"Hinglish" Language - Modeling a Messy Code-Mixed Language

"Hinglish"

Hinglish is a linguistic blend of Hindi (very widely spoken language in India) and English (an associate language of urban areas) and is spoken by upwards of 350 million people in India

Messy language

1. Geographical variation
2. Language and phonetics variation
3. No grammar rules
4. Spelling variation
5. 3000 examples only !!

Text Augmentation

1. Synonym Replacement
2. Random Insertion
3. Random Swap
4. Random Deletion

Data

Table 2: Examples in the dataset

Hinglish and English Data		
Label	HOT	English
Non-Offensive	Hum sab ghumne jaa rahe hain? http://t. @username1	We all are going outside? http://t... @username1
Offensive	<redacted content>! Mujhe mat sikha:/	<redacted content>! Do not teach me:/
Hate Inducing	<redacted content> terrorist Akbaar kill SaveWorld	<redacted content> Kill terrorist Akbaar SaveWorld

Table 1: Annotated Data set

Hinglish and English Data		
Label	HOT	English
Non-Offensive	1121	7274
Offensive	303	4836
Hate Inducing	1765	2399
Total	3189	14509

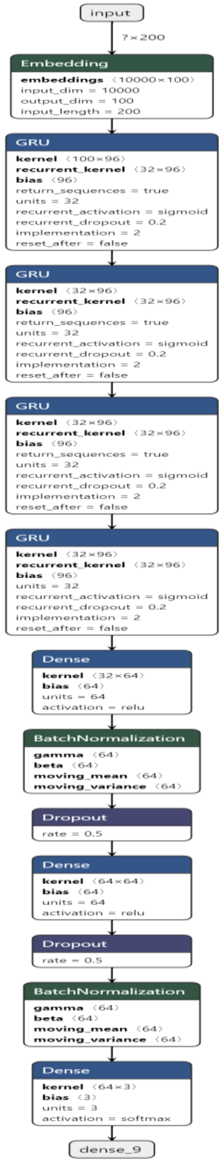
End to End Process

Pre-Processing → Translation Transliteration → Train/Test Split → Train EDA → Model Training/Evaluation

Hyperparameters and Training

1. Learning rate : **0.01**, .001, .003. 0.005
2. RNN types – LSTM, BiLSTM, **GRU**, SimpleRNN.
3. Pre-trained embeddings with fine tuning: **True**., False
4. FC Dense layers: 3, **2**, 1, 0
5. Recurrent Drop out: 0.2, **0.4**
6. RNN units: **Stacked** , Single
7. Embedding dimensions: 50, **100**, 200
8. **Early Stopping, Model Checkpoint, LR Decay, LR Reduce on plateau.**
9. Keras – Sequential API

Model Architecture



Results

Network	BiLSTM/32x2 FC64x2_Dense_3 Recurrent Drop Out(DO)						BiLSTM/32x2 FC64x1_Dense_3 Recurrent Drop Out(DO)						BiLSTM/32x1 FC64x2_Dense_3 Recurrent Drop Out(DO)					
Labels	DO - 0.2			DO - 0.4			DO - 0.2			DO - 0.4			DO - 0.2			DO - 0.4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Non-offensive	0.68	0.69	0.69	0.72	0.54	0.62	0.72	0.65	0.69	0.73	0.65	0.69	0.71	0.59	0.65	0.72	0.75	0.73
Hateful	0.4	0.75	0.52	0.37	0.72	0.49	0.62	0.42	0.5	0.46	0.59	0.52	0.34	0.77	0.48	0.56	0.51	0.53
Offensive	0.87	0.73	0.8	0.84	0.81	0.82	0.79	0.88	0.83	0.84	0.85	0.85	0.86	0.76	0.8	0.86	0.86	0.86
Accuracy	0.72			0.72			0.76			0.76			0.7			0.79		

Network	BiLSTM/32x1 FC64x1_Dense_3 Recurrent Drop Out(DO)						BiLSTM/32x4 FC64x1_Dense_3 Recurrent Drop Out(DO)						GRU/32x4 FC64x1_Dense_3 Recurrent Drop Out(DO)					
Labels	DO - 0.2			DO - 0.4			DO - 0.2			DO - 0.4			DO - 0.2			DO - 0.4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Non-offensive	0.76	0.71	0.73	0.75	0.67	0.71	0.7	0.63	0.66	0.75	0.67	0.71	0.64	0.81	0.72	0.67	0.85	0.75
Hateful	0.61	0.49	0.54	0.51	0.57	0.53	0.47	0.65	0.55	0.51	0.57	0.53	0.42	0.58	0.48	0.64	0.46	0.54
Offensive	0.83	0.88	0.85	0.84	0.88	0.86	0.83	0.83	0.83	0.84	0.88	0.86	0.92	0.73	0.81	0.9	0.8	0.85
Accuracy	0.79			0.78			0.75			0.78			0.73			0.78		

Loss Function

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}1_{y_i\in C_c}P_{model}[y_i \in C_c]$$

Vivek Kumar Gupta