Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03058-6 OPEN

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

Ayan Sengupta[1], Soham Das[2], Md. Shad Akhtar[2] & Tanmoy Chakraborty[1,3]✉

The advent of globalization and adaptation to multiple cultures has emanated a fusion of Hindi and English, casually known as *Hinglish*. The phenomenon of mixing multiple languages (such as Hindi and English) within a single utterance is often called *code-mixing*. Lately, code-mixed Hinglish has emerged as a dominant conversational language for Hindi-speaking citizens both online (on social media platforms) and offline. Although previous studies investigated such linguistic traits of Hinglish over the past few years, some pertinent questions still need to be answered: How did Hinglish evolve? And, what are the factors behind the evolution of Hinglish? Does the fusion of English impact all Hindi words similarly? To this end, we explore the empirical and statistical shreds of evidence behind the rise of Hinglish on social media such as Twitter. We show that adopting Hinglish depends on several socio-economic and demographic factors. We further formulate dynamic models to explore the socio-economic factors driving the growth of Hinglish, derive the future growth of Hinglish in the upcoming years, and estimate the propensity of users to change their linguistic preferences. Our study highlights that the Hinglish population has evolved steadily between 2014 and 2022, with an annualized growth rate of 1.2%, and the usage of Hinglish on Twitter has increased annually by 2%. Further, we find that the impact of Hinglish evolution is not uniform across different word groups and affects the contextual meaning of different words differently. Although our findings are specific to the Indian Hinglish community, our study can be generalized to understand the evolution and dynamics of other code-mixed languages, such as Spanish-English or Chinese-English.

[1] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India. [2] Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi 110020, India. [3] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi 110016, India. ✉email: tanchak@iitd.ac.in

Introduction

India is a land of diversity. Diverse linguistic origins and diverse ecology have given birth to over 700 languages in India (Gazette 2014), out of which over 100 are spoken as the mother tongue by natives. As different cultures interchange, they give rise to multilingualism (Mallikarjun 2019). In fact, according to the 2011 census, over 26% of the Indian population is multilingual (TOI 2010). Spoken by over 500 million Indians, Hindi is among the most popular choices for bilingual and multilingual speakers. Hindi is also a dominant global language, the third most-spoken language worldwide with 602 million speakers (counting both first and second-language speakers) (International 2022; Barath 2019). An undeniable impact of British colonization in India is the interface between Hindi and English (Annamalai 2004). English was introduced in India by the British primarily for administrative and educational purposes. In the early days of English adoption, the elitists mainly adopted it as a symbol of modernity and societal status. Over the years, Indians have owned English in their vernaculars, amalgamating English and other Indian languages. This hybridization has also impacted the linguistics and morphology of both foreign and domestic languages. There are several instances of English words originating from Hindi or Urdu, such as 'chutney' (a thick sauce of Indian origin that contains fruits, vinegar, sugar, and spices, which is used as a condiment), 'khaki' (a light yellowish-brown cloth usually made of cotton or wool), and 'mantra' (a mystical formula of invocation). Similarly, loanwords can also be found in Hindi that originated from English : 'botal' (bottle), 'kaptaan' (captain), and 'tamatar' (tomato).

This cultural fusion has manifested in mixing Hindi and English, giving birth to a new hybrid language of 'code-mixed Hinglish.' In general, code-mixing (aka code-switching) (Nilep 2006) is a phenomenon that occurs when two or more languages are used together in a single utterance. It is a commonly observed phenomenon in multilingual societies and is widely spread in many language pairs such as Hinglish (Hindi-English) (Nema and Chawla 2018; Parshad et al. 2016), Spanglish (Spanish-English) (Otheguy and Stern 2011), Dutch-English (Roelofs 2019), and Chinese-English (Zhang 2012).

Several linguistic theories have been proposed (Poplack 1980; Sankoff and Poplack 1988) to understand the emergence of code-mixed languages from monolinguals. Generally, code-mixing emerges from a dominant language (aka matrix language) and a non-dominant foreign-embedded language (Myers-Scotton et al. 2002). In most cases, code-mixed texts follow the matrix language's syntactic structure.

Example 1: *tumhara problem kya hai?* (Hinglish)
"What is your problem?" (English)
Example 2: *I don't like daal chawal.* (Hinglish)
"I don't like rice and curry." (English)
Example 3: *Dr. Santosh ji ka* आशीवार्द प्राप्त हुआ (Hinglish)
"Received the blessings of Dr Santosh ji." (English)

As shown in example 1, Hindi is the matrix language, and English is the foreign-embedded language. On the other hand, in example 2, English is the matrix language, and Hindi is the embedded language. Being an informal conversational language, code-mixing can be observed in many conversation settings, a popular one being online social networks (Yusnida et al. 2022). Due to ease of access, most users prefer using romanized scripts for written communications, leading to *script-mixing*. Transliterations in script-mixing have become a common way to write Hinglish on online platforms, and it has allowed users to express themselves in their preferred language, regardless of the script. In example 1, we find the Hindi words *'tumhara'*, *'kya'*, *'hai'* transliterated to Romanized script. On the other hand, in example 3, both Romanized and Devanagari scripts are used for the Hindi text. These diverse linguistic properties of code-mixed languages need to be studied to understand the emergence of derived synthetic languages from natural languages.

Several computational studies (Srivastava et al. 2020; Pratapa et al. 2018) have been conducted to understand how code-mixed texts are formed. Computational linguists have also attempted (Bhatia and Ritchie 2016; Nema and Chawla 2018; Mabule 2015; Thara and Poornachandran 2018) to understand the linguistic variabilities of Hindi-English code-mixed language. Significant efforts have been made for building computational systems for several applications; sentiment analysis (Joshi et al. 2016), parts-of-speech tagging (Singh et al. 2018), named-entity recognition (Priyadharshini et al. 2020), hate detection (Sreelakshmi et al. 2020), and sarcasm detection (Aggarwal et al. 2020) from Hindi-English code-mixed texts. Notable contributions have also been made to develop computational systems (Chakravarthi et al. 2022; Wang et al. 2018) to solve semantic, syntactic, and pragmatic tasks for other code-mixed languages.

Like any other language, code-mixing as a language shows evolutionary traits. Societal, cultural, and evolutionary changes drive the evolution of human languages. Kothari and Snell (2011) delved into the history and origin of the Hinglish language, and the different extents of its integration into various geographical, cultural, and social strata. The authors also explored the potential of Hinglish to grow further as a regular conversational instrument among Indians and in Indian media. Today, Hinglish has become a popular choice of communication for the young Indian population from different demographic and cultural backgrounds. Not only within India, Hinglish has also emerged as a popular language among the Indian diaspora in the US and the UK (Baker 2015). Code-mixing is a phenomenon that has been shown to evolve, bringing changes in the interacting languages and the populations speaking them. The dynamics of such a system can be understood through language competition models, which establish a relationship among the fractions of the population using each participating language. In this work, we answer the following questions: What are the empirical and statistical pieces of evidence of the evolution of Hinglish code-mixing in social networks? Does code-mixing have a similar impact on different word groups? What are the drivers behind the evolution of Hindi-English code-mixing? And, can they be used to predict the adaptation of Hindi-English in the upcoming years?

To our knowledge, ours is the first large-scale computational study addressing the evolution of Hinglish code-mixed language and the drivers behind the evolution. We conduct thorough empirical and statistical analyses to understand the evolution of the Hinglish code-mixed language and the evolutionary dynamics of different linguistic groups. Beyond analyzing Hinglish's current prevalence and characteristics, we look at the phenomenon's temporal evolution in the social media context for Indian users. Social media users often use code-mixing to engage a wider audience and appear relatable. We hypothesize that the use of code-mixing has followed a trend over the years, correlating to the trend of other socio-economic factors. We attempt to find and analyze the strongest of such connections. Our analyses are based on a dataset we collected from Twitter, spanning 2014–2022. Twitter (Twitter underwent rebranding in July 2023 and has since been known as X), one of the most popular online forums among the young urban population, is a medium where users prefer hybrid code-mixed language over monolinguals for a wider audience and for expressing creativity and humor. We chose our population of study from Twitter and collected 262, 578 tweets posted by 16, 710 unique Twitter handles between January 2014 and September 2022. Unlike existing Hinglish code-mixed corpus, we retain the user information (although anonymized) to understand the micro traits of code-mixed evolution. Using a list of annually recorded socio-economic and demographic features of India, we conduct correlation analysis with these trends and create a dynamic model that can predict the growth or decline of code-mixing in the coming years, given certain constraints. Instead of modeling the language evolution purely using the ordinary differential equations (Parshad et al. 2016) or game-theory-based approaches (Nowak and Krakauer 1999), we adopt an econometric technique on the available proportional data while including the exogenous socio-economic and demographic variables recorded annually. Our proposed dynamic model considers the dynamics among Hindi, English, and Hinglish languages and captures the dependence of different exogenous socio-demographic and economic macro features on language evolution. We conduct linguistic analysis to understand the impact of code-mixing on different words their linguistic properties and how they evolve (Srivastava et al. 2020). Our analyses conclude that Hinglish adoption has evolved consistently between 2014 and 2020 with an annualized growth rate of 1.2% and will grow beyond 2023 at an even higher rate of 2.98%. Artifacts collected in this work can aid in computational research on code-mixed language. We also hypothesize and, through our analysis, highlight the role of Bollywood films and actors in the widespread adoption of Hinglish in the Indian community. Code-mixing affects the hierarchy of society differently and, therefore, emerges as a personalized language instead of a language with universal recognition. Our current work sheds light on these aspects and leaves room for researchers to explore more about the personalization of this cultural phenomenon. Insights from our work can also aid in building conversational systems – chatbots and virtual assistants in code-mixing languages. These applications can leverage the code-switching patterns found in real-life conversations to break the language barrier and reach wider audiences. Our work shows the existence of an evolutionary trend in code mixing in the Indian social media context and extracts several patterns about it. It provides arguments for the need to work with newly collected data to capture the ever-evolving semantics of the Hinglish language in natural language models. From our results, we argue the inability to treat natural language processing (NLP) tasks on code-mixed data the same as the tasks on monolingual corpora. We find various characteristic patterns of the Hinglish phenomenon, which are integral to determining and predicting the semantics of the language. Our work is not a comprehensive analysis of the language but provides a basis for deeper analysis that new research works on code-mixed datasets should perform to eliminate erroneous, outdated assumptions of older research and models.

Related work

In one of the earliest studies on code-mixing, Joshi 1982 explained how the dynamics between two grammatical systems lead to the hybrid language of code-mixing. Motivated by this study, Myers-Scotton 1997 proposed a matrix language frame (MLF) model, which theorized the effect of the matrix language and embedded language in forming code-mixed language. Other popular theories on code-mixed language formation include equivalence constraint (EC) (Poplack 1980) and functional head constraint (Di Sciullo et al. 1986). The EC theory expands the concept of code-mixing beyond lexical substitutions and highlights the roles of context-free grammars of the matrix and embedding languages. Several attempts (Bromham et al. 2015; Nowak and Krakauer 1999; Parshad et al. 2016; Abrams and Strogatz 2003; Nie et al. 2013; Walters 2014; Patriarca and Heinsalu 2009) have been made to study the effect of exogenous socio-linguistic and geographical features on the spread of code-mixing and language competition phenomena. Variations can also be observed in the amount of code-mixing observed in utterances on average. Nema and Chawla 2018 discussed the adoption and impact of Hinglish in Indian media, its causes and effects, and various sociological aspects. Analyses of different modes of code-switching have been explored in Bollywood scripts (Si 2011). The language dynamics of Hindi and English were discussed via several models and assumptions, such as ordinary differential equation (ODE) and partial differential equations (PDE) models (Parshad et al. 2016), reaction-diffusion models (Walters 2014), and control theoretical approaches (Nie et al. 2013). Despite these linguistic and computational studies, there has been no large-scale empirical study on the evolution of Hinglish code-mixing that unearths the socio-demographic aspects behind code-mixing. Moreover, most existing studies on the emergence of code-mixed language are based on social interactions and, therefore, overlook the impact of social media on linguistic evolution. We attempt to model the evolution of Hinglish code-mixing on both linguistics and exogenous socio-economic and demographic factors. In contrast to existing studies, we conduct our analyses on curated social media data instead of census data, allowing us to perform both aggregated and fine-grained studies. Our study aims to shed light on the language dynamics in the Indian social media space, where people might be motivated to present and converse differently than in other day-to-day interactions, given different sizes of audiences.

Methods

Dataset collection and labelling. To understand the temporal trend of linguistic preferences of Indians on Twitter, we used Twitter academic API (Twitter 2023) (Application Programming Interface) to fetch 260k tweets related to 'Cricket', 'Bollywood', and 'Politics' from 2014 (see *SI Appendix*, Section 1.1). We did not have any specific criteria for choosing the keywords for our queries, except for using terms and themes widely discussed in the Indian social media context. Cricket, for example, has a much wider consumer space in India than most other sports (Economist 2014). We filtered tweets posted by users in the Mumbai and Delhi metropolitan regions, two regions with predominantly Hindi-speaking populations. While multiple Indian states satisfy the previous condition, we arbitrarily chose Delhi and Mumbai to obtain a sizable number of tweets as representatives of the Hindi, English, and Hinglish-speaking populations in the Indian social network scenario. Starting from the year 2014, Twitter automatically tags the language of a tweet. We considered tweets with tagged language ID 'hi' (Hindi) and 'en' (English). This led us to 2,62,578 tweets from 16,710 Twitter handles (users), averaging ~16 tweets per user. We used a pre-trained language model open-sourced with (Sagorsarker 2020) to perform word-level language identification and parts-of-speech (PoS) tagging (see Section 2.3 and Table 3 of *SI Appendix*). A popular metric to quantify the degree of code-mixing is the code-mixing index (CMI) (Gambäck and Das 2014). For a text with n tokens (including Hindi, English, language-invariant words, and symbols) with n_{hi} Hindi and n_{en} English words, we defined the code-mixing index (CMI) as:

$$CMI = 1 - \frac{\max(n_{hi}, n_{en})}{n}. \tag{1}$$

CMI quantifies the extent of code-mixing in the text; having a higher CMI indicates more mixing patterns in a text. Therefore, a text with an equal number of Hindi and English words has a

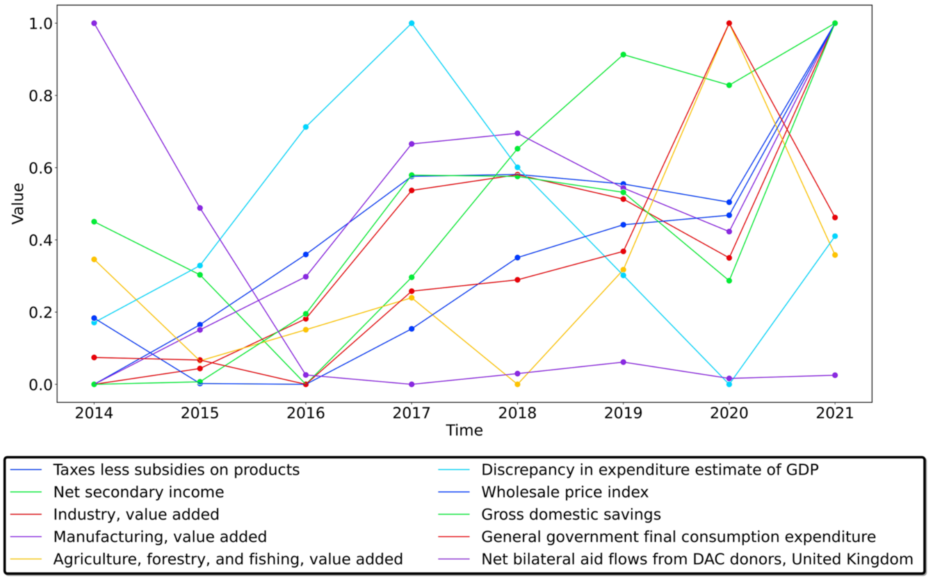

Fig. 1 Socio-economic feature values between 2014 and 2021. All the features are scaled between (0, 1).

higher CMI than a text skewed more toward any particular language. We adopted the following logic to determine if a text is written in monolingual Hindi, English, or Hinglish.

$$\begin{cases} \text{Monolingual Hindi} & \text{if } CMI<0.5 \text{ and } n_{hi}>n_{en} \\ \text{Monolingual English} & \text{if } CMI<0.5 \text{ and } n_{en}>n_{hi} \\ \text{Hinglish code-mixed} & \text{if } CMI \geq 0.5 \end{cases}$$

To determine the linguistic preference of a user, we adopted a similar logic, where we calculated the mean CMI and the total number of Hindi words and English words used by the user in a quarter.

Dynamic model. To understand the trends expressed in the monolingual Hindi (*hi*), monolingual English (*en*), and Hindi-English code-mixing (*cm*) populations (fraction of total population) and the role of various socio-economic factors in influencing these trends, we developed a dynamic econometric model. Towards this, we curated 1,442 countrywide socio-economic indicators (*SI Appendix*, Section 1.2) between 2014 and 2022. We selected 10 features based on Spearman's rank correlation with the three population trends (see Section 2.1 and Tables 1 and 2 of *SI Appendix*). The filtered indicators directly or indirectly influence the living standards, internet activity, and other behavioral traits of the Indian populace (*SI Appendix*, Table 1). Rather than the actual values of the exogenous variables, we hypothesize that the rate of change in the linguistic populations is influenced by the temporal trends captured in these variables (see Fig. 1).

For the three population ratios represented by *hi*, *en*, and *cm* respectively, we define the relationship as

$$\begin{aligned} \frac{\partial hi}{\partial t} &= en \cdot p_{en\rightarrow hi} + cm \cdot p_{cm\rightarrow hi} - hi \cdot (p_{hi\rightarrow en} + p_{hi\rightarrow cm}) + b_{hi} + \Delta X_t \cdot W_{hi} \\ \frac{\partial en}{\partial t} &= hi \cdot p_{hi\rightarrow en} + cm \cdot p_{cm\rightarrow en} - en \cdot (p_{en\rightarrow hi} + p_{en\rightarrow cm}) + b_{en} + \Delta X_t \cdot W_{en} \\ \frac{\partial cm}{\partial t} &= en \cdot p_{en\rightarrow cm} + hi \cdot p_{hi\rightarrow cm} - cm \cdot (p_{cm\rightarrow en} + p_{cm\rightarrow hi}) + b_{cm} + \Delta X_t \cdot W_{cm} \end{aligned} \tag{2}$$

Here $p_{i\rightarrow j}$ is the transition probability from population i to population j. Therefore, it follows, $\sum_{j=hi,en,cm} p_{i\rightarrow j} = 1$, 7D1 $\forall\, i \in \{hi, en, cm\}$. The dynamic system calculates the yearly rate of change in each population. Using the transition probabilities, we determined the inflow and outflow rates of each population. The bias terms b_{hi}, b_{en}, and b_{cm} denote the prior probabilities of each user being in Hindi, English, and Hinglish populations, respectively. For all the exogenous features (see Section 1.2 of *SI Appendix*), we calculated the year-wise rate of change, denoted by ΔX. The weight vectors W_{hi}, W_{en}, and W_{cm} denote the importance of each of these features on the Hindi, English, and Hinglish populations, respectively.

We assumed that the temporal trend of the population fractions only depends on the current fractions and the change in the exogenous features. The inflow of the net Indian population (combining Hindi, English, and Hinglish) on Twitter is captured through ΔX_t. Using the total number of Twitter users per year as the actual population count does not work for two reasons. Firstly, the net population of India will directly depend on the birth, death, and migration rates. However, the number of Twitter handles extracted in our analysis depends on the subset of tweets returned by our API. Secondly, the number of Twitter handles available in a particular year would also not represent the Indian population. Hence, we worked with only the fractional populations to isolate a study on the relative trends of the three linguistic categories. The method used for identifying a particular Twitter handle as code-mixed, monolingual English, or monolingual Hindi for a year is independent of the volume of tweets extracted.

We trained our dynamic system with ordinary least square (OLS) regression. As exogenous features are of different scales, we standardized (scikit-learn developers 2023) these variables based on their values between 2014 and 2021. Using 2014 as the base year $t = 0$, we calculated the proportion of Hindi, English, and Hinglish population in 2015 using the dynamic system defined in Equation (2). Henceforth, we calculated the population proportions in the subsequent years 2016 to 2022, which are used to fit the system parameters. Using our dynamic model and assuming that the exogenous factors stay the same in future years as in the last recorded year in our dataset, we predicted the future population trends for the three language-specific populations.

Word representation model and retention rate. We obtained a vectorized representation of each word using the context-based language model, Word2Vec (Mikolov et al. 2013). Word2Vec uses neural architecture to learn representations of words from each sentence that are algebraically viable. We used the Gensim

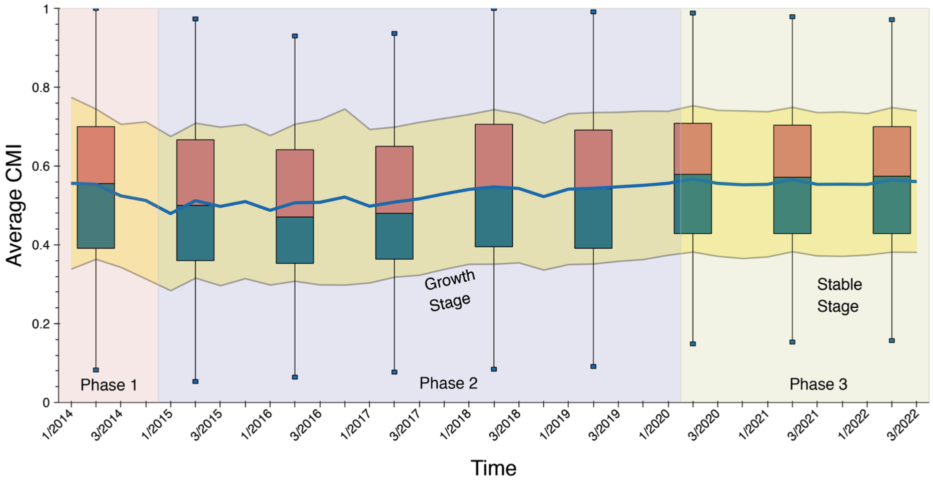

Fig. 2 Distribution of code-mixed index (CMI) calculated on tweets collected from the Indian Hindi-speaking population between 2014 and 2022. We highlight three different temporal phases based on the trend in CMI. The shift in CMI distribution also shows how the perception of code-mixing has shifted at the population level.

toolkit (Řehůřek 2022) to train the Word2Vec model on tweet corpora collected each year. We obtained a 100D vector representation for each word in every year between 2014 and 2022. To define each word's context, we used a window of size 4 (i.e., considering all 2-hop neighbors). To avoid overfitting the models, we considered only the words that appear at least 10 times in the corpus. The contextual similarity between any pair of words can be calculated using the dot product between their vectorized representations. We defined the neighborhood of each word by considering the closest 25 nodes. For each word w and its neighbor $N_w^{(t)}$ in tth year, we defined 'retention rate' as:

$$Ret_w^{(t)} = \frac{|N_w^{(t)} \cap N_w^{(t+1)}|}{|N_w^{(t)}|}. \quad (3)$$

According to the principles of contextualized word representation (Mikolov et al. 2013; Liu et al. 2020; Sezerer and Tekir 2021), the meaning of a word can be determined based on its context, i.e., its neighboring words. Therefore, a word that does not change its neighborhood structure in subsequent years has a higher retention rate.

Results

Evidence of Hinglish evolution. We conducted regression analyses on CMI calculated on tweets posted by different users. We highlighted the year-wise CMI distribution and the temporal drift in median CMI over the years in Fig. 2. The drift in year-wise CMI distribution suggests splitting the entire period into several smaller phases (time periods with consistent trends). We adopt the Fisher-Jenks break point algorithm (*SI Appendix*, Section 2.2) to split the entire period details into three phases, with the first phase spanning January 2014 to December 2014 (containing 1502 tweets with a median CMI of 0.52). The second phase spanned January 2015 to March 2020 (43,877 tweets with median CMI 0.53), and the third phase spanned April 2020 to September 2022 (217, 120 tweets with a median CMI 0.57). We observe that median CMI increased 0.2% per year between 2014 and 2022 with an intercept (bias) of 0.506 (see Table 1a). In phase 2, CMI increases significantly with an annualized growth rate of 1.2%. We observed the goodness-of-fit of the regression model with a high adjusted R^2 of 0.755 and F-statistics of 59.41 with $p < 0.001$. On the other hand, after 2020, CMI stabilizes, which can be empirically justified with a low adjusted R^2 and a slope of 0.00.

Based on the CMI value, we divided the tweets into three categories: Monolingual Hindi, Monolingual English, and Hinglish Code-mixed. The overall trend in terms of the usage of Hinglish code-mixing language shows a similar upward trend (see Fig. 3A). The yearly number of tweets has increased 12-fold in the third phase i.e., after 2020. The proportion of usage of code-mixing has increased from 42% to 60% between 2015 and 2020. After 2020, however, the usage of code-mixing has remained stable at ~60%. The use of monolingual Hindi and English steadily decreased between 2015 and 2022. Table 1b highlights the results of the regression study to quantify these growth rates further. Between 2014 to 2022, usage of monolingual English on Twitter among Indians has decreased at a steady rate of 1.2% per year. On the other hand, code-mixing has increased with a growth rate of 2%. Compared to these, monolingual Hindi usage has remained almost constant, with a prevalence of 26.6%. For each user, we computed the total number of monolingual and code-mixed tweets, based on which we determined the most popular language for each user. Similar to the previous analyses, we analyzed the users and the evolution of their linguistic preferences over the years (see Fig. 3B). It shows the quarter-wise trend in the number of users preferring monolingual Hindi, monolingual English, and Hinglish in their communication on Twitter. Based on regression analysis (highlighted in Table 1c), we conclude that Hinglish has always been the most popular choice mode of written communication among Indian Twitter users (44.9% preferred Hinglish in 2014). The proportion of users preferring Hinglish has increased to 56.3% after 2020 with a steady growth rate of 1.2% year-wise. On the contrary, the proportion of users preferring monolingual English has decreased steadily from 23.3% to 11.2% with a rate of 1.6%.

Linguistic changes due to Hinglish evolution. In the previous subsection, we highlighted the empirical evidence behind the evolution of Hinglish code-mixed language in terms of overall usage, linguistic preferences of users, and the CMI value. To understand the extent of the impact of English inclusion on Hindi, we conducted a fine-grained analysis of different Hindi words. We observed the popularity of Devanagari scripts on Twitter, growing from 35% in 2014 to 82% in 2022. Adverbs like आज (today), अब (now) are more likely to be used in Devanagari than in Romanized scripts. We conducted fine-grain analysis on the top Hindi and English words used in tweets over the past few

Table 1 Regression analyses in different time Phases.

Phase	Time Period	Adjusted R^2	F Statistic	Regression Coefficients	
				Sl.	Inter.
Phase 1	1/2014 – 4/2014	0.871	21.290	−0.016	*0.561
Phase 2	1/2015 – 1/2020	0.755	*59.410	*0.003	*0.490
Phase 3	2/2020 – 3/2022	−0.125	0.003	0.000	*0.558
Overall	1/2014 – 3/2022	0.473	*31.560	*0.002	*0.506

(a) Regression analysis result on CMI.

Phase	Time Period	Adjusted R^2			F Statistics			Regression Coefficients					
		Hi	En	CM	Hi	En	CM	Hi		En		CM	
								Sl.	Inter.	Sl.	Inter.	Sl.	Inter.
Phase 1	1/2014 – 4/2014	0.772	0.126	0.839	11.160	1.430	16.670	0.030	*0.171	0.014	*0.149	−0.039	*0.628
Phase 2	1/2015 – 1/2020	0.077	0.752	0.745	2.680	*61.520	*59.510	−0.002	*0.294	* – 0.005	*0.192	*0.008	*0.459
Phase 3	2/2020 – 3/2022	−0.024	0.540	0.048	0.788	11.550	1.456	0.001	*0.244	−0.002	*0.092	0.002	*0.630
Overall	1/2014 – 3/2022	−0.023	0.793	0.532	0.222	*131.600	*39.680	0.000	*0.267	* – 0.003	*0.190	*0.005	*0.491

(b) Regression analysis results on the number of tweets.

Phase	Time Period	Adjusted R^2			F Statistics			Regression Coefficients					
		Hi	En	CM	Hi	En	CM	Hi		En		CM	
								Sl.	Inter.	Sl.	Inter.	Sl.	Inter.
Phase 1	1/2014 – 4/2014	0.263	−0.442	0.745	2.068	0.081	9.781	0.038	0.175	−0.004	0.189	−0.021	*0.563
Phase 2	1/2015 – 1/2020	−0.030	0.675	0.484	0.419	*42.620	*19.740	0.000	*0.263	* – 0.005	*0.219	*0.005	*0.445
Phase 3	2/2020 – 3/2022	−0.121	0.060	−0.001	0.027	1.575	0.988	0.000	*0.297	−0.002	*0.112	0.002	*0.563
Overall	1/2014 – 3/2022	0.083	0.733	0.410	4.084	*94.340	*24.600	0.001	*0.266	* – 0.004	*0.233	*0.003	*0.449

(c) Regression analysis results on the number of speakers.

All the time periods are defined on a quarterly basis, i.e., 1/2014 denotes the first quarter (January to March) of 2014. In each of these studies we use the time periods as the independent variable – (a) uses CMI as the dependent variable, (b) uses the temporal proportion of Hindi, English, and Hinglish tweets as the dependent variables, and (c) uses the temporal proportion of speakers preferring Hindi, English and Hinglish as the dependent variables. Sl. denotes the slope (rate of change) and Inter. denotes the intercept (bias) from the fitted regression models. To highlight the goodness-of-fit of these models, we report the adjusted R^2 and F statistics (* denotes $p < 0.001$).

years (see Fig. 4). Notably, in Hindi, the most commonly used words are nouns, pronouns, and adverbs, while in English, the most frequently used words tend to be nouns or adjectives. Additionally, English words that occur frequently are often politically related, whereas Hindi words tend to be used in non-specialized contexts. We highlighted the most frequent words occurring in a switching context to understand the switching patterns between different linguistic groups. A *switching context* is a window where a word from a different language group occurs between words from other languages. For example, in the text "*Mujhe aaj market jaanaa hai*" (I have to go to the market today), the English word 'market' is in a Hindi context with window size 2, as all the 1-hop neighbors ('aaj', 'jaanaa') are Hindi. The word 'market' has a Hindi context even for a window size of 4. We observed the most frequent Hindi words that occur in the English context and frequent English words that occur in the Hindi context (see Fig. 5). We noticed that switched words are majorly written in Romanized scripts. This confirms that Indian users are more susceptible to code-mixing than script-switching. We further observed that Hindi verbs are more likely to be used in the English context, whereas English nouns are more likely to be used in the Hindi context. To understand how these Hindi word meanings have evolved, we calculated the *retention rate* of each Hindi word to quantify the extent of retaining meaning under English influence. We categorized these Hindi words with their parts-of-speech (PoS) and categorically analyzed the probability of retention over the years (see Fig. 6). We observed that proper nouns (PROPN) have the highest probability (0.23) of retaining meaning. On the contrary, nouns have the least probability (0.14) of retaining meaning. For all 14 PoS categories, the retention probability increases between 2017 and 2022. Among all the words, the words associated with the topic of cricket have the highest average retention rate of 0.35. On the other hand, the political and entertainment-related words have significantly lower retention rates of 0.19 and 0.20, respectively. Sports-related words such as 'cricket', 'team', and 'desh' (Hindi for the country) have the highest retention rate, indicating the linguistic preferences of this population group.

We further illustrate the evolution of Hindi words under the influence of English by highlighting two words in the context of other Hindi words – 'government', a noun, and 'khan', a proper noun, and a popular surname in India. To understand how Hinglish evolution impacted the meaning of individual words, we obtained their neighbors (words with high cosine similarity between their vector representations) and analyzed how their similarity values changed over the years (see Fig. 7A). We observed that the similarities between the word 'government' and its neighbors, such as 'Delhi' and 'central', increase with time. On the other hand, for the word 'khan', the similarity and neighborhood structure remain stable over the years (see Fig. 7B). Additionally, we analyzed the neighborhood structures of these words and their evolution over the years. Each year, we highlighted the neighbors retained in the neighborhood in the next year. Blue nodes denote the neighbors retained over the

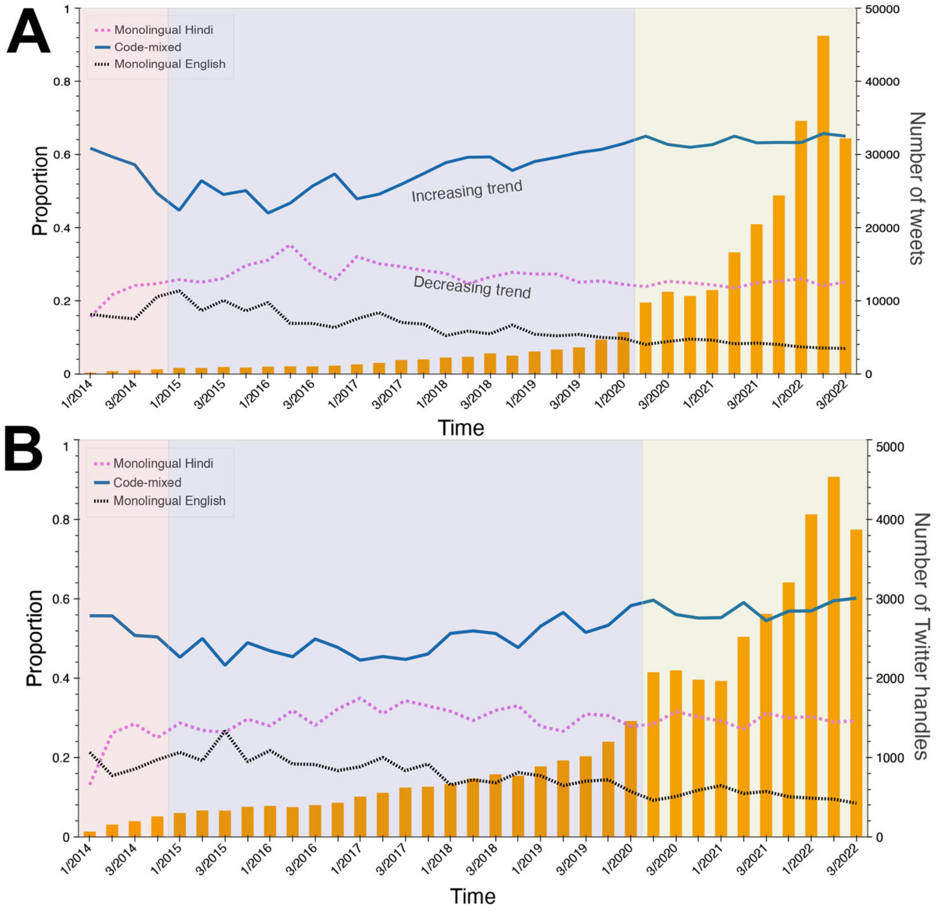

Fig. 3 The proportion of monolingual Hindi, English, and Hindi-English code-mixed usage with volume. A. Proportion of tweets made with different languages. The volume (secondary y-axis) shows the increase in the total number of tweets made by the Indian Hindi-speaking population. Both the quarterly and daily time-series data show an upward trend in code-mixed usage between 2015 and 2020 before saturating in 2021. **B**. Proportion of users from different language groups. The upward trend in both volume and code-mixing population shows the adaptiveness of the Indian Hindi-speaking population in terms of using code-mixing language on Twitter.

years, whereas green nodes denote the new neighbors. Thus, having more green nodes indicates that the word's meaning changes and therefore changes its neighbors.

On the other hand, having more blue nodes indicates a more stable neighborhood structure. For the word 'government', the blue nodes were reduced between 2018 and 2020, and more green nodes were introduced in the later years. We observe very few green nodes and more blue nodes for the word 'khan', suggesting that the word preserves its neighborhood structure and has not evolved. This analysis shows the micro-level linguistic changes due to Hinglish evolution.

Drivers behind Hinglish evolution. We highlighted the importance scores and Spearman's correlation of the exogenous features (elaborated in Section 2, *SI Appendix*) with different population trends in Fig. 8. We observed that the wholesale price index (WPI) has the highest correlation of 0.86 with the extent of code-mixing, followed by net secondary income (correlation of 0.83) and government consumption expenditure. Our proposed econometric model suggests that key economic indicators such as agriculture value-added and bilateral aid flows are the most important factors behind the growth of the Hinglish population.

We further highlighted the transition probabilities between different linguistic groups in Fig. 8C. The probability of changing a user's linguistic preference from Hindi to Hinglish is 0.43. On the other hand, a user preferring English has a higher chance of moving to Hinglish (probability of 0.78) than to Hindi (probability of 0.14). Contrarily, a user preferring Hinglish has a high probability (0.74) of remaining in the Hinglish population. If the user plans to switch to monolingualism, the chances are very high that the user will prefer Hindi (conditional probability of 0.98) to English. On average, a user has a prior probability of 0.36 (shown in Fig. 8D) being preferred to Hinglish, higher than monolingual Hindi and English.

Forecasting the growth of code-mixing on Indian Twitter. We obtained the probability of users being in different linguistic groups in the future years from the dynamic model, which we highlight in Fig. 9. Our dynamic model obtains a root mean squared error (RMSE) of 0.029. We further used an ablation of our model without considering the exogenous variables. This model resembles the ODE model proposed by Parshad et al. 2016 and achieves an RMSE of 0.045. The relative improvement of our model (55% lower RMSE) compared to its ablation version can be attributed to the influences of the exogenous variables on the dynamics of different language groups. We observe that the Hinglish population is predicted to rise with a steady annualized growth rate of 2.97%. The population's preference for monolingual Hindi will remain constant. However, the population preferring monolingual English will decrease between 2022 and

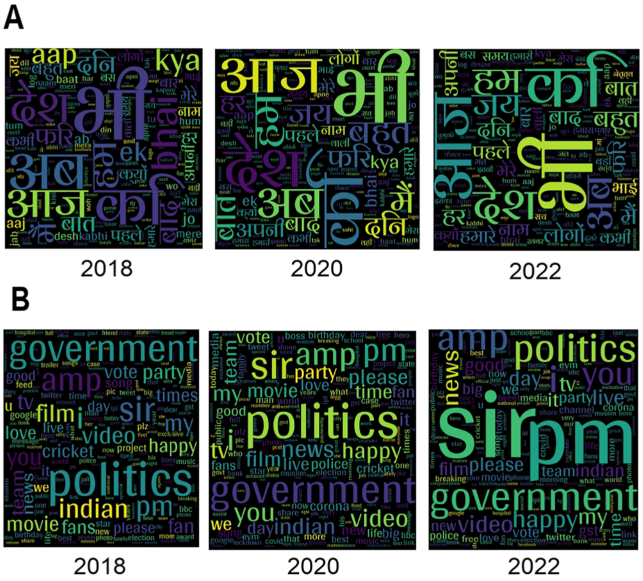

Fig. 4 Popular Hindi and English words over the years. Frequent Hindi words are more likely to be written in Devanagari than Romanized ones. The topical difference between Hindi and English usage is clearly visible. **A** The most prominent Hindi words are either pronouns or adverbs. **B** On the other hand, the frequent English words are mostly nouns and adjectives and are majorly used in political tweets.

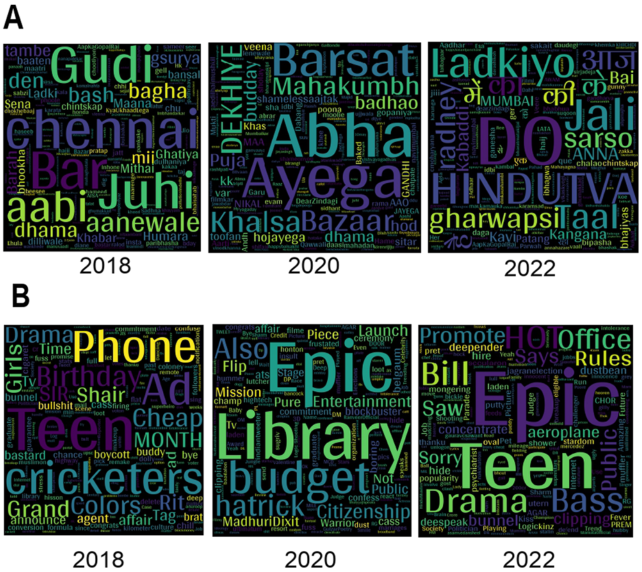

Fig. 5 Popular Hindi and English context-switching words. We observed words surrounded by words from other languages. We highlighted popular Hindi words (**A**) occurring in English contexts, i.e., Hindi words that frequently co-occur within other English words. Similarly, frequent English words occurring in Hindi contexts are highlighted in **B**.

2025 with an annualized rate of 2.98%. A sudden drop in the code-mixed population is observed for the year 2022, which can be attributed to the break in the collected data. However, it is interesting that our fitted mathematical model can resemble the historical trends observed in future years.

Discussion

Our work aims to analyze the state of the English, Hindi, and code-mixed (Hinglish) languages and populations concerning the Indian population engaging with social networks. The datasets were curated from Twitter as a proxy of the Indian population to obtain its general linguistic trends and insights. Our empirical analyses suggest that Indians tend to prefer Hinglish over monolingual Hindi or English to communicate on Twitter. Monolingual Hindi has been the second preferred mode of communication. Although the population preferring Hinglish was 26% more in 2014 than in 2015, we could not use 2014 as the base year due to the lack of tweets during the early months. Considering 2015 as the base year, we observe that the propensity for speaking in Hinglish increases steadily between 2015 and 2020, after which it stabilizes. The extent of code-mixing has also increased steadily since 2015 - 2020. Also, instead of treating the year 2022 in retrospect, we used it in a prospective analysis, as the dataset was curated until August 2022. We observe that Hinglish evolution has left a different impression on different linguistic groups. The historical average CMI on political tweets is 0.51, which is significantly higher than tweets covering Bollywood (0.47) and sports (0.43). The most frequent English words are used in political contexts, indicating that political figures tend to use Twitter as a medium to reach out to more English-proficient demographics. Twitter has become a platform for social and political activism, and Hinglish has made it easier for Indians to express their opinions and debate important issues. Bollywood, the Hindi film industry, has played a significant role in the evolution of Hinglish (Dixit 2016). Bollywood films have been a significant source of entertainment and cultural influence in India. Hinglish usage on topics related to Bollywood has grown with an annualized rate of 1.7%. On the other hand, the annualized growth rates of Hinglish on political and sports-related tweets are 1.5% and 1.4%, respectively. The regression analysis highlights how Bollywood influences the adaptability of Hinglish and helps to popularize and standardize the language. Famous Bollywood actors – Shah Rukh Khan, Aamir Khan, and Akshay Kumar are often at the center of discussions among Hindi movie-loving communities. The high coherence among the keywords 'khan', 'Shahrukh', 'Aamir', and 'Akshay' (see Fig. 7B) is the empirical evidence behind this. Further, as highlighted in the previous section, the coherence has remained stable over the years, indicating Bollywood's consistent influence on Hinglish's linguistic evolution.

To understand the key drivers behind the evolution of Hinglish, we analyzed the socio-economic and demographic indicators over the years and their impact on different linguistic groups. As described in the previous section, the economic indicators – bilateral aid flows, agriculture value-added, net secondary income, and government consumption expenditure play a positive role in the rise of Hinglish. Higher secondary income leads to higher purchasing power parity (PPP), leading to more access to the internet, which naturally gives rise to Hinglish in social media. Education and access to information are also important factors driving the evolution of Hinglish. We further report the year-wise values of these socio-economic factors in Fig. 1. Due to the economic slowdown during COVID-19, there was a dip in most of the economic indicators during 2020 - 2021 (Chaudhary et al. 2020). However, the volume of tweets was high during this period, perhaps due to the countrywide lockdown (Lancet 2020) and more time to spare. As Indians are inherently biased towards Hinglish (as observed in Fig. 8C, D), a higher number of tweets automatically leads to more code-mixing on Twitter. It is safe to assume that this will lead to Hinglish evolution in the future. Historical trends of linguistic evolution are captured in our statistical study. With the assumption that exogenous variables remain constant after the last recorded year, code-mixing follows an increasing trend after a sharp dip. This implies that under the current conditions,

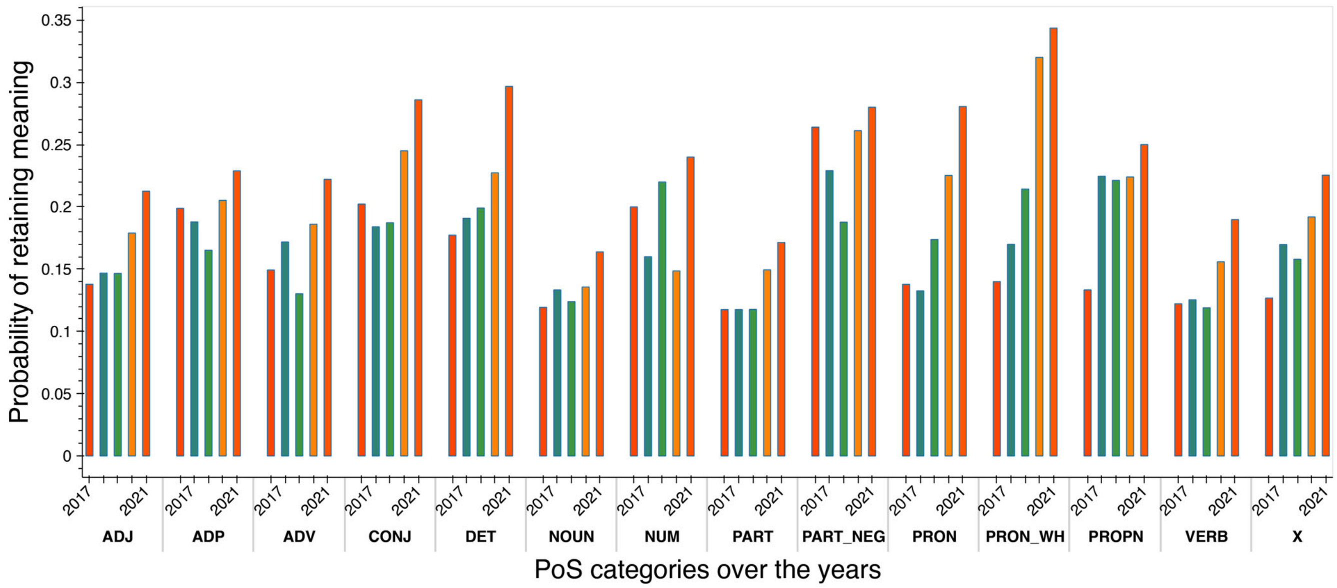

Fig. 6 Linguistic evolution of Hindi-English code-mixed language observed on Twitter. We calculated the retention rate of words between the period 2017 to 2021 and categorized them based on the parts-of-speech (PoS) tags. Proper nouns, determinants, and conjunctions exhibited high retention over the years, whereas nouns, verbs, and adjectives exhibited low retention.

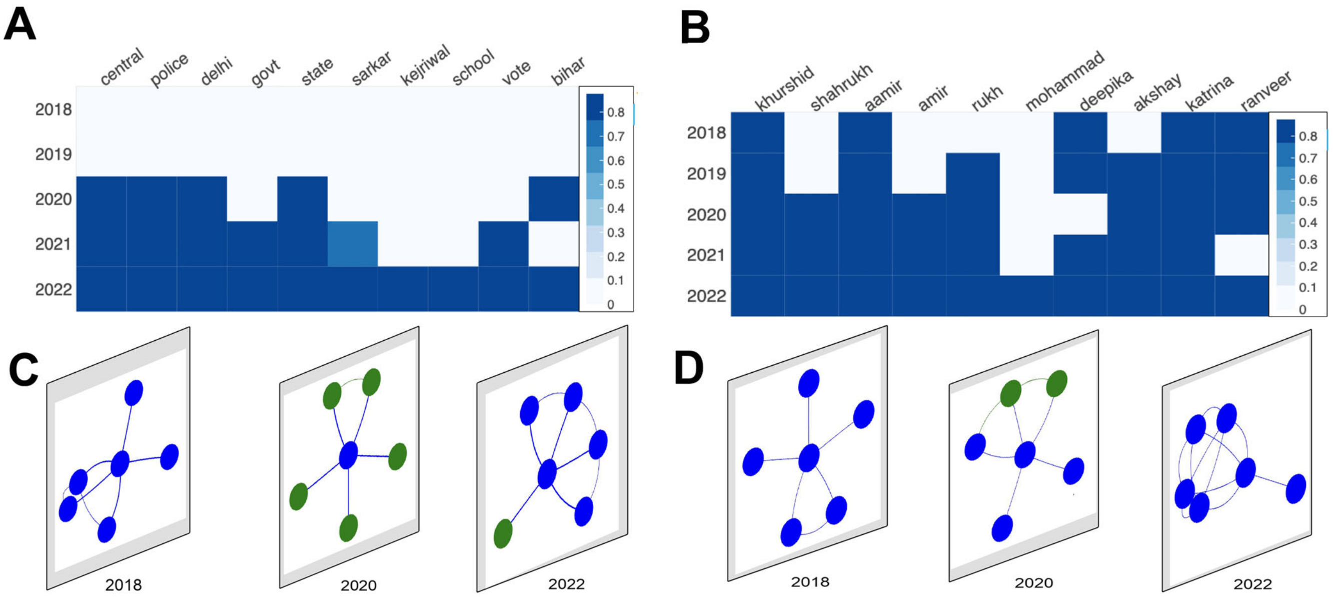

Fig. 7 Linguistic evolution of selected Hindi and English words. We visualize the linguistically similar words to `government', a noun (**A**), and `khan', a proper noun (**B**) between 2018 and 2022. We analyze the top five neighbors of the words `government' (**C**) and `khan' (**D**) to understand the linguistic evolution of these words. `Blue' nodes denote the neighbors retained over the years, `green' denotes neighbors added in a particular year, and `red' denotes neighbors from the previous year that are dropped in the next year. More red nodes for the word `government' highlight nouns' low retention rate. The neighborhood networks of `khan' are observed to be denser as compared to the networks of `government'.

the code-mixed population that the linguistic environment in India can support is lesser than the present proportion and will increase steadily to a lower asymptotic value of around 50% of the population.

Hinglish is a linguistic phenomenon derived from the cultural fusion of Hindi and English. Indians, being multilingual at large, have always been inclined towards using this hybrid language, at least on social media platforms. The code-mixed language was widely adopted in 2015 - 2020, which can be attributed to the economic growth factors. However, the switching pattern differs across all the linguistic categories but depends on the context in which the words are used. These results imply how fusion languages are derived and adopted among communities. With more engagement in social platforms comes more ethical and societal responsibilities. Fake and harmful content detection (Goel et al. 2023) from social platforms has become an urgent need for the hour. Identifying harmful materials from social media content requires natural language understanding capabilities at different linguistic hierarchies. In this work, we lay the groundwork for several fundamental properties of Hinglish

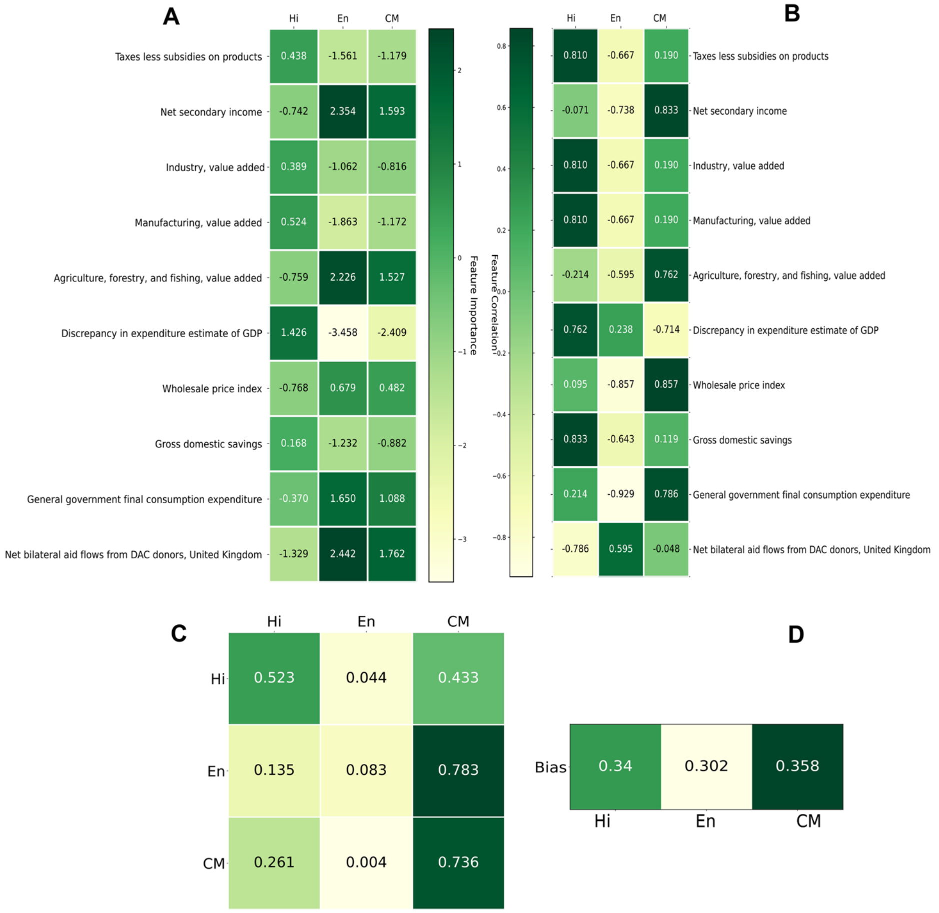

Fig. 8 Importance of different socio-economic features on linguistic preferences. We calculated feature importance (**A**) and Spearman's rank correlation (**B**) between the features and the proportion of users speaking in monolingual Hindi, English, and Hindi-English code-mixed language. We highlighted the top 10 features in terms of the highest absolute correlation with any of the two language groups. **A**. We observed the importance of each feature on different populations obtained from the dynamic model. Each importance value signifies the change in the proportion of different linguistic groups with a unit change in the feature. **C** We highlighted the transition probabilities between Monolingual Hindi (Hi), Monolingual English (En), and Code-mixed (CM) populations. Transition probabilities are not symmetric, i.e., the transition probability En → CM is not same as the transition probability CM → En. A high transition probability CM → CM indicates the propensity of CM users to stay in the CM population. **D** The bias term denotes the probability of a user being in each group without prior knowledge. Generally, users are more likely (0.36) to prefer Hinglish over monolingual Hindi or English.

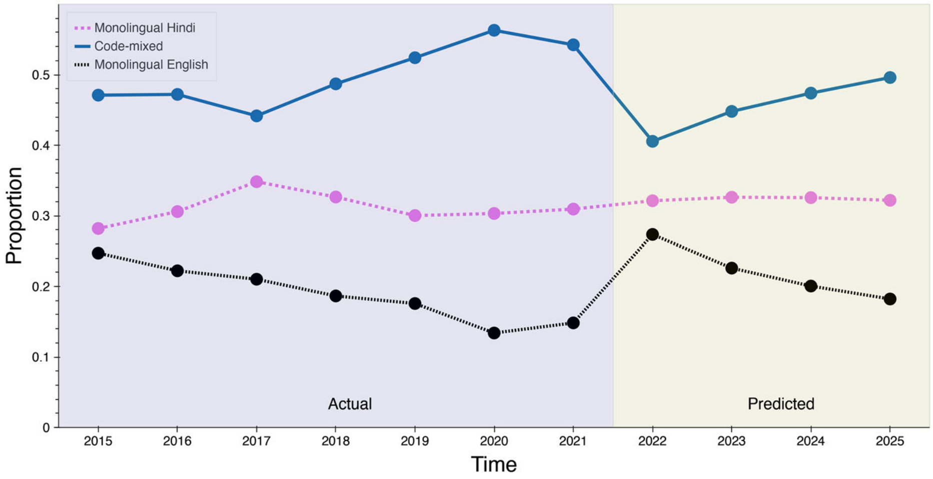

Fig. 9 Predicted proportion of Monolingual Hindi, English, and Code-mixed populations in 2022 - 2025. A steady growth in the code-mixed population can be observed after 2022, with a consistent decline in the monolingual English population. The sudden dip in the code-mixed population in 2022 is due to discontinuity in the training data.

code-mixed language and shed some light on how the language could evolve. By understanding how fake-news peddlers and hatemongers make use of code-mixing for information propagation, one can build robust computational systems to tackle these issues. Observing the results of our dynamic model and the changing semantics of words over time, we conclude that any existing datasets for code-mixed utterances will fail to capture the semantics of the ever-evolving Hinglish language. Therefore, generating new code-mixed datasets might prove helpful in keeping large language models and other NLP applications updated to consumer needs.

Data availability

All data used in this study are shared at Harvard Dataverse (can be found at https://doi.org/10.7910/DVN/BIUUW4). We collected the year-wise socio-economic features from a data bank maintained by the World Bank at https://data.worldbank.org/country/india.

Received: 18 October 2023; Accepted: 12 April 2024;
Published online: 10 May 2024

References

Abrams DM, Strogatz SH (2003) Modelling the dynamics of language death. Nature 424(6951):900–900

Aggarwal A, Wadhawan A, Chaudhary A, Maurya K (2020) “did you really mean what you said?”: Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), pages 7–15, Online. Association for Computational Linguistics

Annamalai E (2004) Nativization of English in India and its effect on multilingualism. Journal of Language and Politics 3(1):151–162

Baker S (2015). Will we all be speaking Hinglish one day? https://www.britishcouncil.org/voices-magazine/will-we-all-be-speaking-hinglish-one-day

Barath H (2019) Indian initiatives aim to break science’s language barrier. Nature 571(7764):289–289

Bhatia TK, Ritchie WC (2016) Multilingual language mixing and creativity. Languages 1(1):6

Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ (2015) Rate of language evolution is affected by population size. Proceedings of the National Academy of Sciences 112(7):2097–2102

Chakravarthi BR, Priyadharshini R, Muralidaran V, Jose N, Suryawanshi S, Sherly E (2022) Dravidiancodemix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. Language Resources and Evaluation 56(3):765–806

Chaudhary M, Sodani P, Das S (2020) Effect of covid-19 on the economy in India: Some reflections for policy and programme. Journal of Health Management 22(2):169–180

Di Sciullo A-M, Muysken P, Singh R (1986) Government and code-mixing1. Journal of linguistics 22(1):1–24

Dixit P (2016) Hinglish as a hybrid language: An analytical study. International Journal of Research and Analytical Reviews 3(1):162–167

Economist T (2014) Why Indians love cricket. https://www.economist.com/the-economist-explains/2014/02/04/why-indians-love-cricket

Gambäck B, Das A (2014) On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, pages 1–7

Gazette T M (2014) 780 languages in India. https://www.milligazette.com/news/9789-780-languages-in-india/

Goel V, Sahnan D, Dutta S, Bandhakavi A, Chakraborty T (2023) Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS Nexus*, 2(3)

International S (2022) What are the top 200 most spoken languages? https://www.ethnologue.com/insights/ethnologue200/ (2022)

Joshi A (1982) Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*

Joshi A, Prabhu A, Shrivastava M, Varma V. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491 (2016)

Kothari R, Snell R. (2011) *Chutnefying English: The Phenomenon of Hinglish*. Penguin Books India

Lancet T (2020) India under covid-19 lockdown. Lancet 395(10233):1315

Liu Q, Kusner MJ, Blunsom P (2020) ‘A Survey on Contextual Embeddings’. CoRR abs/2003.07278. https://arxiv.org/abs/2003.07278

Mabule DR (2015) What is this? Is it code switching, code mixing or language alternating? Journal of Educational and Social Research 5(1):339

Mallikarjun B (2019) Multilingualism in 21st century india. Language India 19(09):148–176

Mikolov T, Chen K, Corrado G, Dean J (2013) ‘Efficient Estimation of Word Representations in Vector Space’. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, edited by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1301.3781

Myers-Scotton C (1997) *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press

Myers-Scotton C et al. (2002) *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press on Demand

Nema N, Chawla JK (2018) The dialectics of hinglish: A perspective. Applied Linguistics Papers, (25/2) 37–51

Nie L-F, Teng Z-D, Nieto JJ, Jung IH (2013) Dynamic analysis of a two-language competitive model with control strategies. Math Probl Eng

Nilep C (2006) “code switching” in sociocultural linguistics. *Colorado research in linguistics*

Nowak MA, Krakauer DC (1999) The evolution of language. Proceedings of the National Academy of Sciences 96(14):8028–8033

Otheguy R, Stern N (2011) On so-called Spanglish. International Journal of Bilingualism 15(1):85–100

Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N (2016) What is india speaking? exploring the “hinglish” invasion. Physica A: Statistical Mechanics and its Applications 449:375–389

Patriarca M, Heinsalu E (2009) Influence of geography on language competition. Physica A: Statistical Mechanics and its Applications 388(2-3):174–186

Poplack S (1980) Sometimes I’ll start a sentence in Spanish y termino en Espanol: toward a typology of code-switching1. Walter de Gruyter, Berlin/New York Berlin, New York

Pratapa A, Bhat G, Choudhury M, Sitaram S, Dandapat S, Bali K (2018) Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553

Priyadharshini R., Chakravarthi BR., Vegupatti M. McCrae JP (2020) Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Pages 68–72. IEEE

Řehůřek R (2022) Word2vec embeddings. https://radimrehurek.com/gensim/models/word2vec.html

Roelofs M (2019)*The Features, Roles, and Functions of Dunglish in Modern Dutch Society*. B.S. Thesis

sagorsarker codeswitch-hineng-lid-lince (2020) https://huggingface.co/sagorsarker/codeswitch-hineng-lid-lince

Sankoff D, Poplack S (1988) Code switching. *InH. von Ulrick Ammon, Norbert Dittmar, and Klaus J. Mattheier (eds) Sioclolinguistics–Soziolinguistik*, 2: 1174–1180

scikit-learn developers sklearn.preprocessing.standardscaler (2023) https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sezerer E, Tekir S (2021) ‘A Survey On Neural Word Embeddings’. CoRR abs/2110.01804. https://arxiv.org/abs/2110.01804

Si A (2011) A diachronic investigation of Hindi–English code-switching, using Bollywood film scripts. International Journal of Bilingualism 15(4):388–407

Singh K, Sen I, Kumaraguru P (2018) A Twitter corpus for Hindi-English code mixed pos tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 12–17

Sreelakshmi K, Premjith B, Soman K (2020) Detection of hate speech text in Hindi-English code-mixed data. Procedia Computer Science 171:737–744

Srivastava A., Bali K. Choudhury M (2020) Understanding script-mixing: A case study of Hindi-English bilingual Twitter users. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 36–44 (2020)

Thara S Poornachandran P (2018) Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388. IEEE (2018)

TOI (2010) Indiaspeak: English is our 2nd language. https://timesofindia.indiatimes.com/india/Indiaspeak-English-is-our-2nd-language/articleshow/5680962.cms

Twitter (2023) Twitter API academic research access. https://developer.twitter.com/en/products/twitter-api/academic-research

Walters CE (2014) A reaction-diffusion model for competing languages. Meccanica 49(9):2189–2206

Wang C, Cho K Kiela D (2018) Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158 (2018)

Yusnida D, Muliawati I, Rezeki HS (2022) A descriptive analysis of code-mixing types used by youngsters in online communication. Journal of English Teaching and Linguistics 3(2):98–108

Zhang W (2012) Chinese-english code-mixing among China's netizens: Chinese-english mixed-code communication is gaining popularity on the Internet. English Today 28(3):40–52

Author contributions

AS, MSD and TC conceived the experiments, AS and SD conducted the experiments, and all the authors analyzed the results, and wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03058-6.

Correspondence and requests for materials should be addressed to Tanmoy Chakraborty.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.