Differentiating Code-Borrowing from Code-Mixing

Neha Prabhugaonkar, Sai Peketi, Kavita Ganeshan & Unnikrishnan Sureshkumar
Cuddle Inc.
{neha.prabhugaonkar,sai.peketi,kavita.ganeshan,unni.krishnan}@cuddle.ai

ABSTRACT
In linguistics, code-switching and borrowing are two separate and concepts. Code-switching occurs when an individual speaks two distinct languages and switches between the two while doing conversation with another person who will also have knowledge about both languages. The code-switching is effortless, and demonstrates a strong mastery of both languages.

Author Keywords
Code-Borrowing, Code-Switching, Code-Mixed, Twitter

INTRODUCTION
When an individual uses two languages and switches back and froth between the two in the same sentence with fluency, it is called code-switching. This implies that the speaker is comfortable using both languages. The individual can think bilingually and can use both the languages effectively for communication.

On the other hand, we say code-borrowing happens when an individual is using one language which is a primary language and mixes the words or phrases from secondary language. Here, he/she speaks one language, and alters vocabulary from another to fit the primary language.

The main task of the data challenge was to differentiate code-borrowing from code-mixing. To differentiate code-borrowed words from code-mixed data we implemented several metrics which includes Unique User Ratio, Unique Tweet Ratio, Weighted Class Average Model, Inverse Model, TF-IDF and Code Switched Model. this report gives the desciption about each of these models used for ranking 230 words.

CORPUS CREATION AND ANNOTATION

Corpus Creation
For the creation of corpus, we used the Twitter API to download the tweets. Total number of tweets provided for the data challenge was 258757. Followed by downloading of tweets, we annotated the tweets using the classes mentioned in the section below.

Annotation
Each of the downloaded tweet was classified into one of the following categories:

- **English:** Tweet is labelled as English if almost every word (i.e. $>$ 90 percent) in the tweet is tagged as En.
- **Hindi:** Tweet is labelled as Hindi if almost every word (i.e. $>$ 90 percent) in the tweet is tagged as Hi.
- **CME:** Code mixed tweet but the majority (i.e. $>$ 50 percent) of words in the tweet is tagged as En.
- **CMH:** Code mixed tweet but the majority (i.e. $>$ 50 percent) of words in the tweet is tagged as Hi.
- **CMEQ:** Code mixed tweet having an equal number of words tagged as En and Hi words respectively.
- **Code-Switched:** There is a trail of Hindi words followed by a trail of English words or vice versa. Following is the details about the data classification:

Tags	Number of Tweets
ENGLISH	5474
HINDI	72
CMH	9137
CME	187998
CMEQ	2249
CS	20962

MODELS USED

Model 1: Unique User Ratio (UUR) Model
The unique user ratio was computed using the same formula given in the task description [1], i.e.

$$UUR(w) = (U_{hi} + U_{cmh})/U_{en},$$

where *U_hi* is the number of users who have used the word w in their *Hindi* tweets at least once. Similarly, U_{cmh} and U_{en} represent the number of users who have used the word *w* in their *CMH* tweets and in *English* tweets at least once respectively.

Model 2: Unique Tweet Ratio (UTR) Model
Unique Tweet Ratio of a word *w*[1] is given by:

$$UTR(w) = (T_{hi} + T_{cmh})/T_{en},$$

where T_{hi}, T_{cmh}, and T_{en} represent the number of *Hindi* tweets, *CMH* tweets and *English* tweets in which word w is present respectively.

Model 3: UTR*UUR Model
We used the UTR and UUR values to compute the borrowed index. The candidate words were later sorted in the descending order of UUR and UTR values to get respective rank lists.

Model 4: Weighted Class Average Model
101755 tweets are filtered out of 225892 tweets where at least one of the 230 words is present. Then the tweets are classified as :

- Pure English (PE): Where the entire tweet consists of words tagged as EN, Other or NE.
- Pure Hindi (PH): Where the majority are Hindi words and consists of at most one English word
- Start English and later Hind (SE): Have a sequence of English words then followed by the Hindi Words.
- Start Hindi and later English (SH): Have a sequence of Hindi words initially followed by the English words.
- Code Switched more than twice (CS): Having a pattern of EN/HI/EN/HI or vice-versa

At most two English words (EWI): Have at most two English words in between the tweets where 95 percent of the words consists of Hindi words. So the metric is calculated using

*WCAM(word) = PH*0.25 + EWI*0.25 + (SE+SH) * 0.2 + CS*0.1 / PE*,

where the PH is the number times word is used in the PH tweet, EWI is the number times word is used in the EWI tweet, SE is the number times word is used in the SE tweet, SH is the number times word is used in the SH tweet, CS is the number times word is used in the CS tweets and PE is the number of times word is used in the PE tweets. Higher the value of the WCAM the word is more likely to be borrowed.

Model 5: Inverse Model
Similar to Model 4, the Inverse metric is calculated using the formula,

$Metric$ = 0.5 * $IUUR(w)$ + 0.5 * $IUTR(w)$, where

$IUUR(w) = 1 + log(U/UUR)$, where $U = U_{hi} + U_{cmh} + U_{en}$ and

$IUTR(w) = 1 + log(T/UTR)$, where $T = T_{hi} + T_{cmh} + T_{en}$

This metric is used to compute the borrowed index. The candidate words were later sorted in the descending order of the metric values to get respective rank lists.

Model 6: TF-IDF
Term Frequency (TF) and Inverse Document Frequency (IDF) reflects how vital a word is to a particular document or corpus. The 230 words that are given were used and ranked based on the TF-IDF metric.

Term Frequency is calculated as $1 + log(count)$ where count is number of times the word occurred in the tweets.

Inverse Document Frequency is calculated as

$$1 + log(len(tokenized\_tweets)/(sum(contains\_token)))$$

where the $tokenized\_tweets$ is the total number of tweets and $contains\_token$ is the number of times the word appears in the tweets

Model 7: Code Switched Model (CS)
As per [3], words that are part of tweets with code switching are candidates for borrowing. This model finds the ratio of frequency of words tagged as English to the tweets where code switching occurs at least once.

Code Switch Ratio of a word *w* is given by:

$$CSR(w) = N_{en}/T_{cs},$$

where N_{en} and T_{cs} represent the frequency of word tagged as *English* and the number of code switched tweets in which word w is present respectively. Code switching includes switch from *English* to *Hindi* and vice versa.

Majority-Voting Based Bagging
In [2], Dietterich et al. discussed about the philosophy of using majority voting algorithm for bagging several contender classification algorithms.

REFERENCES

1. Data Challenge CODS 2017. `https://ikdd.acm.org/cods2017/data-challenge.html`.
2. Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, Springer (2000), 1–15.
3. Gualberto A. Guzman, Jacqueline Serigos, B. E. B. A. J. T. Simple Tools for Exploring Variation in Code-Switching for Linguists. 12–20.