Authors :
Raghav Kapoor, Yaman Kumar,
Kshitij Rajput, Rajiv Ratn Shah,
Ponnurangam Kumaraguru,
Roger Zimmermann

MIDAS@IIITD
Multimodal Digital Media Analysis Lab

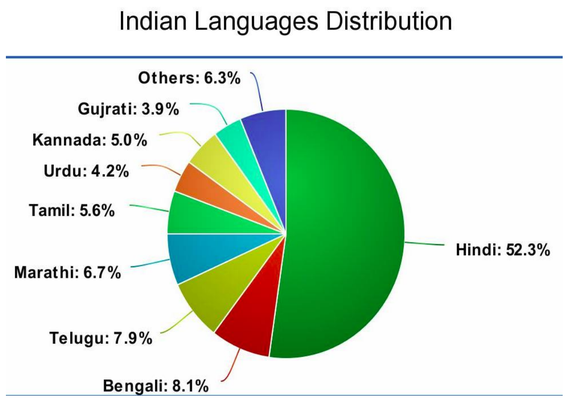
Mind Your Language : Abuse and Offense detection using Code-Switched Languages

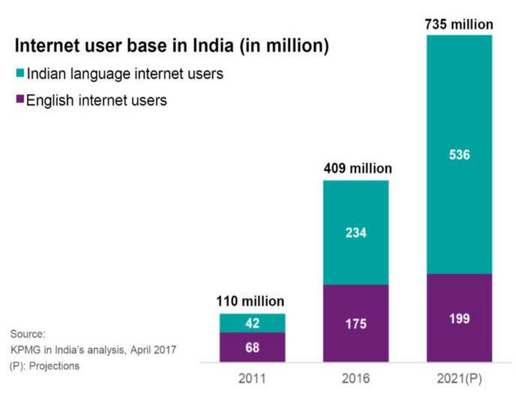
Research Goal

This paper lays out an A.I. based system that recognizes and vacuums out the illicit use of "Hinglish" (Hindi + English) code-switched language on social media platforms that induces a sense of hate and animosity for the users

Its significance is marked by the fact :

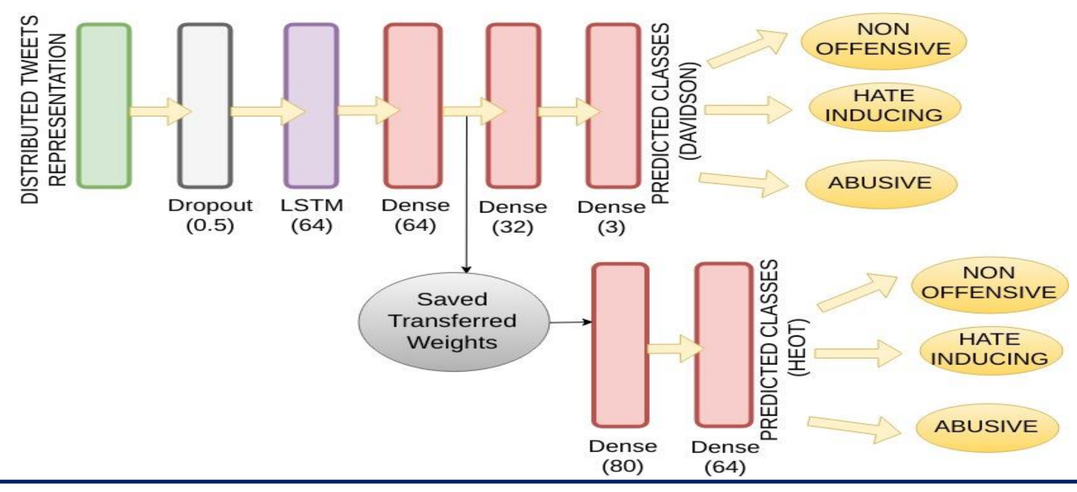
- **In multilingual societies (Eg: India), code-switched language is most popular**
- **Indian Internet Users crossed 500 million**
- **Tackles the difficulty of non-fixed grammar, vocabulary and semantics of the language pair**





System Architecture

- **Preprocessing** involves **transliteration** (change of script) to English and **translation** (conversion of words)
- **Input to model** is the embeddings previously trained on Glove and Word2vec
- **Dataset used :** HEOT dataset, Davidson tweets dataset
- **Ternary classification model** (abusive, benign, hate inducing) using **LSTM** and **transfer learning** to utilize the benefits from weights trained before.
- **Accuracy, F1, Precision, Recall** are used as metrics to evaluate the credibility of results.



Workflow Model

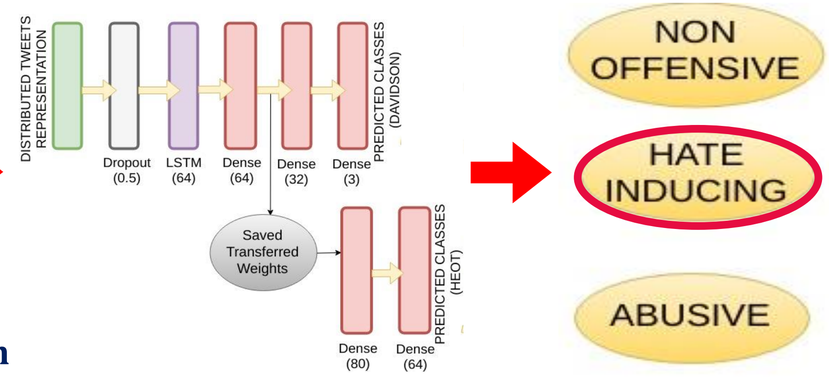
हम उन्हें बर्बाद कर देंगे

ham unhe barbaad kar denge

(Indic-transliteration python library)

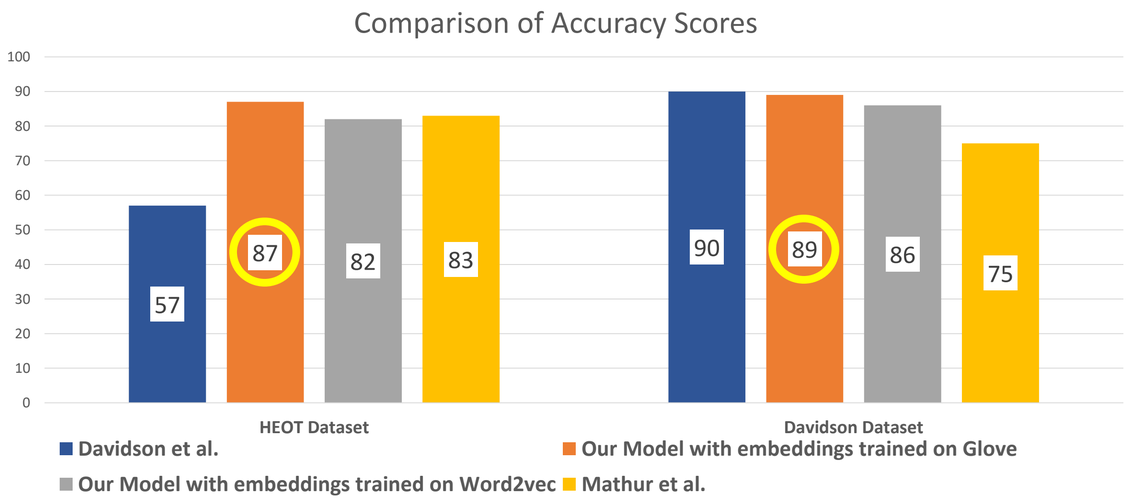
we will destroy them

(Xlit Crowd Conversion Dictionary)



Comments from Social Media → Comments in Hindi → Trans-literation to English → Translation to English → Passes through our system → Highly Accurate Results

Results



"STATE OF THE ART" RESULTS FOR HINGLISH

Model	Precision Score	
	HEOT Dataset	Davidson Dataset
Davidson et. al	0.573	0.91
Our Model with embeddings trained on Glove	0.868	0.885
Our Model with embeddings trained on Word2Vec	0.814	0.859
Mathur et. al	0.802	0.672

- Best results for our model trained on Glove embeddings on HEOT
- Comparable results on Davidson English tweets dataset.

Applications

- **Detect False Propaganda** by Political Groups in **Elections**
- **Youtube/Netflix Subtitles –** "Auto-beep" offensive language
- **Online Social Media -** Report Defamatory Pages and comments
- **Feedback analytics** for better user experience.
- **Real time "clean-chat" facility.**
- **Censor board** – Auto-eliminate abusive content.

Watch Demo Video Now!!