Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives

Shivang Chopra – Delhi Technological University
Ramit Sawhney – Netaji Subhas Institute of Technology
Puneet Mathur – University of Maryland, College Park
Rajiv Ratn Shah – MIDAS, IIIT-Delhi

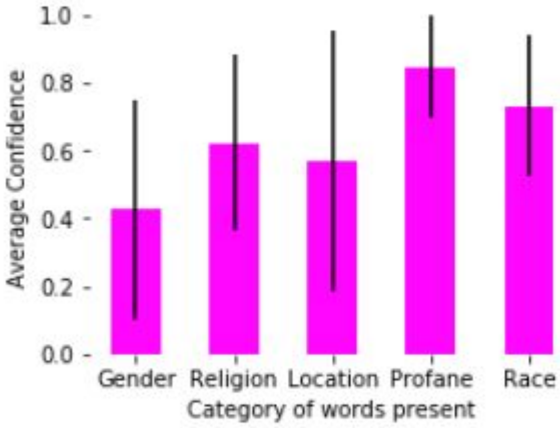
MIDAS@IIITD
Multimodal Digital Media Analysis Lab

Context & Objective

- Model & Analyze the linguistic aspects of **low-resource** code-switched; Hindi-English **offensive** social media text.
- Empirically validate the effectiveness of exploiting linguistic homophily (connected people speak similarly).
- Discover and normalize bias across **race, religion & gender** in **Hinglish.**

Challenges

- Loose syntactic and semantic structure
- Diverse categories of variable biases
- Community based author profiling



Contributions

- Hindi-English Code-Switched --- Modeling & Processing
- Author profiling --- Graph Embeddings
- Bias --- Identification & Elimination

Ethical Considerations and Limitations

- **Privacy**: Informed Consent, no intervention.
- **Interpretation**: Interpretation of offensive behaviour might be highly subjective.
- **Demographics:** Impairs generalizability --- narrow tightly coupled communities.
- **Medium-Specificity:** Limited to Twitter, but generalizable across social media.
- **Assessment Granularity:** Non-binary, simplification but scalable.

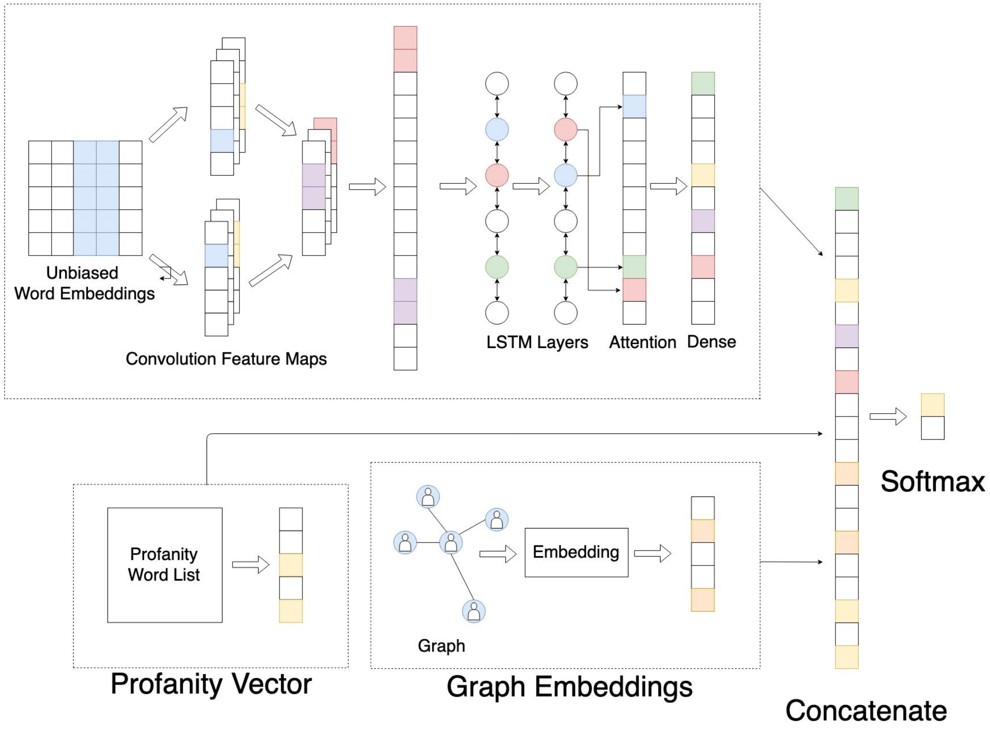
Code Switching: Hinglish & Challenges

- To validate the proposed hypothesis, we use two datasets,HS (Bohra et al. 2018) and HEOT (Mathur et al. 2018b).

Mere musalmaani bhaiyo ko id mubarak. (Id mubarak to my Muslim brothers)	Non-offensive
Vo to congress ka kutta ban chuka hai saala. (He has become a dog for the congress party.)	Offensive
He is a Muslim aadmi. (He is a Muslim man.)	Non-Offensive

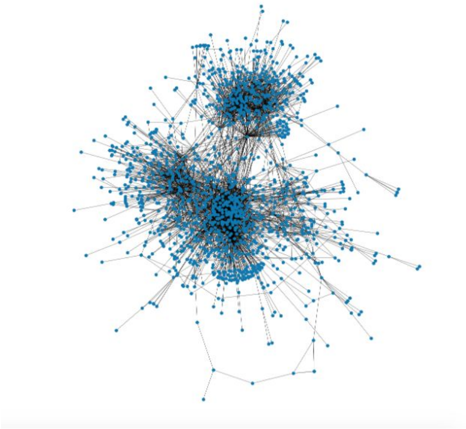
Table 1: Some examples from the dataset

Architecture



Linguistic Homophily and Graph-based Author Profiling

- Tightly coupled communities have high influence on offensive behaviour.
- Incorporating this homophily using Node2Vec.
- Significant performance enhancement observed.



Modelling the Social Graph

$G_{Mentions}$ A mentioned B in a tweet

G_{Quotes} A quotes B’s tweet

$G_{repliedTo}$ A replied to B’s tweet

Gathered from the tweets from the dataset and the historical tweets of users

Statistic	Value
Number of nodes	3005
Number of edges	4448
Average degree	2.9604
Maximum path length	11
Largest Connected component	1481

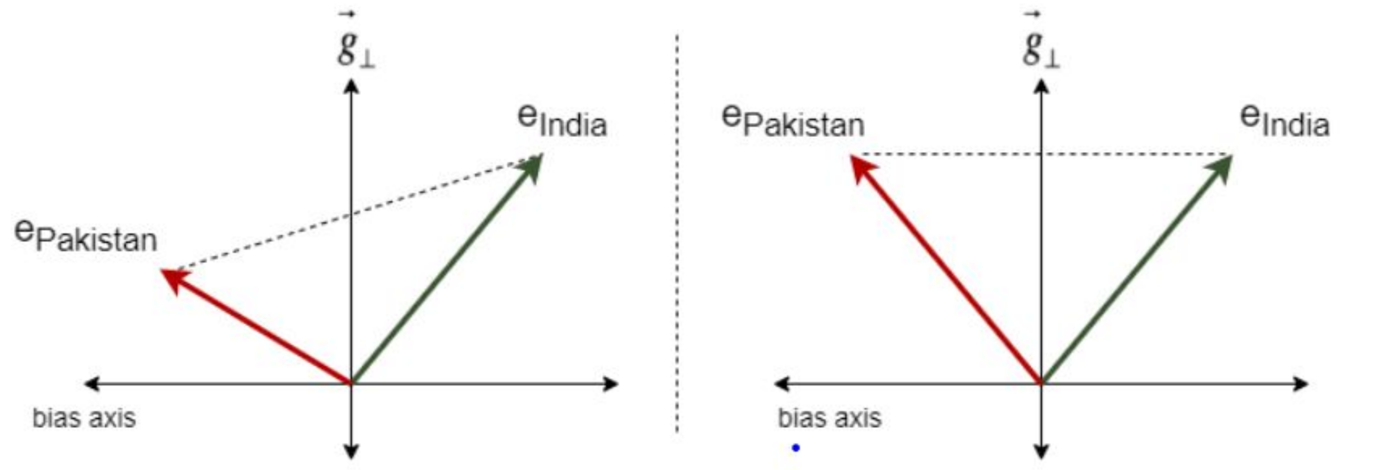
Table 2: Graph statistics for HS dataset

Pair Generation

- Clustering Words
- Making pairs
- Expert segregation

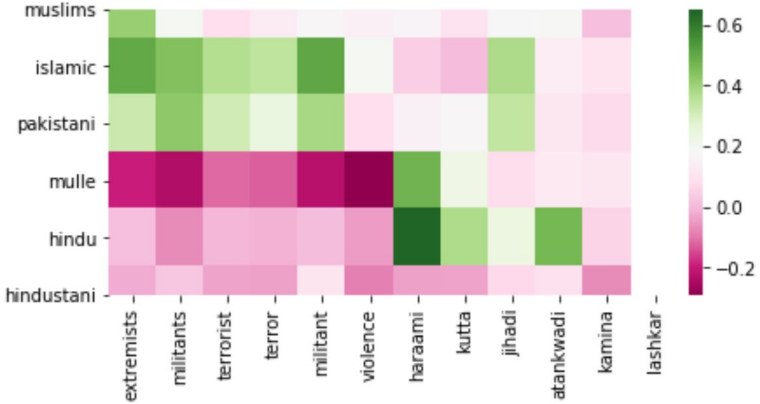
Hindu	Muslim
India	Pakistan
Bhai (Brother)	Behen (Sister)
Mom	Dad
Kutta (Dog)	Kutiya (Bitch)
Chacha (Uncle)	Chachi (Aunt)

Bias Correction

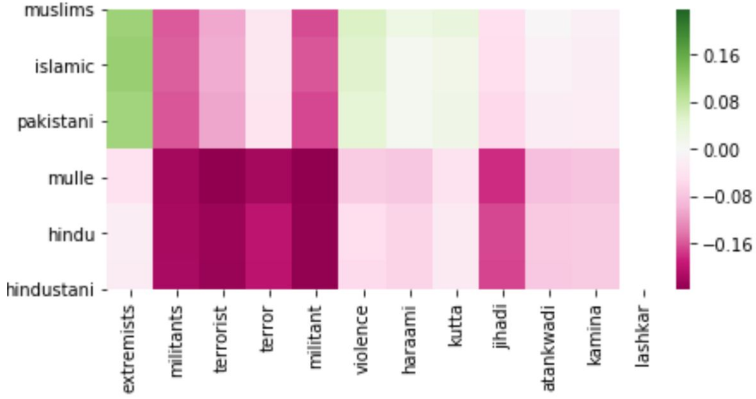


Neutalization and Equalizing

Analysing Debiased Word Embeddings



(a) Pre-Debiasing



(b) Post-Debiasing

Results

	Model	HS		HEOT	
		F1	Acc	F1	Acc
Text Comparison	CNN	0.49	0.51	0.70	0.76
	CNN + LSTM	0.60	0.60	0.69	0.70
	CNN + BiLSTM	0.54	0.56	0.72	0.76
	BiLSTM	0.58	0.59	0.61	0.63
	BiLSTM + Attn	0.62	0.62	0.71	0.77
	CNN + BiLSTM + Attn	0.62	0.62	0.72	0.76
Graph + Text Ablation	CNN + node2vec	0.50	0.52	NA	NA
	CNN + LSTM + n2v	0.61	0.61	NA	NA
	CNN + BiLSTM + node2vec	0.57	0.57	NA	NA
	BiLSTM + n2v	0.59	0.59	NA	NA
	BiLSTM + Attn + node2vec	0.62	0.63	NA	NA
	CNN + BiLSTM + Attn + node2vec	0.63	0.64	NA	NA
	CNN + BiLSTM + Attn + DeepWalk (DW)	0.67	0.71	NA	NA
	PV + node2vec	0.52	0.63	NA	NA
PV Incorporation	CNN + BiLSTM + Attn + PV	0.64	0.71	0.77	0.85
	CNN + BiLSTM + Attn + PV + DW	0.73	0.78	NA	NA
Debiasing Ablation	CNN + BiLSTM + Attn + PV + POSDeb	0.64	0.70	0.70	0.73
	CNN + BiLSTM + Attn + PV + MBE	0.68	0.72	**0.86**$^{+}$	**0.87**$^{+}$
	CNN + BiLSTM + Attn + PV + DW + MBE	**0.76**$^{+}$	**0.78**$^{+}$	NA	NA
Comparative	LSTM + Transfer Learning [Kapoor et al. 2018]	0.71*	0.74*	0.79*	0.87
	CNN + Transfer Learning [Mathur et al. 2018b]	0.69*	0.72	0.71*	0.83*
	Statistical ML [Bohra et al. 2018]	0.62*	0.71	0.70*	0.76*
	Hierarchical LSTM [Santosh and Aravind 2019]	0.48	0.71	0.52*	0.63*
	Ours	0.76	0.78	0.86	0.87

Table 3: Ablation and comparative results in terms of Accuracy and F1 score. NA : No results due to unavailability of data * : Replication of baselines + : Statistically significant results

Conclusion

Developed a robust classifier for Hindi-English hate speech detection

Incorporated bias elimination to improve robustness and ensure fairness

Extensive Qualitative and Quantitative Analysis performed.

Utilized social network graphs for author profiling to enrich the classification model

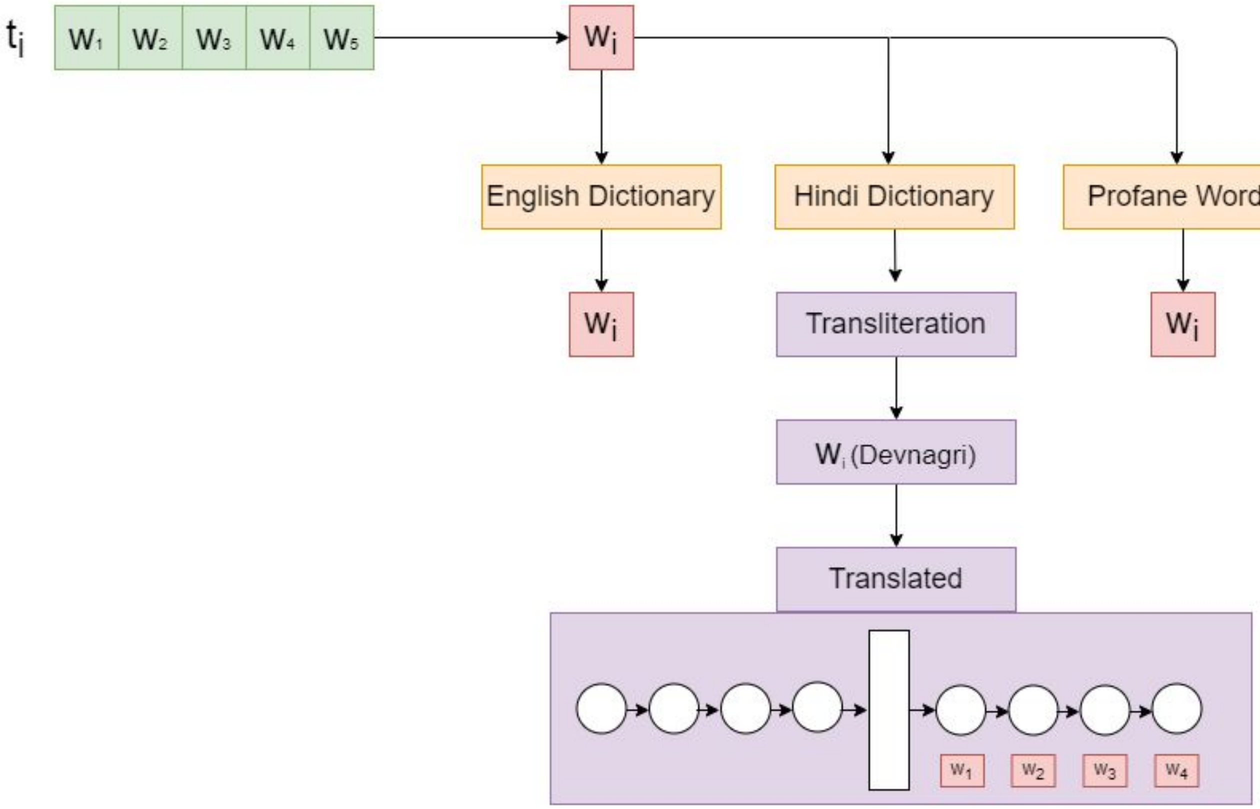
Thank You

Linguistic Backbone and Profanity Vector Augmentation

- BiLSTM-based architecture used as the backbone network.
- Profanity Vector incorporated along with the tweet latent vector.

$$\mathbf{PV}^{(j)} = \begin{cases} 0 & \text{if } p_j \in t_i \\ 1 & \text{if } p_j \notin t_i \end{cases}$$

Data Preprocessing



Qualitative Analysis