Code-Borrowedness of English words in Hindi Language

Ram Mohan[1], Muhammad Arif[1], Jobin Wilson[1], Santanu Chaudhury[2] and Brejesh Lall[2]

R&D Department[1]
Flytxt, India.

Department of EE[2]
Indian Institute of Technology, Delhi.

Introduction

The goal of IKDD-CoDS 2017 data challenge is to develop a metric to rank a set of candidate words according to the likeliness of them being code-borrowed from English to Hindi, by analyzing social media conversations.

Ground Truth

The user preference towards usage of an English word as opposed to equivalent Hindi word in a Hindi sentence, is determined through a survey. The ground truth ranking metric for a word is the difference between number of responses preferring English usage of a word as opposed to its Hindi usage. A sample survey question is as below.

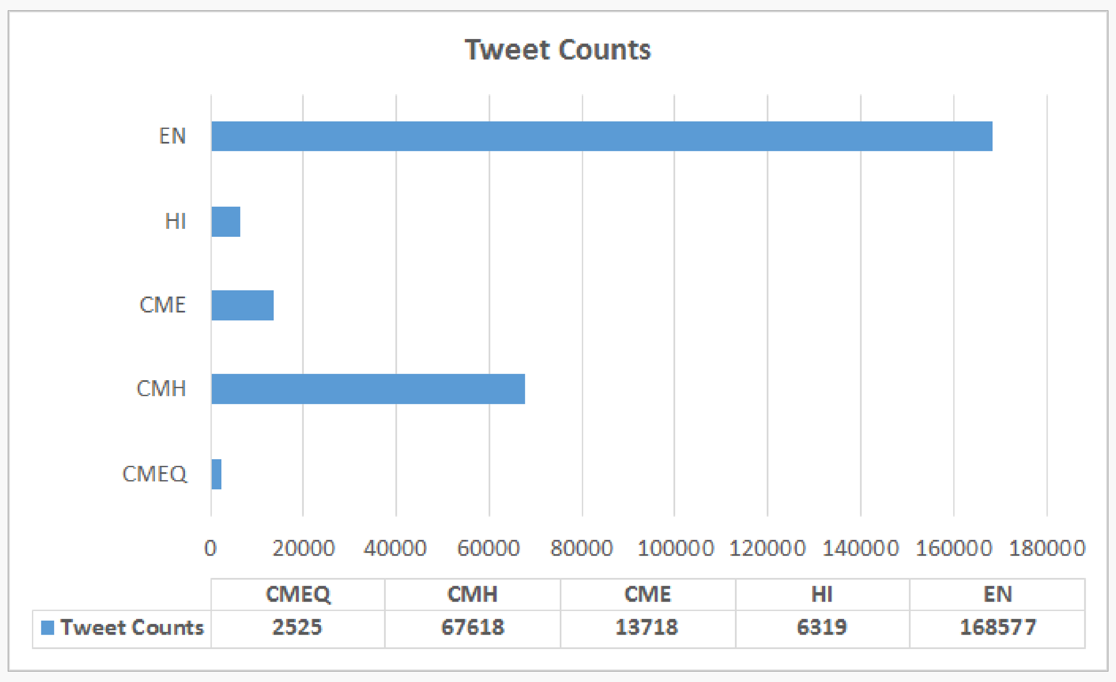
○ आपका आंसर रॉंग है
○ आपका आंसर गलत है
○ None

Dataset

A social dataset consisting of approximately 0.24 million twitter conversations are provided for estimating the likeliness of code-borrowedness. Tweets are tagged at the word level using tags such as Hindi(HI), English(EN), Named Entity(NE) and Other(OTH). Based on the word tags present within, the tweets are tagged as described in table below. Here ENC denotes English word count in the tweet whereas HIC represents Hindi word count.

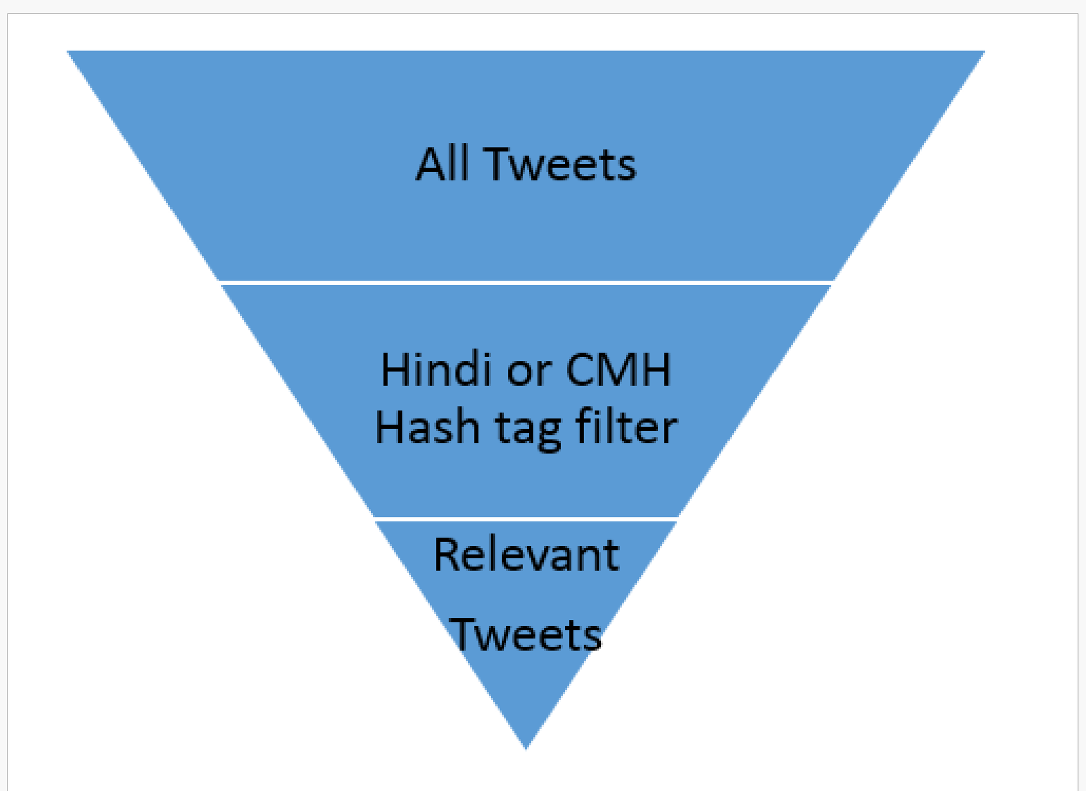
Tweet Tag	Rule
EN	ENC / (ENC + HIC) > 0.9
HI	HIC / (ENC + HIC) > 0.9
CME	ENC / (ENC + HIC) > 0.5
CMH	HIC / (ENC + HIC) > 0.5
CMEQ	ENC / (ENC + HIC) = 0.5

Distribution of tweets tagged to each category is as depicted below.

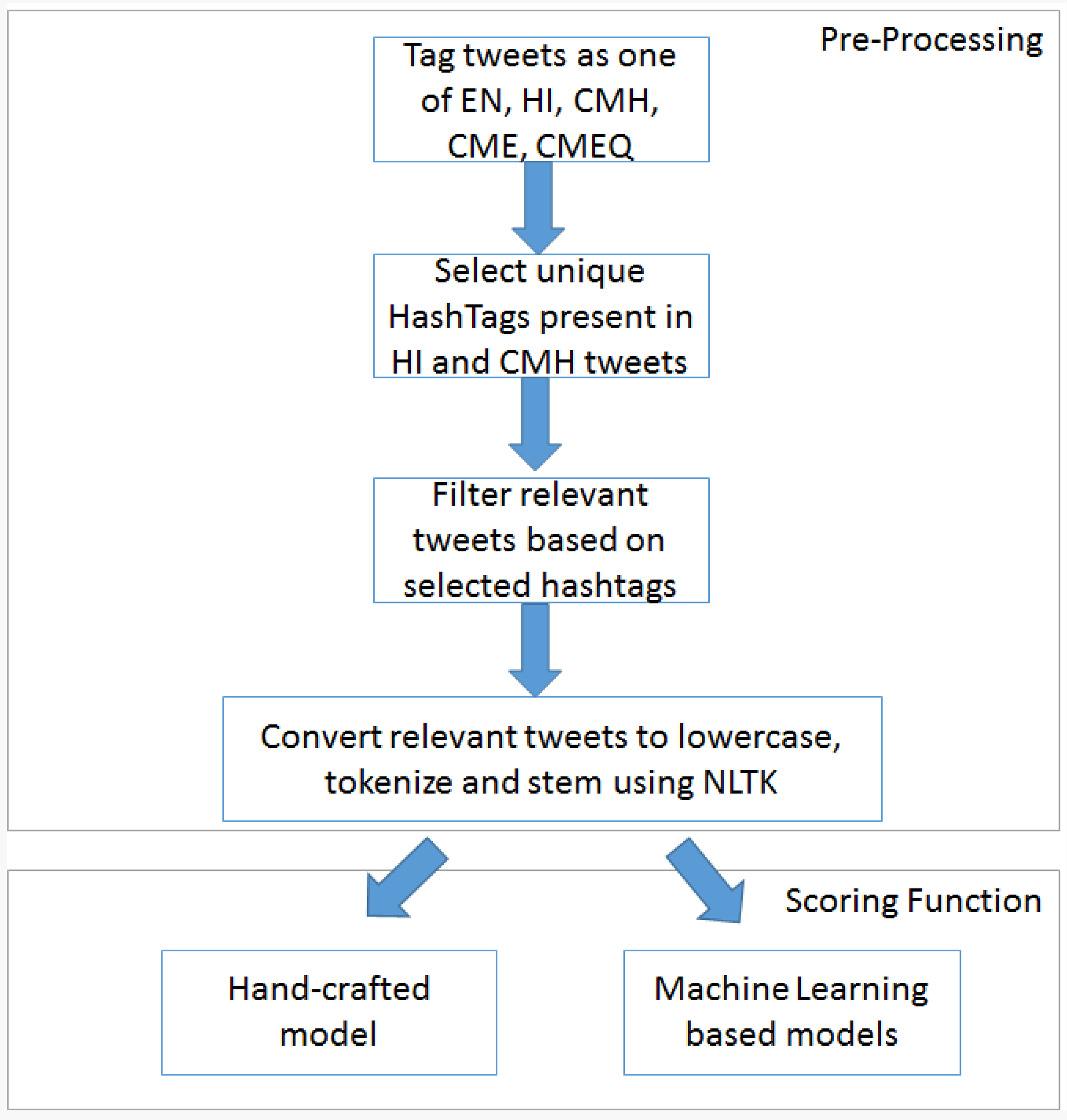


Filtering Relevant Tweets

English tweets were significantly more in the dataset as compared to Hindi and CMH tweets and hence, English usage of a word is more likely to be encountered. However, identifying relevant tweets which are more indicative of code-borrowedness is essential. Since tweets are associated with hashtags, we first identify relevant hashtags associated with tweets tagged as HI or CMH. We filter tweets with these hashtags, to generate a set of *relevant tweets*

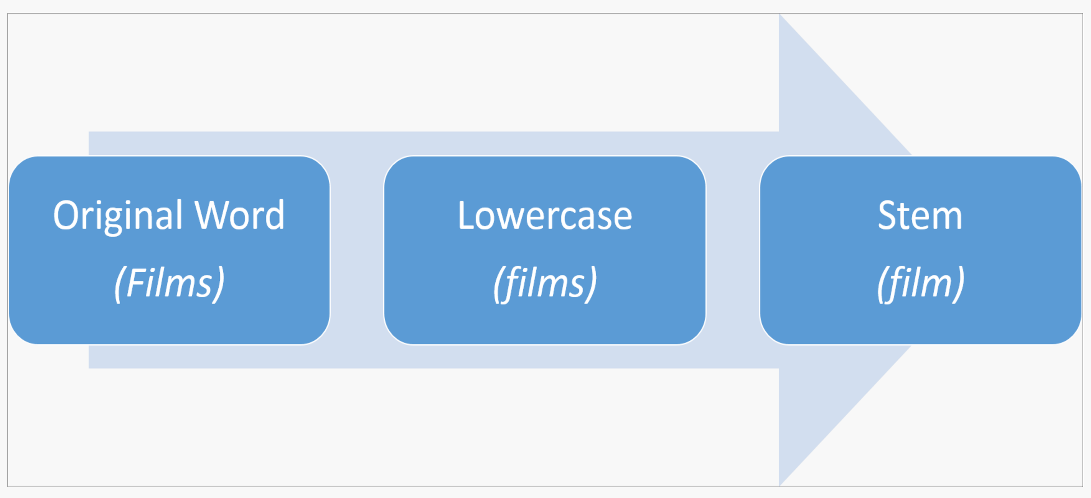


Proposed Model



Tweet Tokenization and Stemming

Several variants of a root word exists in the tweets. To derive accurate word statistics, we used stemming. For instance, our pre-processing would transform words such as film, films etc to its root form 'film'.



Features

The following features were extracted from the *relevant tweets*.

Name	Description
RTU_{hi}	Number of users who have used the root word in HI tweets
RTU_{en}	Number of users who have used the root word in EN tweets
RTU_{cmh}	Number of users who have used the root word in CMH tweets
RTU_{cme}	Number of users who have used the root word in CME tweets
RTU_{cmeq}	Number of users who have used the root word in CMEQ tweets
RTT_{hi}	Number of HI tweets containing the root word
RTT_{en}	Number of EN tweets containing the root word
RTT_{cmh}	Number of CMH tweets containing the root word
RTT_{cme}	Number of CME tweets containing the root word
RTT_{cmeq}	Number of CMEQ tweets containing the root word

Machine Learning Based Models

In our machine learning based approach, we used supervised learning algorithms for function approximation. The target variable was the ground truth metric and the input features were the ones described in the Feature list table. Our machine learning models are listed below.

- Ordinal Regression: In this model, we attempted to learn a function to directly rank words.
- Linear Regression: In this model, weights for input features were learnt and code-borrowedness score was modelled as a weighted linear combination of the input features.
- Neural-Networks: In this approach, code-borrowedness score was modelled as $y = \frac{w1'X+b1}{w2'X+b2}$.
- Non-Linear Regression: Kernalized-Ridge regression and Support Vector Regression(SVR) models were used for function approximation.

Hand-crafted Model

Our hand crafted model calculated the code-borrowedness index of a word using the formulas listed below.

$$RHTUR(w) = \frac{RTU_{hi} + RTU_{cmh}}{RTU_{en}}$$

$$RHTTR(w) = \frac{RTT_{hi} + RTT_{cmh}}{RTT_{en}}$$

$$BorrowednessScore(w) = \frac{RHTUR(w) + RHTTR(w)}{2}$$

RHTUR(w) reflects user level preferences of usage of a word in a particular language whereas *RHTTR(w)* indicates a global usage preference. Since the number of conversations from users can vary significantly in social media datasets, we factored in both these preferences to estimate the code-borrowdness..

Results

Model	Spearman Correlation Coefficient
Ordinal Regression	-0.03
Linear Regression	0.67
Non-Linear Regression	0.73
Neural-Network	0.47
Hand-Crafted	0.7915

Conclusion

In this work, we evaluated several models to compute code-borrowedness index of words from English to Hindi based on twitter conversations. We observe that identifying the relevant tweets indicative of code-borrowedness of words is essential to derive features capable of ranking words appropriately in this context. Further, the limited availability of ground truth data inhibits machine learning approaches to effectively learn a function approximation capable of predicting code-borrowedness of words. Our proposed hand-crafted model deriving BorrowednessScore(w) from simple features calculated from *relevant tweets* performed significantly better compared to the other models considered.

Source Code

The complete source code for our model and a detailed write-up is available in our github repository - https://github.com/flytxtds/ikdd-cods-2017-iitd