

Social, economic and demographic factors drive the emergence of Hinglish code-mixing on social media (Supplementary Information)

Ayan Sengupta¹, Soham Das³, Md. Shad Akhtar³, and Tanmoy Chakraborty^{*1,2}

¹Department of Electrical Engineering, Indian Institute of Technology Delhi, New
Delhi India – 110016

²Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New
Delhi India – 110016

³Department of Computer Science and Engineering, Indraprastha Institute of
Information Technology, New Delhi, India – 110020

1 Dataset

1.1 Tweet Data

Approximately 260k tweets were crawled from 2014 to 2022 using a Twitter API and a set of Boolean queries to extract tweets on popular topics in the Indian context, such as cricket, Bollywood, and politics. Cricket, a popular sport in India, was chosen as a topic of interest, along with Bollywood (the film industry based out of Mumbai) and politics. The Twitter API calls were invoked every day between 1st January, 2010 and August, 2022. However, no tweets were returned for the 2010-2013 period. The data returned by the API contained the date and time when the tweet was published, the author ID, the user ID, the tweet text, a “lang” property denoting the language identified in the tweet by Twitter’s language identification model, and other properties left unused in our work.

1.2 Exogenous Factors

A set of 1442 socio-economic indicator features were extracted for India from the World Bank data bank¹ collected yearly from 1960 till 2021. The values for the year 2014 onwards were used in our work.

2 Methodology

2.1 Feature Selection

We inspected Spearman’s rank correlation with the target values to select a subset of exogenous features from the 1442 feature set. First, we created three sets of features, each with the absolute value of Spearman’s rank correlation greater than 0.75 with at least one of the three temporal population trends. This gave us 10 features as shown in Table 1. Their Spearman’s rank correlation statistic and p -values are noted in Table 2. Since the input size was small (< 10 years of data), we used a permutation test to calculate the p -value instead of reporting the asymptotic p -value.

*Corresponding author: tanchak@iitd.ac.in

¹<https://data.worldbank.org/country/india>

Table 1: Descriptions for the exogenous features.

Feature Index	Feature Description
1	Agriculture, forestry, and fishing, value added (% of GDP)
2	Discrepancy in expenditure estimate of GDP (current LCU)
3	General government final consumption expenditure (% of GDP)
4	Gross domestic savings (current US\$)
5	Industry (including construction), value added (current US\$)
6	Manufacturing, value added (current US\$)
7	Net bilateral aid flows from DAC donors, United Kingdom (current US\$)
8	Net secondary income (Net current transfers from abroad) (current US\$)
9	Taxes less subsidies on products (current US\$)
10	Wholesale price index (2010 = 100)

Table 2: Spearman’s rank correlation of features with the population trends.

Feature Index	Hindi		English		Code-mixed	
	Correlation	<i>p</i> -value	Correlation	<i>p</i> -value	Correlation	<i>p</i> -value
1	-0.214	0.630	-0.595	0.128	0.762	0.034
2	0.762	0.040	0.238	0.584	-0.714	0.054
3	0.214	0.623	-0.929	0.003	0.786	0.030
4	0.833	0.015	-0.643	0.093	0.119	0.798
5	0.810	0.024	-0.667	0.084	0.190	0.656
6	0.810	0.021	-0.667	0.082	0.190	0.674
7	-0.786	0.022	0.595	0.139	-0.048	0.941
8	-0.071	0.883	-0.738	0.048	0.833	0.016
9	0.810	0.022	-0.667	0.082	0.190	0.645
10	0.095	0.823	-0.857	0.011	0.857	0.009

2.2 Fisher Jenks Algorithm

The Fisher-Jenks algorithm is a widely used and effective technique for data classification, especially when dealing with continuous data. The algorithm works on one-dimensional data only. It starts by sorting the data values in ascending order. Then arbitrary breaks are introduced in the data to form clusters, and the sum of within-class variances is calculated. The aim is to find the set of breaks giving the least sum of within-class variances, and this is done with an iterative approach of choosing different breaks in every iteration. In this way, a set of breaks with a minimal within-class variance is obtained.

2.3 PoS tag analysis

PoS tagging is a natural language processing task that assigns the words in a text corpus to categories based on the parts of speech they correspond to. For every word, the meaning and context of its usage in the sentence are considered to assign it a PoS tag. The complete list of PoS tags can be found at the Universal PoS tags website². The subset of tags used in our analysis, along with their definitions and usage examples, is provided in Table 3.

²<http://universaldependencies.org/u/pos/index.html>

Table 3: PoS tags, definitions, and examples.

PoS tag	Full form	Definition	Example Usage
ADJ	Adjective	A word describing a noun or pronoun	He is a dutiful husband.
ADP	Adposition	A word that shows a relationship between a noun or pronoun and other words in a sentence.	He ran towards the building.
ADV	Adverb	A word describing a verb, adjective, or another adverb	She fought ferociously .
CONJ	Conjunction (coordinating or correlative conjunction)	A word or words joining two words, phrases, or clauses of equal syntactic importance	The gardener went to the lawn and trimmed the shrubs.
DET	Determiner	A word determining and introducing a noun or noun phrase	The plane landed properly.
NOUN	Noun	A word that names a person, place, idea, or thing	I read the newspaper every day.
NUM	Numeral	A word expressing a number or quantity	I have spent a hundred hours on this.
PART	Particle	A function word that must be associated with another word or phrase to impart meaning and that does not satisfy definitions of other universal parts of speech. For each language, there is a very restrictive set of words labeled as particles.	I want to party.
PART_NEG	Particle	A particle that indicates negation	The food was not good.
PRON	Pronoun	A word that takes the place of a noun or noun phrase	He is doing just fine.
PRON_WH	Wh-pronoun	A pronoun introducing wh- questions such as "what", "why", "how" etc.	Why would you say that?
PROP	Proper noun	A noun for a specific place, person, or thing.	I live in India .
VERB	Verb	A word expressing an action, occurrence, or state of being.	When you see them, you just run .
X	Other	A word that is unintelligible or does not fit in any category	Hmmm , that's interesting...