TWITTER SENTIMENT ANALYSIS ON HINDI IMPOSITION ISSUE

Ojas Arora, Pranav Iyengar and Rochan Mohapatra

CS-2378/POL-2070-1

May 10, 2022

Introduction

Following Amit Shah's comments that Hindi should be made the official language of India and that more people from other states should speak in Hindi, Twitter saw a frenzy of tweets.[1] It saw mixed reactions from both Hindi and non-Hindi speakers.[2] It also attracted much disparagement from the leaders of non-Hindi speaking states. Language remains an essential socio-cultural indicator in Indian politics, and given the history of language conflicts in India, such a reaction was anticipated. However**,** with Twitter emerging as an essential tool for running propaganda campaigns and aiding narrative formation, it plays an essential role in understanding political receptivity in online spaces.[3] Subsequently, it has also become a potent platform to influence individual opinions' (for the better or worse). Given the presence of targeted political bot accounts that only function to add to the ongoing political trends on Twitter, understanding the use of Twitter as a platform for opinion-making and shaping vis-à-vis political discourse is essential.

To understand how Twitter users react to political changes, this paper, with the help of tweet mining analyses the tweets with relevant hashtags between 28th April and 7th May. These hashtags are: #StopHindiImposition, #OneNationOneLanguage, #HindiLanguage and #HindiIsNotNationalLanguage. The tweets are only mined between the said dates because of the limited access on the Twitter developer's account. For this analysis, the sample size of tweets is >100,000. The analysis helps us gauge Twitter users' sentiment trajectory pertaining to the language issue in question. Correspondingly, we operationalised the analysis by reading the mined tweets, calculating their bot score,[4] and tracing their geographic locations. Following this, we concluded that Twitter users from Southern India were more opposed to Hindi being declared as the national language of India. On the other hand, people from Northern India were either indifferent or supportive of the idea. Furthermore, contrary to the assumption, Twitter bot accounts did not have a polarising effect on other Twitter users in this case. Besides that, the paper will give political background to sufficiently understand and conceptualise the political relevance of language politics in India.

---

[1] Caldarelli, Guido, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. "The Role of Bot Squads in the Political Propaganda on Twitter." *Communications Physics* 3, no. 1 (May 11, 2020). https://doi.org/10.1038/s42005-020-0340-4.

[2] We choose these as two categories because it makes the conceptualisation of the data accessible. It also helps us in analysing the data better.

[3] Gaisbauer, Felix, Armin Pournaki, Sven Banisch, and Eckehard Olbrich. "How Twitter Affects the Perception of Public Opinion: Two Case Studies." Research Gate. https://www.researchgate.net/publication/344072776_How_Twitter_affects_the_perception_of_public_opinion_Two_case_studies.

[4] Bot score indicates the likelihood of the user account being a bot. https://twitter.com/botometer?lang=en

Methodology

We pulled as many tweets as possible from the past week on different hashtags, such as those mentioned above (see *Introduction*), using Tweepy.[5] We created multiple .csv files, keeping only tweets with geolocation on, limiting our dataset from over 100,000 tweets to below 500.

After pulling the data, we cleaned it by removing rows with missing values, and duplicates. After doing this, we then tried to geolocate the data using Mapbox,[6] using the coordinates provided by Twitter. However, we then ran into the issue that the coordinates given by Twitter were those of a bounding box and not a precise location. Hence, we then moved on to hand-label each tweet - whether from the South/East or North/West of India. This segregation was done as people in the North/West tend to speak Hindi whilst those in the South/East speak regional languages[7].

After hand-labelling each tweet by location, we then analysed the bot score of each user. To do this, we made use of the Botometer Pro API.[8] We passed each username of those who had tweeted to the API, and the API returned multiple values, such as CAP, fake follower percentage etc. We only kept those that were needed - universal and English bot scores. These are scores rated by the API based on the language used by each user. Hence, if the user's primary language of tweeting was English, we assigned the English bot score to the user; otherwise, the universal bot score.

Finally, we moved on to the sentiment analysis of each tweet. We first tested two different packages with different approaches – FlairNLP [9] and NLTK Vader. [10] We ran each library on the first 50 tweets of the dataset and then analysed the results manually. We concluded that Flair gave us a more accurate sentiment analysis, and hence, we ran sentiment analysis using Flair on the complete dataset.

After gathering and calculating all the data, we ran two regressions. First, we ran a regression on bot score and sentiment. Next, we ran a regression on sentiment and geolocation. The analysis of these regressions will be written in *Data Analysis.*

Assumptions and Limitations in Methodology

An assumption we had is that people who are active on the issues are more likely those who have a negative sentiment.

The first limitation we found was tweet scraping. As we did not have access to an Academic Researcher Twitter API, we could not my tweets from the historical database, only the past week (a past week from mining time). Furthermore, no matter how many tweets we pulled, there was a scarce number of users with their geolocation on. Hence, our dataset was whittled down.

---

[5] Tweepy is a Python package that allows you to use the Twitter API. https://docs.tweepy.org/en/stable/api.html
[6] Mapbox is a developer-friendly mapping and location cloud platform. https://docs.mapbox.com/api/overview/
[7] Misra, Satish. "Battle for the Hindi heartland: Will it favour the BJP again?" India Matters https://www.orfonline.org/expert-speak/battle-for-the-hindi-heartland-will-it-favour-the-bjp-again-50031/
[8] Botometer monitors Twitter account activity and assigns a score depending on how likely they are to be bots.https://botometer.osome.iu.edu/api
[9] https://documenter.getpostman.com/view/5353571/TzsbKTAG
[10] NLTK gives a comprehensive introduction to language processing programming. https://www.nltk.org/

The second limitation we came across was that of sentiment analysis. Sentiment analysis failed when there were non-English characters in the tweet and also when there were non-English words typed in English. Furthermore, sentiment analysis will only tell us the tone of the tweet. The negative or positive sentiment won't tell us whether the tweet is towards or against the imposition of Hindi.

For example, one tweet said, *'Those who are opposed imposition of Hindi language. Let's start communicating with each other in English. We will educate English to our brothers and sisters from Hindi speaking people so that they too can become comfortable in English. Jai Bhim! #HindiIsNotNationalLanguage.*' Sentiment analysis deemed this tweet as positive, however, this tweet is going against the imposition of Hindi, thereby making the sentiment misleading. Along those same lines, another user tweeted: '*'if you don't know hindi language , so from today don't dubbed your movies in hindi, if you have any standard 👎 @KicchaSudeep #HindiLanguage #Hindi #Hindustani https://t.co/0bgSrrRhcJ*' The user is trying to say something positive towards Hindi in a negative context, and again, the sentiment analysis becomes misleading.

The final limitation we found was that of bot scores. We tried to approach the bot scores as a binary idea - those above a bot score threshold would be deemed a bot. However, this would give us many false positives and a few false negatives, as these scores are indicative of how likely one might be a bot, not a definitive amount. Essentially, we are getting a likelihood (probability) of a user being a bot. Hence, we might wrongly identify a user as a bot or not identify a bot as one.

We notice a few limitations when extending the research to other scenarios. India, unlike many countries, has a vibrant and dense culture and diversity, which may not be the case in other nations for similar comparisons. Additionally, the number of people on Twitter is a set of people who are primarily above the poverty line, limiting the extent of the analysis to only those.

Political Context

Many believe that having a common language is a must for national integrity. However, India has repeatedly proved that wrong. The Indian constitution officially recognises twenty-two languages, and there are more than a hundred regional languages spoken in India. Language in India has two connotations: cultural and literal. It is used as a medium to communicate and express effectively, and at the same time, it entails a cultural meaning because of shared history. As Paul Friedrich puts it, language becomes increasingly politicised by linguistic as well as political groups to overshadow other problems in the society, i.e., unemployment, increased crime, poor health indicators, et cetera.[11] In simpler terms, it has become a crutch for political leaders to hide their policy failures — and it has worked brilliantly for them.

Given that language also plays a vital role in disseminating policy and welfare measures, it is more likely that language spoken and understood by the majority of the population gets an edge. Thereby, it gives the champions and professors of the Hindi language an edge. Notwithstanding, having only one national language can increase political tensions and create

[11] Friedrich, Paul. "Language and Politics in India." *Daedalus*, Current Work and Controversies—2 (Summer, 1962), 91, no. 3 (1962): 543–59. https://doi.org/https://www.jstor.org/stable/20026727, see p. 546.

unnecessary hurdles for the already burdened bureaucracy. The gruesome clashes in Tamil Nadu and other southern states over Hindi imposition in 1965, agitation in parts of Assam, West Bengal, Manipur, et cetera should serve as a reminder for the political pundits to not try to clamp down on multilingualism.[12] Furthermore, it also refreshes demographic anxieties amongst non-Hindi speakers. One cannot ignore the weightage that speech entails in identity politics in India. The separation of states based upon linguistic boundaries in 1956 following the recommendation by the 'State Re-organisation Committee' of 1953[13] reflects that it has always been prudent to facilitate linguistic differences.[14] For it helps maintain the social fabric of the society and mitigates any risk of social division in the society.

The argument that Hindi is the most spoken language in India is distorted, as it does not account for the various dialects and Khari *bolis* in the region. Moreover, as a category, Hindi has been often used blanketly to categorise people who speak Hindustani, Urdu, Haryanvi, Marwadi, et cetera.[15] It provides a false representation of the Hindi-speaking population, and it also conflates different dialects and regional languages. However, with less awareness about these intricacies, the same has been normalised.

Given Bhartiya Janta Party's apparent obsession with the Hindi language, annotating it as the national language will have varied political implications for the party. BJP has used the Hindi language as a cultural tool to score political mobilisation and pander to the Hindu population.[16] [17] Furthermore, it has helped them create a cultural doctrine that can be deployed time and again to influence people and engage in soft politics. Touting Hindi as the national language sits perfectly well with their Hindu nationalist brand of politics. Undoubtedly, the BJP government has been able to use this as a ploy to deviate attention from other issues of importance. The use of the Hindi language as a soft-politics tool, among other things, contributes significantly to the electoral ambitions of the BJP.[18] However, in so doing, BJP is playing with fire and neglecting the larger picture.

Indeed, BJP has significantly penetrated the Indian political system and steadily rose the political ladder. Nonetheless, there is still a considerable chunk of voters who do not fully subscribe to their brand of politics and do not support the idea of Hindi as the national language, either. Making Hindi the national language would mean risk alienating its voter base in South

---

[12] Darbhamulla, Sruthi. "Explained: The Anti-Hindi Imposition Movements in India.". The Hindu, April 23, 2022. https://www.thehindu.com/news/national/explained-the-anti-hindi-imposition-movements-in-india/article65346376.ece.

[13] "State Reorganisation Commission Report of 1955_270614: MHA." Ministry of Home Affairs. https://www.mha.gov.in/sites/default/files/State%20Reorganisation%20Commisison%20Report%20of%201955_270614.pdf.

[14] Hegde, Pandurang. "India's Multilingualism Faces Threats." Deccan Herald. DH News Service, March 3, 2021. https://www.deccanherald.com/opinion/in-perspective/indias-multilingualism-faces-threats-957845.html.

[15] Friedrich, Paul. "Language and Politics in India."., see p. 550.

[16] Bhatia, Sidharth. "Hindi, the New Hindutva Weapon of Polarisation." The Wire. The Wire. https://thewire.in/politics/amit-shah-hindi-polarisation.

[17] Handa, Ekta. "Why the BJP Should Get over Its Hindi Obsession." India Today, September 14, 2017. https://www.indiatoday.in/fyi/story/india-hindi-national-language-official-constitution-1022720-2017-07-06.

[18] Bhatia, Sidharth. "Hindi, the New Hindutva Weapon of Polarisation." The Wire. The Wire.

India, Northeast India, parts of Eastern India, and people who take pride in their language. While language is not the most deterministic factor in accounting for political clout and success, it certainly shapes voters' decisions. We have seen it in Tamil Nadu in the past[19] [20] Therefore, any effective decision to diminish the significance of other regional languages is tantamount to an attack on a person's personality, as language constitutes an essential part of a person's shared and individual identity.[21] It is no lie that Hindi and power politics are closely intertwined and will remain for times to come.[22] The same would garner more political traction with easy availability and accessibility to social media, print media, and electronic media — all of which contribute significantly to opinion formation and discourse.

Amit Shah's comments have already attracted massive backlash from Asom Sahitya Sabha and Manipur's Meitei Erol Eyek Loinashillon Apunba Lup (MEELAL), civil societies in the North-eastern region.[23] Many celebrities like Kicha Sudeep, A.R. Rahman, Divya Spandana, Chiranjeevi, et cetera took to Twitter and various other social media platforms to express contention against the Shah's comments.[24] Contrastingly, many celebrities also approved and extended support to Shah's suggestions. Even though Twitter is not representative of what Indians feel about a political issue — in this case, the Hindi imposition issue — it helps us gauge how social media giants like Twitter contribute to opinion formation and enabling discourse. Additionally, it should also not be assumed that the opinion of celebrities or people with a massive following on Twitter always has a defining influence on opinion formation.

Data Analysis

---

[19] Sivapriyan, ETB. "Language and Identity Politics Making a Comeback in Tamil Nadu." Deccan Herald. DH News Service, September 11, 2020. https://www.deccanherald.com/national/south/language-and-identity-politics-making-a-comeback-in-tamil-nadu-885690.html.

[20] Rao, V. Venkata. "LANGUAGE POLITICS IN INDIA." *The Indian Journal of Political Science* , July-September , 31, no. 3 (July 1970): 203–21. https://doi.org/https://www.jstor.org/stable/41854382., see p.211.

[21] Rao, V. Venkata. "LANGUAGE POLITICS IN INDIA."., see p. 208.

[22] Friedrich, Paul. "Language and Politics in India."., see p. 552.

[23] Agarwala, Tora. "Northeast Groups Oppose Centre's Hindi Move, Call It an 'Imposition'." The Indian Express, April 11, 2022. https://indianexpress.com/article/north-east-india/assam/northeast-groups-oppose-centres-hindi-move-call-it-an-imposition-7862587/.

[24] Jain, Jhalak. "From Ajay Devgn to Kiccha Sudeep, Celebs Divided over 'National Language'." TheQuint, May 6, 2022. https://www.thequint.com/neon/now-rolling/ajay-devgn-kiccha-sudeep-celebs-divided-over-hindi-as-national-language

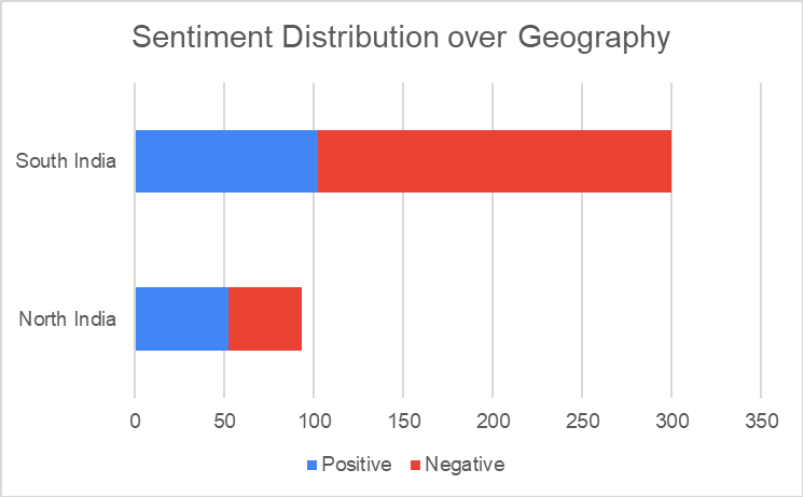

Figure 1

In *Figure 1*, we conducted an analysis based on tweets' geographical locations. It should be noted that tracing the geographical locations of Twitter users is a difficult task, and as a result, the number of tweets analysed for this experiment is less. That said, out of the total tweets examined, over 74.4% of them are from South India (Tamil Nadu, Karnataka, Kerala, Andhra Pradesh and Telangana), and 22.8% are from other states in India (Punjab, Haryana, Bihar, Uttar Pradesh, Madhya Pradesh, Jharkhand, Chhattisgarh et cetera), and 2.8% are from outside India. It reflects that people from Southern India not only tweeted more actively and passionately, but they also vehemently opposed the idea of Hindi as the national language. On the flip side, people from Northern India were more likely indifferent or supportive of the picture.

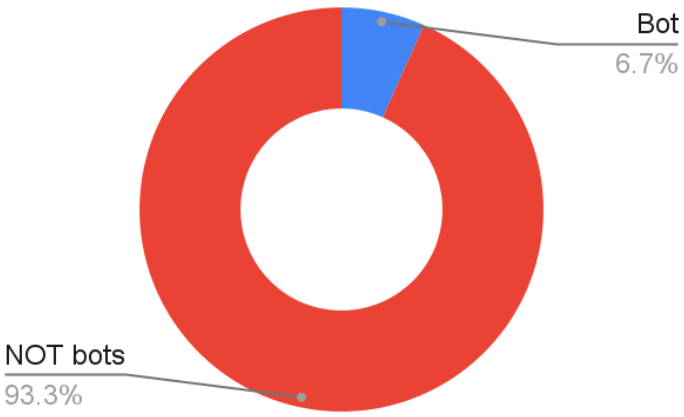

Figure 2

In *Figure 2*, we analysed the users being bots or humans (done using BotoMeter). We used a binary classification method, classifying any user above a bot score of 3.6 as a bot. We assume that anyone over this threshold is almost definitely a bot and continue under this assumption.

By this method, contrary to our expectation, a staggering 93.3% of the tweets were made by humans rather than bots. Contrary to initial the assumption, Twitter bot accounts did not have a polarising effect on other Twitter users in this case.

Regression

```
. regress finalsentiment botscore

      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(1, 473)       =      0.20
       Model |  .138038784         1  .138038784   Prob > F        =    0.6519
    Residual |  320.412133       473  .677404086   R-squared       =    0.0004
-------------+----------------------------------   Adj R-squared   =   -0.0017
       Total |  320.550171       474  .676266184   Root MSE        =    .82305

------------------------------------------------------------------------------
finalsenti~t | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    botscore |   .0166542   .0368934     0.45   0.652    -.0558409    .0891494
       _cons |  -.2753958   .0617363    -4.46   0.000     -.396707   -.1540845
------------------------------------------------------------------------------
```

We hypothesised that bots would be more likely to have radical tweets - tweets with more negative or positive sentiment. Hence, those who follow bots would be likely to be polarised. Therefore, we regressed the sentiment and the bot score to test the hypothesis. Our findings suggest that if the bot score goes up, the sentiment also goes up – a positive correlation. However, looking at the T values, we can conclude that the results are statistically insignificant, and there is much noise. Hence, we can assume that the bot score does not significantly correlate with the sentiment, and therefore, our hypothesis is false.

```
. binnorrth |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        359       78.04       78.04
          1 |        101       21.96      100.00
------------+-----------------------------------
      Total |        460      100.00

. regress finalsentiment binnorrth

      Source |       SS           df       MS      Number of obs   =       460
-------------+----------------------------------   F(1, 458)       =     21.79
       Model |  13.9859811         1  13.9859811   Prob > F        =    0.0000
    Residual |  293.928656       458  .641765625   R-squared       =    0.0454
-------------+----------------------------------   Adj R-squared   =    0.0433
       Total |  307.914637       459   .67083799   Root MSE        =    .8011

------------------------------------------------------------------------------
finalsenti~t | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   binnorrth |   .4212282   .0902317     4.67   0.000     .2439087    .5985477
       _cons |  -.3547579   .0422806    -8.39   0.000    -.4378459   -.2716699
------------------------------------------------------------------------------
```

We also hypothesised that tweets from the North/West would be more positive, as they would support the imposition of Hindi, while tweets from the South/East would be more negative. To prove this hypothesis, we ran a regression between sentiment and geolocation. We made geolocation a binary variable - 1 for NW and 0 for SE. We then ran the regression, and our findings show a correlation between the two variables. People from the NW are more likely to tweet positively about this than those from the SE, proving our hypothesis.

Conclusion

It is apt to conclude that Twitter users in Southern India were more opposed to Hindi being India's official language. North Indians, on the other hand, were either neutral or enthusiastic about the proposal.

Figure 3

The tweets' analysis and geographical division between North and South India point towards the same. As seen in *Figure 3*, the number of tweets from the Southern states is much higher than that of the Northern states (over 36% of the tweets are from Karnataka alone).

Our initial hypothesis was that the presence of bots would be an essential factor in the polarisation of tweets. However, after running the regression, it was found that bots do not have a notable correlation with the sentiment. In addition, contrary to popular belief, Twitter bot accounts did not polarise other Twitter users in this scenario.

The next premise was that the tweets from the Northern states would be more likely to have a positive sentiment while the tweets from the Southern states would have a negative

sentiment. On running the regression, the premise was validated, and it was further proved that people from the South are more likely to tweet about the said issue.

The analysis of the tweets mined majorly yields the result that people who are troubled by a policy are the ones who voice their opinion actively. Additionally, it is not only the number of tweets that indicated a negative sentiment but also the sheer disparity between the sentiments and the absence or scarce presence of external players (like bots) that further strengthens the point that bots' are limited.

Work Distribution

Section	**Worked by**
Tweet Mining, Bot Score Calculation, Sentiment Analysis	Pranav
Introduction & Conclusion	Rochan
Political Context, Background and Citations	Ojas
Analysis & Implications	Pranav, Rochan, Ojas
Regression	Pranav

Bibliography

Agarwala, Tora. “Northeast Groups Oppose Centre's Hindi Move, Call It an 'Imposition'.” The Indian Express, April 11, 2022. https://indianexpress.com/article/north-east-india/assam/northeast-groups-oppose-centres-hindi-move-call-it-an-imposition-7862587/.

Bhatia, Sidharth. “Hindi, the New Hindutva Weapon of Polarisation.” The Wire. The Wire. Accessed May 9, 2022. https://thewire.in/politics/amit-shah-hindi-polarisation.

Caldarelli, Guido, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. “The Role of Bot Squads in the Political Propaganda on Twitter.” *Communications Physics* 3, no. 1 (May 11, 2020). https://doi.org/10.1038/s42005-020-0340-4.

Darbhamulla, Sruthi. “Explained: The Anti-Hindi Imposition Movements in India.” Return to frontpage. The Hindu, April 23, 2022. https://www.thehindu.com/news/national/explained-the-anti-hindi-imposition-movements-in-india/article65346376.ece.

Friedrich, Paul. “Language and Politics in India.” *Daedalus*, Current Work and Controversies—2 (Summer, 1962), 91, no. 3 (1962): 543–59. https://doi.org/https://www.jstor.org/stable/20026727.

Gaisbauer, Felix, Armin Pournaki, Sven Banisch, and Eckehard Olbrich. “How Twitter Affects the Perception of Public Opinion: Two Case Studies.” Research Gate . Accessed May 8, 2022. https://www.researchgate.net/publication/344072776_How_Twitter_affects_the_perception_of_public_opinion_Two_case_studies.

Handa, Ekta. “Why the BJP Should Get over Its Hindi Obsession.” India Today, September 14, 2017. https://www.indiatoday.in/fyi/story/india-hindi-national-language-official-constitution-1022720-2017-07-06.

Hegde, Pandurang. “India's Multilingualism Faces Threats.” Deccan Herald. DH News Service, March 3, 2021. https://www.deccanherald.com/opinion/in-perspective/indias-multilingualism-faces-threats-957845.html.

Jain, Jhalak. “From Ajay Devgn to Kiccha Sudeep, Celebs Divided over 'National Language'.” TheQuint, May 6, 2022. https://www.thequint.com/neon/now-rolling/ajay-devgn-kiccha-sudeep-celebs-divided-over-hindi-as-national-language.

Misra, Satish. “Battle for the Hindi heartland: Will it favour the BJP again?”. India Matters, May 9, 2022. https://www.orfonline.org/expert-speak/battle-for-the-hindi-heartland-will-it-favour-the-bjp-again-50031/

Ojas Arora, Pranav Iyengar, Rochan Mohapatra. “Consolidated Database.” (May 5, 2022). https://docs.google.com/spreadsheets/d/104IeO2GQN3ABL0At_nSpTTGmFzNMIFCIPYOAHDXIyhY/edit?usp=sharing.

Rao, V. Venkata. “LANGUAGE POLITICS IN INDIA.” *The Indian Journal of Political Science* , July-September , 31, no. 3 (July 1970): 203–21. https://doi.org/https://www.jstor.org/stable/41854382.

Sivapriyan, ETB. “Language and Identity Politics Making a Comeback in Tamil Nadu.” Deccan Herald. DH News Service, September 11, 2020. https://www.deccanherald.com/national/south/language-and-identity-politics-making-a-comeback-in-tamil-nadu-885690.html.

“State Reorganisation Commisison Report of 1955_270614: MHA.” Ministry of Home Affairs . Accessed May 8, 2022. https://www.mha.gov.in/sites/default/files/State%20Reorganisation%20Commisison%20Report%20of%201955_270614.pdf.