Early Detection of Depression in Multi-lingual Code-Switching

Brij Rokad, Dhaval Bhanderi, Rutu Gandhi, Shrinidhi Adke

Introduction
- Code-switching is a linguistic phenomenon that encapsulates the way multilingual people communicate.
- Exploring the effects of code switching on the early detection of depression

Methods
1. Dataset Scraping and preprocessing
We scraped about 821 tweets with the following hashtags: #depression, #depressed, #happiness
Preprocessing steps like the removal of URLs, emoticons and username are performed.

2. Corpus Creation and Annotation
Language Identification: each word is given one of the tags "eng", "hin" or "other".
Causal Language Annotation: This is the language that the emotion is expressed in.

3. Feature Extraction
- Negation words: a count is maintained for the number of negation words per sentence
- Probability of positivity and negativity = $max\frac{a}{a+b}, \frac{b}{a+b}$
- Character N-Gram
- Word N-Gram
- TF-IDF

4. Classification
Training Datasets: Trained the models on Hindi and English and Hinglish Review datasets as well as on depression dataset.
Machine Learning Algorithms: We trained the multitude of models including Logistic Regression, MLP, SGD, Decision Tree and Naïve Bayes.

5. Results and Discussions:
Out of all the models, we have achieved the highest accuracy of 94% for ensemble model.

Ex: माहि का तो band बज जाएगा अभी.
T: Mahi will get f***ed up now.

Ex: साला यह जनम तोह Waste ही हो गया मेरा .
T: Shit, my birth has become waste.

6. Future Work
- Incorporate the human emotions into natural language processing
- Combine multi-lingual classification with facial expression.

References
- "Corpus Creation and Emotion Prediction for Hindi-English Code-Mixed Social Media Text", D. Vijay et. al. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*
- "Sentiment Analysis For Hindi Language", Piyush Arora, Master's Thesis, IIIT, India.

Language Switching affects Psycho-Linguistics aspects of expressing *Depression. A* Bilingual Corpus for studying *Depression* is created. Abstracted Features of different languages plays significant role in classification.

Supplementary Tables

Emotion	Sentences
Depressed	36.55%
Happy	63.45 %

Table 1: Data Distribution

Caused Language	Sentences
English	54 (18%)
Hindi	228 (76%)
Mixed	16 (5%)

Table 2: Caused Language Statistics

Positive/ Negative Reviews	Sentences
English	50%
Hindi	50%

Table 3: Distribution of Review Dataset

Fig 1: Accuracy and F - Score of *Reviews* in Hinglish

Fig 2: Accuracy and F - Score of *Depression* in Hinglish

UNIVERSITY OF GEORGIA