"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

**Model Architecture**

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

**End to End Process**

**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

**Results**

<table>
<thead>
<tr>
<th>Network</th>
<th>BERTFNL</th>
<th>BERTFNL</th>
<th>BERTFNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>Recall</td>
<td>0.86</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>Precision</td>
<td>0.90</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Loss Function**

\[- \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} I_{c \in C_i} P_{model}(y_i \in C_i)\]

Vivek Kumar Gupta