Supplementary Materials of: All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media

[1]Jasabanta Patro, [2]Bidisha Samanta, [3]Saurabh Singh,
[4]Abhipsa Basu,[5]Prithwish Mukherjee, [6]Monojit Choudhury, [7]Animesh Mukherjee
[1,2,3,4,5,7] Indian Institute of Technology Kharagpur, India – 721302 , [6]Microsoft Research India, Bangalore – 500033
{[1]jasabantapatro, [2]bidisha, [3]saurabhsingh, [4]abhipsabasu}@iitkgp.ac.in,
[5]pritspido@gmail.com, [6]monojitc@microsoft.com, [7]animeshm@cse.iitkgp.ernet.in

Figure 1: Some example code-mixed tweets from English-Hindi bilinguals. Hindi words are in italics.

Huge traffic restrictions for PM's visit to #blast site mean deserted roads in #Hyderabad. ''*Itna sanaata kyon hai bhai?*''
Translation: Huge traffic restrictions for Prime Minister's visit to the blast site mean deserted roads in Hyderabad. "Why is there so much silence, bro?"

MMS will go to #HyderabadBlast site to take *jayeja* of area & say *Hazaaron Jawabon Se Acchi Hai Meri Khamoshi* #ThikHai
Translation: MMS (name of a politician) will go to #HyderabadBlast site to take a survey of the area and say "My silence is better than a thousand answers." #ThikHai

A Examples of English-Hindi bilingual tweets

Figure 1 presents examples of some typical English-Hindi bilingual tweets occurring in SMC.

B Tweet type distribution in dataset

The details of the number and % of tweets falling in each of the six tweet categories introduced in section 4 of the main text are presented in table 1.

Type	Number of Tweets	percentage
En	597162	82.35
Hi	23237	3.2
CME	28702	3.95
CMH	36268	5.00
CMEQ	3173	0.44
CS	36637	5.05

Table 1: Number and percentage of tweets in each of the six categories in which the tweets are labeled.

C List of frequent 230 nouns

The list of 230 nouns that we obtain in the first step of our target word selection scheme can be found in following box.

'welfare', 'anniversary', 'tribute', 'box', 'victory', 'thing', 'lot', 'youth', 'need', 'nation', 'birth', 'people', 'muslims', 'god', 'water', 'teacher', ''airport', 'army', 'room', 'answer', 'blood', 'law', 'light', 'chief', 'green', 'office', 'border', 'food', 'university', 'side', 'event', 'health', 'reason', 'city', 'station', 'theatre', 'crore', 'ground', 'college', 'bomb', 'corruption', 'court', 'opposition', 'respect', 'life', 'air', 'rail', 'student', 'government', 'mom', 'aunty', 'weekend', 'age', 'protest', 'guy', 'company', 'bollywood', 'place', 'message', 'friend', 'mind', 'mobile', 'view', 'volunteer', 'moment', 'rest', 'suicide', 'lyrics', 'group', 'death', 'home', 'way', 'brother', 'house', 'blue', 'wedding', 'reaction', 'terrorist', 'person', 'mother', 'press', 'election', 'power', 'question', 'lord', 'birthday', 'president', 'half', 'day', 'internet', 'number', 'service', 'morning', 'waste', 'voice', 'evening', 'night', 'luck', 'son', 'favourite', 'captain', 'video', 'sun', 'body', 'experience', 'family', 'use', 'music', 'date', 'teaser', 'share', 'man', 'paper', 'lunch', 'logo', 'season', 'job', 'game', 'post', 'gift', 'poster', 'film', 'test', 'performance', 'price', 'plan', 'class', 'shot', 'report', 'prime', 'state', 'exam', 'success', 'road', 'form', 'problem', 'check', 'wife', 'boy', 'car', 'heart', 'scam', 'style', 'police', 'issue', 'card', 'country', 'boss', 'party', 'entry', 'uncle', 'politics', 'father', 'parliament', 'work', 'sunday', 'story', 'play', 'request', 'week', 'playlist', 'matter', 'superstar', 'traffic', 'suit', 'woman', 'cool', 'history', 'money', 'bat', 'seat', 'score', 'photo', 'parents', 'decision', 'girlfriend', 'picture', 'month', 'song', 'word', 'school', 'hero', 'degree', 'love', 'train', 'end', 'wrong', 'main', 'scene', 'bank', 'miss', 'king', 'channel', 'face', 'link', 'news', 'media', 'mood', 'book', 'selfie', 'bus', 'status', 'petrol', 'railway', 'budget', 'well', 'development', 'team', 'phone', 'baby', 'sir', 'interview', 'fan', 'trailer', 'year', 'girl', 'time', 'review', 'madam', 'movie', 'minister', 'joke', 'century', 'cup', 'match', 'world', 'temple', 'wicket', 'cricket', 'star'

D Grouping for target word selection

In order to select set of *target word*s we constructed a feature vector for each of the 230 nouns followed by a K-means clustering. The procedure is as follows.

Construction of feature vectors – We represent a context feature for a *target word* as a tuple $\{P_b, P_a\}$ where P_b is the language tag for the word before the *target word* (i.e., the left context) and P_a is the language tag of the word after the *target word* (i.e., the right context). Each of P_b and P_a can be either "E" indicating English, "H" indicating Hindi or "$" indicating the boundary (i.e., beginning or end) of the tweet. Thus, we have *eight* feature combinations of the left and the right contexts of a *target word* – "EE", "HH", "EH", "HE", "$E", "E$", "$H", "H$" while "$$" is not possible. For every *target word*, we compute the percentage of occurrences of each of these combinations.

Note that we compute these percentages from the three different categories of tweets – *CME*, *CMH* and *CMEQ*. Thus, for every *target word* we have a final feature vector of length 24, each entry denoting the percentage of one feature combination in a particular tweet category. We show example feature vectors for some words in figure 2.

K-means clustering – We use the feature representation of the words to cluster them into contextually similar groups. We use K-means clustering (Hartigan and Wong, 1979) for this purpose. We vary the value of K and using the traditional elbow method (Tibshirani et al., 2001) we obtain 15 as the optimal value of K. This process therefore groups the 230 nouns into 15 different clusters.

E Age-wise distribution of survey participants

The age wise distribution of survey participants are presented in figure 3. It was on this basis of this distribution that the participants were classified into two classes (young and old).

F Ranked order histograms of *target words*

The ranked order histograms of 57 candidate words according to LPF, UUR and the baseline metric are given in Figure 4.

Figure 2: Stacked plot representing feature vectors of four different words. Note that the feature vectors of the word pairs (i) "job" and "film" and (ii) "moment" and "protest" are very similar. The fractional counts of the eight combinations for each tweet category should sum up to one; since there are three tweet categories so the total size of the stacked plot is three.

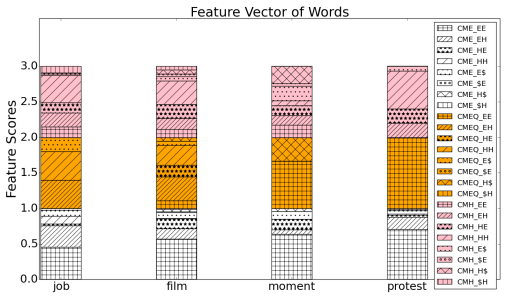


Figure 3: Age distribution of the survey participants.

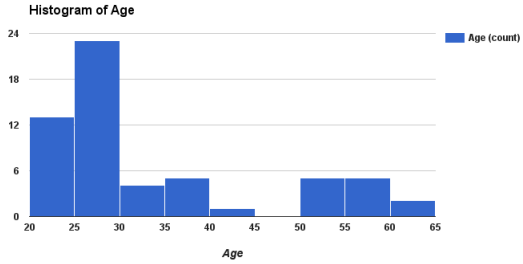
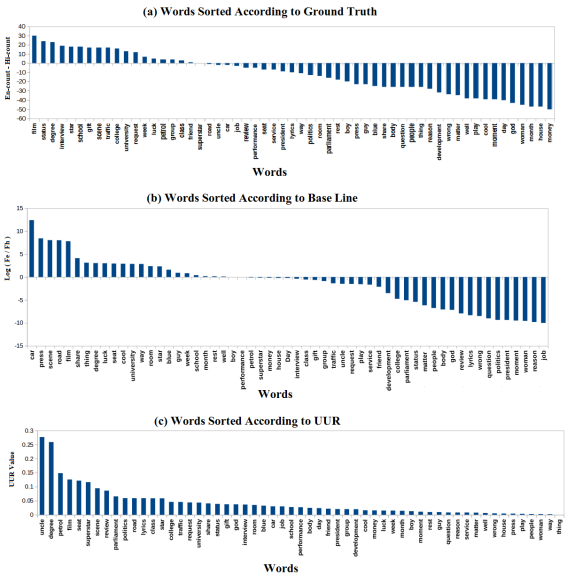


Figure 4: The rank ordered histograms of *target words* ranked by various metrics.



References

J. A. Hartigan and M. A. Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.