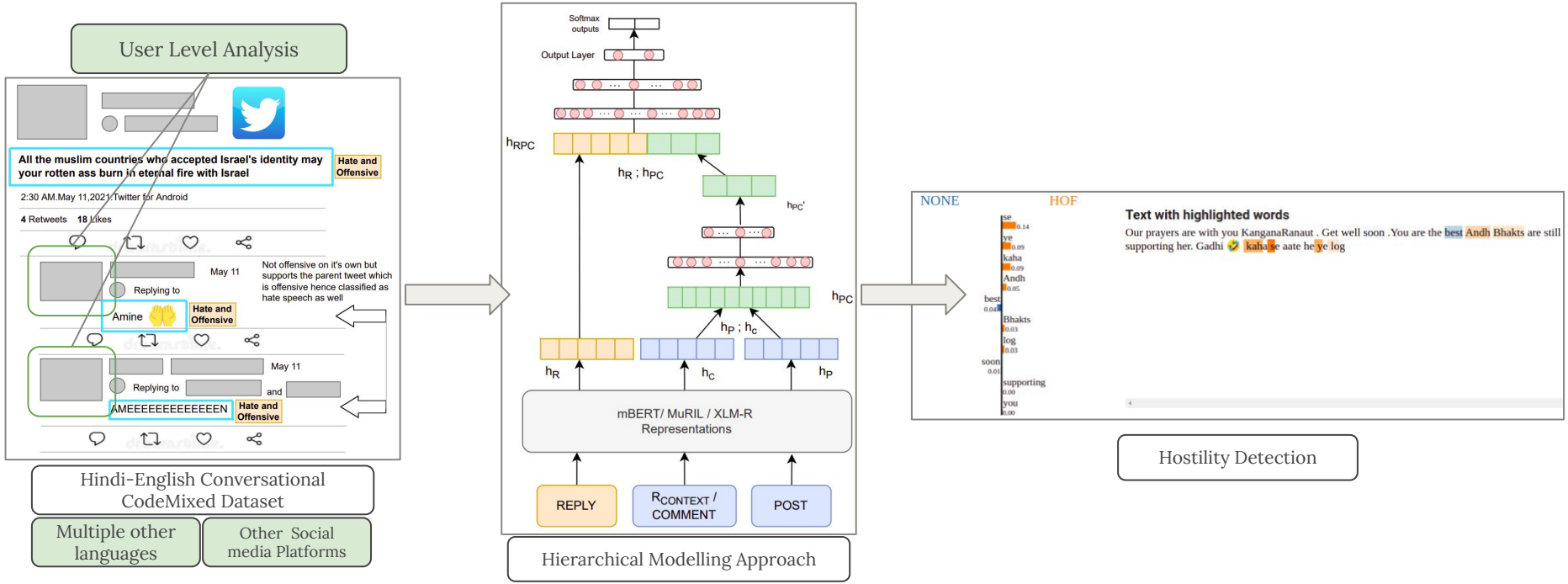
Abalation Type	Setup	Accuracy	F_1 Score
No removal	**Hierarchical**	**0.679**	**0.716**
Component Removal	Setup 1	0.522	0.511
	Setup 2	0.524	0.408
	Setup 3	0.521	0.511
	Setup 4	0.518	0.510
Context Removal	Setup 5	0.519	0.506
	Setup 6	0.519	0.509
	Setup 7	0.512	0.393

Setups 1, 2, 3 is created by removing a linear layer at a time and 4 is created by removing 1, 2 together. Similarly Setups 5, 6 are created by removing context and post and Setup 7 is created by removing both post and context.

Outline

- Introduction
- Problem Statement
- Proposed Model
- Experimental Setup
- Results and Analysis
- Conclusions

Some natural extension of the proposed model would be



- The paper presented a **novel hierarchical neural network architecture** for detecting hate and offensive content in Hindi-English Code-Mixed conversations.

- It exploits the inherent hierarchy of the online social media conversational threads and provides selective and abstractive context for a given utterance to boost the model performance.

References

1. Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring Transformer Based Models to Identify Hate Speech and Offensive Content in English and Indo-Aryan Languages.

2. Mohit Bhardwaj, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility Detection Dataset in Hindi.

3. Arkadipta De, Venkatesh Elangovan, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2021. Coarse and fine-grained hostility detection in Hindi posts using fine tuned multilingual embeddings

4. Chander Shekhar, Bhavya Bagla, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2021. Walk in Wild: An Ensemble Approach for Hostility Detection in Hindi Posts.

5. Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and Offensive Speech Detection in Hindi and Marathi.

6. Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media.

Thanks!

Any questions ?

Reach us at:

- cs21mtech14007@iith.ac.in
- cs21mtech16001@iith.ac.in

Or raise an issue at:

https://github.com/AditiBagora/Hasoc2021CodeMix