Mapping Semantic Domains Across India's Social Media: Networks, Geography, and Social Factors

Gunjan Anand
University of Illinois
Urbana-Champaign,
Urbana, IL 61801 USA
gunjana2@illinois.edu

Jonathan Dunn
University of Illinois
Urbana-Champaign,
Urbana, IL 61801 USA
jedunn@illinois.edu

Abstract

This study examines socially-conditioned variation within semantic domains like kinship and weather using thirteen Indian cities as a case-study. Using bilingual social media data, we infer six semantic domains from corpora representing individual cities with a lexicon including terms from English, Hindi and Transliterated Hindi. The process of inferring semantic domains uses character-based embeddings to retrieve nearest neighbors and Jaccard similarity to operationalize the edge weights between lexical items within each domain. These representations reveal distinct regional variation across all six domains. We then examine the relationship between variation in semantic domains and external social factors such as literacy rates and local demographics. The results show that semantic domains exhibit systematic influences from sociolinguistic factors, a finding that has significant implications for the idea that semantic domains can be studied as abstractions distinct from specific speech communities.

1 Introduction

India is a country with many diverse cultures and languages. This creates interactions between languages, particularly Hindi and English in the Northern regions of India and between Hindi and other languages elsewhere. This paper asks whether such longstanding linguistic and cultural contact changes the character of semantic domains present within thirteen Indian cities. Much previous work views semantic domains as language-specific, so that a language like Hindi has a single semantic map for a domain like kinship. The contribution of this paper is to show that social factors like language contact have a systematic influence on the structure of semantic domains. India provides an ideal case-study because of these longstanding contact situations.

We focus on six semantic domains: weather, kinship, emotion, animals, professions and temporal units. These domains were chosen because of their known variation in lexical granularity between English and Hindi. For example, Hindi distinguishes between paternal and maternal grandfathers lexically, whereas English uses the same term for both relationships. Additional modifiers (*paternal, maternal*) are used in English when necessary. In contrast, Hindi uses the same term for *yesterday* and *tomorrow*, disambiguating based on verb tense and context; English uses distinct words for these two concepts. These examples show how languages can encode conceptual distinctions with differing levels of granularity. A speaker's lexical choices are shaped by grammatical and cultural systems enforced in the lexicon. Our question here is the degree to which linguistic and cultural contact create variation within semantic domains within the same languages.

To investigate this question, we analyze data from Indian social media. This kind of spontaneous, everyday language use provides insight into how different populations lexicalize these six semantic domains. Social media offers a large-scale, naturalistic corpus to capture regional variation. In particular, it allows us to ask whether the same semantic concepts are realized with consistent lexical patterns across cities or whether these patterns diverge due to differences in language contact and social environment. We develop a corpus of over 50 million samples containing a mix of English and Hindi usage across thirteen Indian cities as a means of observing semantic domains across regional populations.

Given population-specific corpora, we need to infer a representation of these semantic domains in order to compare them across populations. We take a data-driven approach based on non-contextual character embeddings from fastText, learning a separate model from each city-specific corpus. These embeddings can be seen as approximations of con-

Proceedings of the 16th International Conference on Computational Semantics, pages 307–320

ceptual structure in which lexical items from the same domain form a neighborhood within the embedding space. This approach to operationalizing a semantic domain as similarities within an embedding space aligns with an opposition theory approach to signs (de Saussure, [1916] 1983). This approach posits that the value of a concept is determined by its contrasting relations within the system of language, particularly how it contrasts with other similar terms. Therefore, embeddings offer a way to operationalize the structure of these semantic domains which can then be used to measure the degree to which these domains vary across speech communities.

Importantly, many concepts in these six domains exhibit co-lexification: there is not a one-to-one mapping between form and meaning. For example, the cases of *paternal/maternal grandfather* (in English) and of *yesterday/tomorrow* (in Hindi) are instances in which one language co-lexifies what the other splits into two separate items. In our bilingual corpus data, however, a speaker is not limited to the co-lexification patterns of either language. We hypothesize that this provides additional flexibility to the mapping between form and meaning within lexical items, allowing them to vary systematically across populations due to sociolinguistic conditions. If this is the case, we would expect that the operationalizations of these semantic domains, created using an embedding space, will also differ across regions in predictable ways.

This paper makes three main contributions: First, we show that these semantic domains, as inferred from corpora, vary significantly across Indian cities in way that corresponds with different levels of overall language contact. Second, we show that these variations are relatively stable across all six domains and are not artifacts within only a single domain. And, third, we show that these variations are not simply random but are significantly related to social and demographic factors. Taken together, these findings suggest that semantic domains are not a single entity shared by all speakers of a language but rather systems which are influenced by social factors like differing degrees of language contact.

After reviewing related work in semantic domains and social factors in Section 2, we present a dataset derived from Twitter/X posts from various Indian cities in Section 3. This dataset contains samples in English, Hindi and Transliterated Hindi. Our method for operationalizing semantic domains using an embedding space is detailed in Section 4, along with the social factors used for later analysis. The analysis of variation in semantic domains across cities is presented in Section 5, with a special focus on the relationship between these variations and external social factors like language contact. We end, in Section 6, by discussing the larger implications of this work on the interface between sociolinguistics and computational semantics. While previous computational work has abstracted away from sociolinguistic factors in the representation of semantic domains, the findings in this paper show that such idealized representations will not capture variations within the speech community.

2 Previous Work

2.1 Computational Approaches to Semantic Analysis

Word embeddings have become a widely used computational method for analyzing contextual relationships between words used in corpus data. Mikolov et al. (2013) suggests that embeddings allow semantic similarity to be mapped and quantified through vector proximity in embedding high dimensional spaces. This is further demonstrated studies such as in Jatnika et al. (2019) and Jin and Schuler (2015) which confirm that words which share similar contexts tend to cluster together in embedding spaces.

As explained by Lai et al. (2015), these models generate word vectors based on surrounding context, allowing semantic relatedness to be inferred by vector proximity. This aligns with opposition theory (de Saussure, [1916] 1983) as if a vector gets its value by the opposition vectors, semantic relatedness can be seen by how close the vectors are. So one would wonder how semantic domains can be seen in embedding spaces and how variant this would be within the domains?

Recent work has focused on applying this framework to study semantic domains. Grand et al. (2018) used embedding spaces to project out semantic domains (e.g. animals, weather, professions), showing how humans mentally organize semantic fields through patterns of usage. However, Antoniak and Mimno (2018) cautions the usage of such frameworks as these results may be sensitive to corpus size and sampling variability which raises concerns about how reproducible and conclusive the results can be. They suggest bootstrapping over multiple samples which is used to check stability of the model in this paper. Similarly, Burdick et al. (2021)

report variation in embedding stability across languages, particularly in morphologically rich contexts - an insight which is important to keep in mind while looking at India's multilingual landscape.

fastText, developed by Bojanowski et al. (2017) represents a significant advancement in creating embedding spaces for words in a corpus. The model has the ability to capture sub-word information which would help in analyzing morphologically rich languages like Hindi. Studies such as Rana et al. (2024) and Thavareesan and Mahesan (2020) have used fastText embeddings to analyze semantic similarity, confirming the model's strength in multilingual environments.

Building on these methods, this study extends prior work by analyzing semantic domain variation across cities in India. It uses fastText embeddings to map word usage onto high dimensional embedding spaces where each lexical item is represented as a vector.

Following Grand et al. (2018)'s framework, we construct semantic domain networks using embeddings. We use Jaccard similarity between k nearest neighbors of the lexical items to detect semantic similarity as supported by Gonen et al. (2020) as a stable and interpretable method for detecting semantic relationships. The study innovates by using these similarities to create a semantic domain structure to enable a more nuanced analysis of cross-regional variation. This use of embedding space to show dialectal and regional differences is seen in Dunn (2023), which demonstrates that the stability of embeddings vary significantly across geographically distinct corpora. This drives our city wise analysis of domain structures, allowing us to visualize how lexical items are used within domains and how this relations can be similar or different across regions.

2.2 Social Factors

The study of language contact has been studied through looking at the processes of borrowing, code-switching, and interference. Current research looks into how the intermingling of languages in a multilingual society has led to more complex language contact. With more speakers becoming multilingual, choosing a specific language for communication can also be linked to social identity (Tajfel, 1979). Therefore, in this study, a person's choice between using English, Hindi or Transliterated Hindi is not only limited to languages they know but can also extend to this theory of which social group they prefer to belong to. Factors which could contribute to this social identity could be social factors such as education, urbanization and gender. This study explicitly considers several social factors and how it impacts language use.

Urbanization: Urban areas are usually more linguistically diverse and have a higher amount of language contact. This is due to higher migration into the cities which leads to more contact between different communities. Furthermore, the more urban the city, the higher the access of to multilingual education and connectivity to the internet and multilingual media. The percentage of urban population is therefore a relevant factor in understanding the prevalence of language mixing on social media. Language contact has actually also been used to study urban cities (Chríost and Thomas, 2008; Peukert, 2013).

Literacy and Education: Higher literacy rates usually suggest an increased access to and engagement with online platforms. Furthermore, education level, especially in India since English is not everyone's native language, influences proficiency in English, impacting it's degree of use in online communication (Bhatt, 2008).

Regional Language Influence: India's diverse linguistic landscape could influence semantic variation in online communication. Khubchandani (1983) emphasizes the role of interference and code switching in multilingual communication. Therefore, we include number of Hindi speakers, whether Hindi is the 1st/2nd or 3rd language of any speakers, whether English is the 1st/2nd or 3rd language of any speakers and whether Hindi is the state's official language or not to our analysis. We do focus on these metrics as we are looking into English and Hindi data and our census data (2011) has limitations.

Gender and Language: Gender has a role to play in how language is used and changes over time are also driven by gender (Gordon, 2003; Eckert and McConnell-Ginet, 2003, 2013). Therefore, this study considers the gender distribution among Hindi speakers, sex ratio of the city and literacy rate by gender as important factors to understand the language use on social media.

3 Data

This study uses a large scale social media corpus containing 49,801,176 English tweets and 5,545,724 Hindi tweets, all originating from India.

It is important to note that the English corpus contains a huge amount of Transliterated Hindi data which is often used by people in the region especially for online communication. To analyze the data, we combined the two corpora into a single embedding space to capture semantic representations across languages. The resulting corpus contains approximately 55 million documents.

3.1 Cities

Each tweet in the corpus includes geo-location metadata, which we will use to study regional variation in lexical semantics. However, the full corpus contains data from 100 cities. To ensure we represent different regions of India and yet maintain complexity, we selected 13 cities with the highest tweet volume. Our thirteen cities are spread across India as shown in Figure 1. It is important to note that we also made sure that we chose cities with a similar level of connectivity with the internet in order to ensure uniformity. Table 1 shows the document count and regional classification for each of the selected cities.

Figure 1: Map of India with states marked for each of the thirteen chosen cities.

3.2 Semantic Domains

We analyze six semantic domains: animals, kinship, weather, professions, emotions and temporal units. These domains were selected based on their cross linguistic variation and sufficient lexical coverage in the corpus due to high usage in day-to-day life. All domains include Hindi, English and Transliterated Hindi lexical terms. The appendix mentions the whole list of lexical items used for this study.

City	Location	Count
Bengaluru	South West	2613502
Mumbai	Western Peninsular	2269945
Delhi	North	2191038
Chennai	South East	1367075
Hyderabad	South	1315379
Kolkata	East	1297674
Thane	Western Peninsular	1126080
Pune	Western Peninsular	1020440
Srinagar	North	989439
Chandigarh	North	967845
Jaipur	North West	663712
Patna	North East	559755
Lucknow	North	545069

Table 1: Count of documents from each of the thirteen chosen cities.

Table 2 shows the number of unique lexical items used per semantic domain. We ensured relatively balanced lexicons across domains to support a more robust comparison of semantic structures.

Domain	Number of Data Points
Kinship	113
Animals	282
Time	158
Professions	191
Weather	133
Emotions	138

Table 2: Number of unique lexical items for each semantic domain.

3.2.1 Kinship

Kinship is a domain in which the Hindi system is significantly more granular than the English system. Key difference include different terms for maternal vs. paternal relatives (e.g. grandparents, aunts, uncles), and specific terms for paternal uncles based on their age relative to the father. Similarly, grandchildren, nieces and nephews are also distinguished lexically based on lineage.

3.2.2 Animals

In the animal domain, Hindi often differentiates between male and female animals lexically, more extensively than in English. For example, Hindi users refer to a male cat as 'billa' (transliterated) and a female cat as 'billi' (transliterated) whereas English users typically use a gender neutral term, "cat".

3.3 Temporal Terms

This domain includes terms used for telling time such as days of the week, months of the year and terms to refer to days such as today, yesterday and tomorrow. A few notable distinctions between Hindi and English include:

1. 'kal' is used for both tomorrow and yesterday in Hindi.

2. 'parso' (transliterated) is used for both day after tomorrow and day before yesterday in Hindi.

3. For time, Hindi has specific words for 1:30 and 2:30 which do not include numerals and specific words for quarter past, half past and quarter to.

The remaining domains offer supplementary data for regional comparison, despite showing less lexical variation.

4 Methodology

The basic approach in this paper is to infer a representation of six semantic domains from population-specific corpora representing different cities in India. Once we have inferred these semantic domains, we compare them to one another in order to quantify variation and then use regression models to understand the relationship between these variations and external factors like language contact.

4.1 Inferring Semantic Domains

Because we are interested in the usage of a bilingual speech community, we combine both the Hindi (in any orthography) and the English data together. Our baseline dataset contains data from over one hundred Indian cities; this is used to infer average or non-population-specific semantic domains. Our test datasets, on the other hand, are drawn from thirteen individual cities. The idea is to compare these city-specific domains to the average domain as a means of quantifying the amount of semantic variation within these domains.

Given these corpora (the baseline corpus and the thirteen city-specific corpora), we then learn character-based embeddings using fastText in order to represent general semantic relationships between lexical items. We do not use pre-trained LLMs for creating these embeddings spaces because there is not sufficient data to do so while representing only city-level populations. Relying on models trained on outside data would risk contaminating the city-level semantic domains with information derived from the broader population.

At this stage we have distinct embedding spaces for each city-level population and for the country as a whole. The next task is to create maps or networks representing each of the six semantic domains using this embedding space. First, we manually curate the lexical items for each domain, drawing from both English and Hindi. These terms are found in Appendix A. For example, the kinship domain includes both English terms like *grandmother* and *aunt* as well as Hindi terms like *parivaar* and *mausa*. We create a network out of this domain-specific vocabulary using Jaccard similarity: each lexical item is a node in the network and the Jaccard similarity quantifies the edge weights between nodes. For instance, we would expect that *grandfather* is closer to *grandmother* than it is to *niece*. Jaccard similarity in this context is calculated by using cosine similarity to retrieve the n nearest neighbors for each word (here, $n = 1000$). High set similarity is then reflecting the fact that two words are located in the same neighborhood within the domain.

From a Saussurian perspective (de Saussure, [1916] 1983), the meaning and value of each word can be taken from the relationships within this graph. In other words, the meaning of *grandfather* is derived purely from its relationships with other items in the same kinship domain. These domains are then jointly defined by (i) using prior knowledge to select the relevant lexical items and (ii) using an embedding space to estimate edge weights.

To summarize, then, we operationalize semantic domains as networks by, first, learning a character-based embedding from each city-specific corpus and, second, using nearest neighbors in this embedding space to calculate the distance between nodes, where a node is a domain-specific lexical item. One challenge with character-based embeddings is that they can exhibit instability, reaching

different neighborhoods across multiple random initializations. We thus conduct a stability analysis to ensure that these inferred networks are reliable representations of each domain.

To test robustness, we re-ran the full pipeline for each city and each semantic domain ten times taking different sub samples of the data. Across all runs, the models consistently produced the highly similar network maps, with only minimal variation. This indicates that the inferred networks are stable.

4.2 Comparing Cities

Once we attained the Jaccard similarity between lexicon items for each domain for each city, we compared the similarity matrix between cities by calculating the mean square difference. This gave us a quantifiable difference between the structures, making it easier to group cities as being similar or different from each other. This resulted in a matrix which contained the mean square difference between the cities. We took the correlation of this matrix and then compared this between domains to see whether cities which have similar domain structures for one domain have similarity in lexicons across domains or not.

4.3 Social Features

After getting a matrix of similarities between cities across domains, we look at social features which could cause certain cities to have similar structures. Social features included the percentage of English (as a 1st, 2nd or 3rd language) and Hindi (as a 1st, 2nd or 3rd language as well as just as the mother tongue) speakers in the city, literacy rates and percentage of urban area/population. For the number of Hindi speakers and literary rates, we further got gendered data. This data is extracted from the Census of India (2011) which was published in 2018 (where city data is not available state data was used). Linear regression was conducted to see how these factors correlated to similarities/differences between the cities and the national average semantic structure.

5 Analysis

Figure 2 shows us the average correlation matrix for the mean square differences in the mappings between cities for our six domains. Here i is the taken as the national average. A correlation close to 1 shows high positive correlation shown by dark red. This signifies that the two cities had similar mean square differences for that domain suggesting a similar structure. A correlation close to -1 shows high negative correlation between the cities shown by dark blue. This shows that the cities have very different structures as their mean square differences compared to other cities in our matrix are not very similar and are quite contrasting. A value close to 0 suggests that there is no correlation between the cities. We want to observe whether there are any significant differences in the structure of semantic domains as operationalized. This would be seen if our correlation matrix has a range of values from -1 to 1 as this would suggest that each city has some difference in structure. However, if we see the same very extreme values that would suggest that all cities are correlated meaning that all structures look the same (uniformity in semantic structures). On the other hand, if we see no correlation (values just ranging near 0), that would suggest that there is no commonality in any of the structures and all cities have a very different way to portray the lexicon in the embedding space. We decided to average out and create one matrix for our analysis. This is because the matrices for each of the domains had similar values. These matrices can be seen in appendix.

5.1 Across Domains

Across domains, we see the following clusters:

1. Bengaluru, Hyderabad, Kolkata, Pune and Thane
2. Delhi, Jaipur, Lucknow and Patna
3. Chennai and Srinagar
4. Chandigarh and Mumbai
5. Mumbai and Thane

Figure 3 shows these clusters. It is important to note that the map marks the states instead of the cities to show neighboring states in an easier manner. We also see a pattern of Chennai having the most correlated structures to the national average and Pune having the least correlated structures. Our analysis shows that the domain structure changes across regions. This could be due to language contact with other languages which occurs in those states and also bleeds to neighboring states. Overall, this suggests that there is a meaningful difference in the structure of different cities and this difference is seen uniformly across domains as our clusters rarely

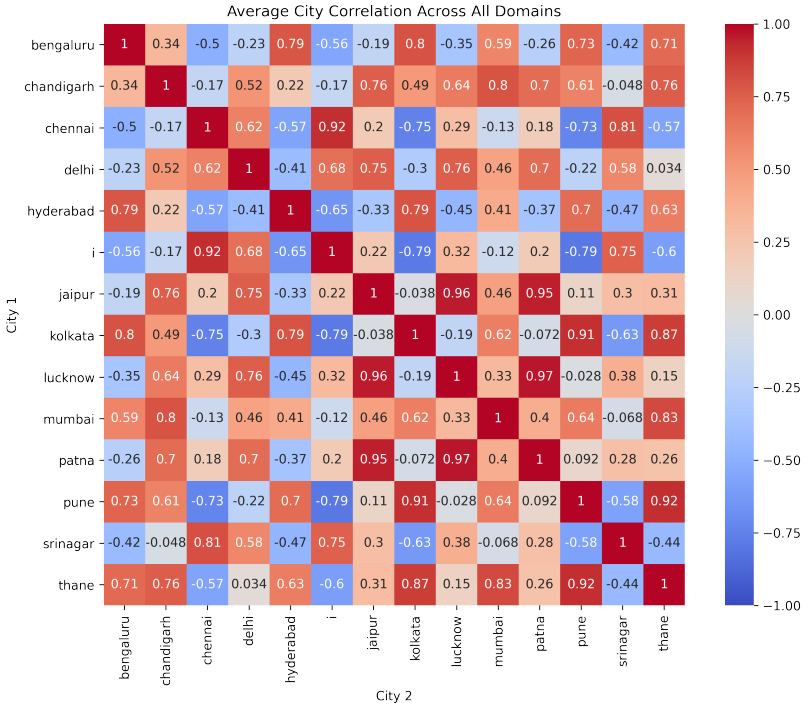

Figure 2: Average Correlation Matrix showing similarities and differences between cities across all domains. Here a value close to -1 suggests negative correlation between those cities (very different mappings) and a value close to +1 suggests positive correlation between those cities (very similar mappings). Here i refers to the national average.

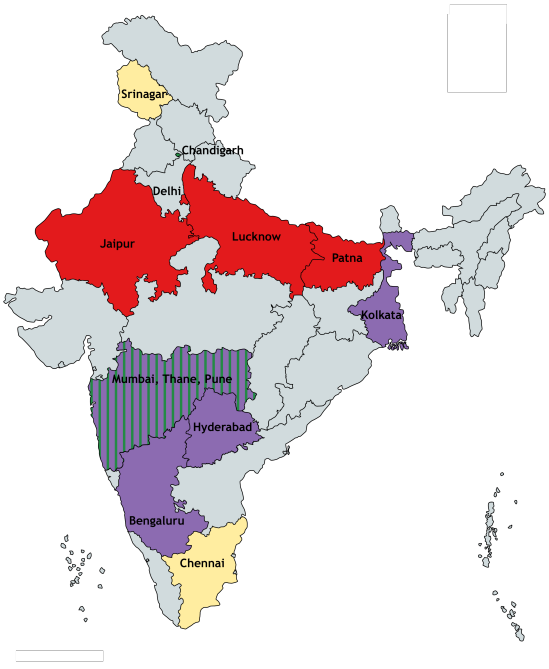

Figure 3: Groupings of positive correlation across domains on the Indian map. Here different color suggests that they have different semantic mappings and same color suggests that they have very similar semantic mappings

change in our analysis of the six domains. This suggests that geographic distance between cities impacts semantic representations.

5.2 Social Factors

We performed linear regression to examine whether a city's deviation from the national average in semantic structure could be explained by social factors. Prior to regression, all social variables were normalized to ensure comparability across scales, especially between large values (e.g. population) and percentage-based features (e.g. literacy rate). Across our domains several social features consistently contributed to predicting semantic similarity/conformity to national average in our linear regression model. These features include:

- **Literacy Rate (Overall)**: Consistently the strongest *positive* predictor across all domains. Higher overall literacy in the city is strongly associated with greater semantic conformity to the national average.
- **Literacy Rate (Male and Female)**: When overall literacy is excluded, male and female literacy show large but opposing effects - male literacy is strongly positive while female literacy is strongly negative. This suggests male

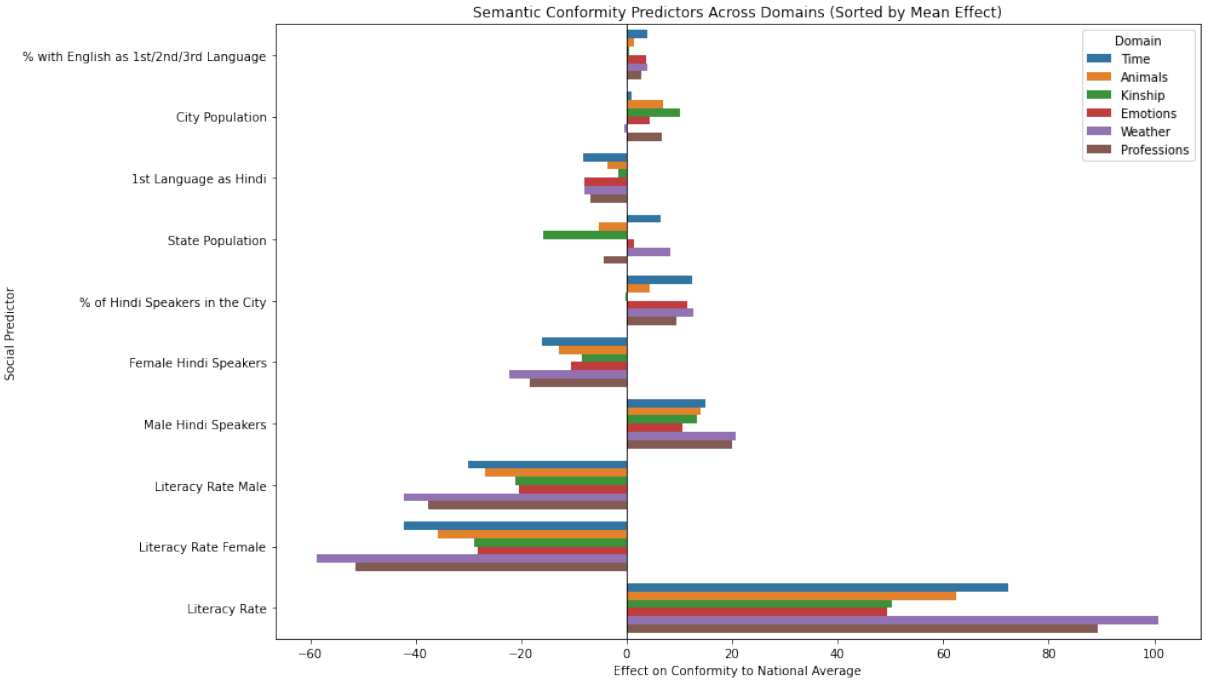

Figure 4: Social predictors of semantic conformity to the national average across domains. Positive values indicate that an increase in the feature contributes to semantic similarity with the national average.

literacy reinforces conformity, while higher female literacy correlates with divergence, consistent with the idea that women may innovate away from norms. The opposing effects also underscore multicollinearity with overall literacy.

- **Male Hindi Speakers**: Strong and consistent *positive* predictor - cities with more male Hindi speakers show greater semantic similarity to national patterns.
- **Female Hindi Speakers**: Strong *negative* predictor - possibly indicating gendered variation in language use and exposure that diverges from national norms.
- **Percentage of Hindi Speakers**: A clear *positive* influence - more Hindi presence overall contributes to semantic conformity.
- **State and City Population**: Population effects are domain -specific. Larger cities generally show a positive effect, suggesting that urban centers mirror national semantic patterns. By contrast, state population often has a negative effect in domains such as Kinship and Professions, likely reflecting the greater rural urban diversity within populous states. Thus, while cities may exert a homogenizing influence, states capture broader variation that diverges from national norms.
- **1st Language as Hindi**: Moderate *negative* effect - cities with Hindi as a first-language are somewhat less similar to the national average.
- **English (1st/2nd/3rd Language)**: Mild *positive* correlation - increased multilingualism including English is weakly associated with conformity.

These results suggest that social variables - particularly literacy, language exposure, and city/state population size - significantly shape how closely a city's semantic patterns align with the national average. The gendered contrast between male and female Hindi speakers and literacy rates, in particular, reveals complex sociolinguistic dynamics and do agree with the idea that woman diverge from norms and hence could be the drivers for language change and regional language usage.

6 Conclusion

This study investigated how the mapping between concepts and lexical items within specific semantic domains varies across geographical regions within India. Our analysis revealed that semantic mappings do vary across regions, offering insight into

language contact and multilingual variation in online communication. It showed consistent clusters of cities which had similar semantic structures. This suggests that geographic proximity influences variation in these representations. These clusters show that semantic similarity is influenced by spatial and social features. Notably, Chennai, despite not being Hindi-dominant, showed the highest similarity to the national average, while Pune showed the least, indicating that the national semantic norm reflects more than just northern, Hindi-speaking patterns.

The linear regression model showed that social features correlate closely with semantic conformity of a domain structure to the national average.

1. **Literacy rates and Number of Hindi speakers** both showed gendered divergence. High **male rates** correlated with a strong *positive* effect on semantic conformity, while **female rates** shows a strong *negative* effect. These opposing effects suggest distinct linguistic networks across gendered populations. Furthermore, it suggests that male population conforms to national average whereas female population might be driving diverse language change.

2. **State and city population effects diverge across domains**: **City population** generally shows a mild *positive* effect on semantic conformity in most domains (e.g., Kinship, Time), suggesting that larger urban centers may trend toward standardized usage. In contrast, **state population** shows a more *inconsistent* or even *negative* effect in domains like Kinship and Professions, indicating that broader regional demographics do not always align with national patterns and reflect greater internal linguistic diversity.

Overall, this study contributes to our understanding of how language contact and social features shape semantic domain structure and lexical semantics in multilingual online spaces. Our methodology - creating semantic networks from embedding spaces and enriching them with social predictors - offers a novel framework for studying semantic variation especially in multilingual settings. By offering a structured approach to examining how semantic variation aligns with regional and social characteristics in multilingual settings, this study can inform more personalized language technologies and educational resources.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Rakesh M. Bhatt. 2008. In other words: Language mixing, identity representations, and third space. *Journal of Sociolinguistics*, 12(2):177–200.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea. 2021. Analyzing the surprising variability in word embedding stability across languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5891–5901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Diarmait Mac Giolla Chríost and Huw Thomas. 2008. Linguistic diversity and the city: Some reflections, and a research agenda. *International Planning Studies*, 13(1):1–11.

Jonathan Dunn. 2023. Variation and instability in dialect-based embedding spaces. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 67–77, Dubrovnik, Croatia. Association for Computational Linguistics.

P. Eckert and S. McConnell-Ginet. 2003. *Language and Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*, 2 edition. Cambridge University Press.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Matthew J. Gordon. 2003. Principles of linguistic change: Social factors, volume 2. *American Anthropologist*, 105:436–437.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings.

Government of India. 2011. Census of india. Technical report, Government of India.

Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Word2vec model analysis for semantic similarities in english words. *Procedia Computer*

Science, 157:160–167. The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society.

Lifeng Jin and William Schuler. 2015. A comparison of word similarity performance using explanatory and non-explanatory texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 990–994, Denver, Colorado. Association for Computational Linguistics.

Lachman M. Khubchandani. 1983. *Plural Languages, Plural Cultures*. University of Hawaii Press, Honolulu.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI Conference on Artificial Intelligence*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Hagen Peukert. 2013. Measuring language diversity in urban ecosystems. *Linguistic Superdiversity in Urban Areas*, pages 75–95.

Abhishek Rana, Akshita Pant, Nikita Rawat, Priyanshu Rawat, Satvik Vats, and Vikrant Sharma. 2024. Semantic similarity analysis using fasttext. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, pages 454–460.

Ferdinand de Saussure. [1916] 1983. *Course in General Linguistics*. Duckworth, London. (trans. Roy Harris).

Henri Tajfel. 1979. An integrative theory of intergroup conflict. *The social psychology of intergroup relations/Brooks/Cole*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

A Appendix

A.1 Lexicon

A.1.1 Kinship Terms

English: grandmother, grandfather, aunt, uncle, mother, father, sister, brother, niece, nephew, daughter, son, granddaughter, grandson, cousin, husband, wife, father-in-law, mother-in-law, brother-in-law, sister-in-law, children, brother-in-law's wife, son-in-law, daughter-in-law

Transliterated Hindi: parivaar, nani, nana, dadi, masi, mausa, mummy, pita, papa, bua, chacha, bhua, tau, tauji, behen, bhai, didi, bhaiya, bhaanji, bhaanja, bhajiti, bhatija, beti, beta, naatin, naati, pota, poti, pati, patni, sasur, saas, devar, nanad, bacche, jeeja, devrani, saala, daamaad, bahu

Hindi: नानी, नाना, दादी, दादा, मासी, मौसी, मामा, मम्मी, माँ, पिता, पापा, बुआ, चाचा, ताऊ, बड़े पापा, बहन, भाई, दीदी, भैया, भांजी, भांजा, भतीजी, भतीजा, बेटी, पुत्री, बेटा, पुत्र, नतिनी, नातिन, नाती, पोती, पोता, पति, पत्नी, ससुर, सास, देवर, ननद, बच्चे, जीजा, देवरानी, साला, दामाद, बहू

A.1.2 Animal Terms

English: Hyena, Dove, goat, Snail, monkey, Mosquito, Crocodile, Lizard, Earthworm, camel, Tortoise, Myna, Turtle, Fish, Caterpillar, Bugs, Birds, Deer, Leopard, Lioness, Sheep, Goose, Pig, Wolf, Seahorse, Bat, mouse, Insect, Bear, Panther, Sealion, Fox, Donkey, Spider, Housefly, elephant, Honeybee, Butterfly, Snake, Gander, Cuckoo, Mongoose, Buffalo, Grasshopper, Hen, Lion, Animal, Aquatic, Kite, Weaverbird, Rabbit, Duck, Alligator, Woodpecker, Chameleon, Squirrel, Eagle, Octopus, Cricket, Pet, Guinea pig, Cow, Giraffe, Tiger, Tigress, Pigeon, Prawns, Whale, Dolphin, dog, Horse, Bird, Shark, Hawk, Parrot, Insects, Hippopotamus, Owl, cat, Jellyfish, Oyster, Mammals, Vulture, Cockroach, Ant, Frog, Crow, Rooster, Wild

Transliterated Hindi: tidda, lomdi, chipakali, madhumakkhii, jiraaf, shernii, hiran, sher, ghonghaa, totaa, safed kabuuTar, kiida, bhaalu, mendhak, gae, chidiyaa, suar, bakri, hathinii, saanp, chuuhaa, haangar, ashtabahu, kauvaa, battakh, murgii, chuhiya, lakadbagghaa, baaz, jhiingur, gini pig, titli, bhed, billi, kabuuTar, paaltoo, bandariya, bakraa, jeliifish, ullu, jhiinga machhli, gauraiyaa, tilacchattaa, vhel, jalsinh, samudri ghodaa, oont, makri, girgit, kharagosh, oontnii, bhediyaa, siip, giddh, daryaai ghodaa, haathi

Hindi: घोंघा, भेड़, पालतू, कीड़ा, हाथी, कछुआ, जंगली, साँप, मुर्गा, ऊँटनी, लोमड़ी, मक्खी, सियार, केंचुआ, तिलचट्टा, घड़ियाल, गधा, चुहिया, बंदरिया, हंसिनी, मधुमक्खी, ख़रगोश, घोड़ा, बया, बकरी, गिद्ध, बिल्ली, कीड़े, बाघ, भालू, जेलीफ़िश, बंदर, तितली, ऊँट, मकड़ी, गिरगिट, बकरा, झींगा मछली, व्हेल, कुत्ता, भेड़िया, हाँगर, चील, चूहा, मैना, तेंदुआ, बत्तख़, गौरैया, समुद्री घोड़ा, भैंस, शेरनी, बिल्ला, गाय, मेंढक, मछली, गिनी पिग, चींटी, कबूतर, जिराफ़, मच्छर, लकड़बग्घा, तारामीन, गिलहरी, छिपकली, जानवर, कुतिया, मुर्गी, सुअर, मगरमच्छ, हथिनी, शेर, दरियाई घोड़ा, जलसिंह, चिड़िया, हिरण, इल्ली, कठफोड़वा, कौवा, अष्टबाहु, झींगुर, नेवला, कोयल, तोता, हंस, स्तनधारी, समुद्री, सूंस, बाघिन, सफ़ेद कबूतर, चमगादड़, सीप, बाज, उल्लू, टि–

ड्डा

A.1.3 Time Terms

English: Second, Minute, Hour, Day, Week, Month, Year, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, January, February, March, April, May, June, July, August, September, October, November, December, Morning, Afternoon, Evening, Night, Midnight, Yesterday, Today, Tomorrow, Day before yesterday, Day after tomorrow, Now, Later, Earlier, O'clock, Half past, Quarter past, Quarter to, Always, Often, Sometimes, Rarely, Never, For a short time, For a long time, Since, Until

Transliterated Hindi: sekand, minat, ghanta, din, saptah, saptaah, mahina, saal, varsh, ravivar, somvar, mangalvar, budhvar, guruvaar, shukravar, shanivar, janavari, pharavari, march, aprail, joon, julai, agast, sitambar, aktoobar, navambar, disambar, subah, dophar, shaam, raat, aadhi raat, kal, aaj, kal, parson, abhi, baad mein, pehle, baje, saade, sava, paune, hamesha, aksar, kabhi-kabhi, shayad hi kabhi, kabhi nahi, thodi der ke liye, lambe samay tak, se, tak

Hindi: सेकंड, मिनट, घंटा, दिन, सप्ताह, महीना, साल, रविवार, सोमवार, मंगलवार, बुधवार, गुरुवार, शुक्रवार, शनिवार, जनवरी, फरवरी, मार्च, अप्रैल, मई, जून, जुलाई, अगस्त, सितंबर, अक्टूबर, नवंबर, दिसंबर, सुबह, दोपहर, शाम, रात, आधी रात, कल, आज, कल, परसों, परसों, अभी, बाद में, पहले, बजे, साढ़े, सवा, पौने, हमेशा, अक्सर, कभी–कभी, शायद ही कभी, कभी नहीं, थोड़ी देर के लिए, लंबे समय तक, से, तक, वर्ष

A.1.4 Weather Terms

English: Sun, Rain, Wind, Snow, Cloud, Weather, Hot weather, Cool weather, Pleasant weather, Weather change, Weather forecast, Seasons, Spring, Winter, Summer, Autumn, Rainy, Temperature, Hot, Humid, Cold, Moisture, Scorching, Sunshine, Sunrise, Sunset, Sky, Cloudy, Rainbow, Drizzle, Storm, Cyclone, Lightning, Fog, Dew, Snowfall, Hail

Transliterated Hindi: Sooraj, Baarish, Hawaa, Baraf, Baadal, Mausam, Garam Mausam, Thanda Mausam, Suhaana Mausam, Mausam Parivartan, Mausam Purvaanumaan, Rituyen, Vasant Ritu, Sardi, Thand, Shishir, Sheet Ritu, Grishm Ritu, Patjhad, Sharad Ritu, Barsaat, Varsha Ritu, Hemant Ritu, Taapmaan, Paara, Aardrataa, Thandak, Nami, Chilchilaatii, Dhuup, Suryodaya, Suryaast, Aasmaan, Aakaash, Boondaabaandi, Phuhaar, Tufaan, Chakravaat, Bijli, Kohraa, Os, Barfbari, Ola Vrishti

Hindi: सूरज, बारिश, हवा, बरफ, बादल, मौसम, गर्म मौसम, ठंडा मौसम, सुहाना मौसम, मौसम परिवर्तन, मौसम पूर्वानुमान, ऋतुएं, मौसम, वसंत ऋतु, सर्दी, ठंड, शिशिर, शीत ऋतु, गर्मी, ग्रीष्म ऋतु, पतझड़, शरद् ऋतु, बरसात, वर्षा ऋतु, हेमंत ऋतु, तापमान, पारा, गर्मी, उमस, आर्द्रता, ठंडक, नमी, चिलचिलाती, धूप, सूर्योदय, सूर्यास्त, आसमान, आकाश, बदली, इंद्रधनुष, बूंदाबांदी, फुहार, तूफ़ान, चक्रवात, बिजली, कोहरा, ओस, बर्फ़बारी, ओला वृष्टि

A.1.5 Emotion Terms

English: blissful, brave, careful, cautious, clever, curious, excited, friendly, glad, good, great, happy, innocent, interesting, optimistic, pleasant, pleased, proud, quiet, satisfied, sensible, serious, surprised, angry, arrogant, awful, bad, bored, crazy, disappointed, exhausted, frightened, sad, guilty, helpless, hurt, lonely, mad, miserable, nervous, shocked, sheepish, silly, strange, terrible, upset

Transliterated Hindi: anandmay, bahaadur, saavadhan, satark, chaalak, utsuk, uttejit, mitravat, prashann, achcha, mahaan, khush, nirdosh, dilchasp, aashavaadi, sukhad, santusht, garvit, shaant, samajhdaar, gambhir, haeraan, naaraaj, abhimaani, daraavana, bura, uba hua, sanki, niraash, thaka, bhayabhit, dukhi, doshi, asahaay, aahat, akela, paagal, abhaaga, ghabaraaya hua, haeran, sharminda, murkh, anokha, bhayaanak, pareshan

Hindi: आनंदमय, बहादुर, सावधान, सतर्क, चालाक, उत्सुक, उत्तेजित, मित्रवत, प्रसन्न, अच्छा, महान, खुश, निर्दोष, दिलचस्प, आशावादी, सुखद, सन्तुष्ट, गर्वित, शांत, संतुष्ट, समझदार, गंभीर, हैरान, नाराज, अभिमानी, डरावना, बुरा, ऊबा हुआ, सनकी, निराश, थका, भयभीत, दुखी, दोषी, असहाय, आहत, अकेला, पागल, अभागा, घबराया हुआ, हैरान, शर्मिंदा, मूर्ख, अनोखा, भयानक, परेशान

A.1.6 Profession Terms

English: Butcher, Florist, Travel agent, Scientist, Gardener, Mason, Pilot, Librarian, Model, Shop assistant, Bus driver, Real estate agent, Lawyer, Cook, Fireman, Poet, Poetess, Soldier, Receptionist, Designer, Fire fighter, Fisherman, Waitress, Actress, Author, Dentist, Shop keeper, Traffic warden, Baker, Journalist, Judge, Actor, Plumber, Secretary, Veterinary doctor, Farmer, News reader, Craftsman, Lifeguard, Photographer, Taxi driver, Carpenter, Optician, Accountant, Teacher, Electrician, Postman, Tailor, Painter, Policeman, Engineer, Hairdresser, Policewoman, Nurse, Doctor, Mechanic, Translator, Politician, Lecturer, Waiter, Workers,

Cleaner, Pharmacist

Transliterated Hindi: Machuaara, Anuvaadak, Chashma Banane Wala, Phoolwala, Naanbai, Sachiv, Shramik, Samachar Paadak, Vaastukar, Sramik, Model, Nalsaaj, Maarjak, Sipaahi, Svaagati, Lekhak, Kavi, Vakil, Aushadhajny, Badai, Baayara, Abhinetri, Yaatra Agent, Abhiyanta, Bas Chalak, Daakiya, Vigyaanik, Sainik, Dukan Sahayak, Sharir Raksak, Rajnitigy, Rajgir, Viman Chalak, Granthaagarik, Bhumi Bhavan Abhikarta, Nyaayadhish, Chitrkar, Abhineta, Kasaai, Mechanic, Shikshak, Dukandar, Shilpkar, Naai, Yaatayaat Nirikshak

Hindi: बायरी, कवियित्री, डाकिया, पत्रकार, वकील, मॉडल, अग्निशामक कर्मचारी, दंत चिकित्सक, दुकानदार, नानबाई, भूमि भवन् अभिकर्ता, रसोइया, बिजली मिस्त्री, बढ़ई, मार्जक, राजगीर, नलसाज, कवि, ग्रंथागारिक, रूपकार, अभिनेत्री, वैज्ञानिक, मुनीम, औषधज्ञ, विमान चालक, बायरा, मैकेनिक, सचिव, दुकान सहायक, स्वागती, यात्रा एजेंट, नाई, अभिनेता, चित्रकार, मछुआरा, माली, शिल्पकार, फूलवाला, लेखक, समाचार पाठक, नर्स, यातायात निरीक्षक, फोटोग्राफर, शरीर रक्षक, अनुवादक, पशु चिकित्सक, श्रमिक, किसान, वास्तुकार, महिला सिपाही, दर्जी, टैक्सी चालक, शिक्षक, अभियन्ता, कसाई, राजनीतिज्ञ, सैनिक, चिकित्सक, बस चालक, चश्मा बनाने वाला, न्यायाधीश, सिपाही, व्याख्याता

A.1.7 Domain wise Analysis

A.2 Domains

Fig. 5 shows the correlation matrices across the thirteen cities for all six domains.

A.2.1 Time

Time matrix shows positive correlation between cities of (taking a cut off of 0.81)

1. Bengaluru, Hyderabad, Kolkata, Pune and Thane
2. Chandigarh and Mumbai
3. Chennai and Srinagar
4. Delhi, Patna, Lucknow and Jaipur

Mumbai and Thane also have high positive correlation but the other members of those groups do not. Compared to our average (India), Chennai and Delhi have high positive correlation and Hyderabad and Kolkata have high negative correlation.

A.2.2 Weather

Weather matrix shows positive correlation between cities of (taking a cut off of 0.81)

1. Bengaluru, Hyderabad, Kolkata, Pune and Thane
2. Chandigarh and Mumbai
3. Delhi, Patna, Lucknow and Jaipur

Again, Mumbai and Thane also have high positive correlation but the other members of those groups do not. Chennai has high negative correlation with (a) and Srinagar has no strong correlations with any groups. Compared to our average (India), Chennai has high positive correlation and Hyderabad, Bangalore, Pune and Kolkata have high negative correlation.

A.2.3 Animals

Animals matrix shows positive correlation between cities of (taking a cut off of 0.81)

1. Bengaluru and Kolkata
2. Chandigarh, Pune and Thane
3. Patna, Lucknow and Jaipur

Again, Mumbai and Thane also have high positive correlation but the other members of those groups do not. Compared to our average (India), Chennai has high positive correlation and Pune has high negative correlation.

A.2.4 Kinship

Kinship matrix shows positive correlation between cities of (taking a cut off of 0.81)

1. Kolkata, Pune and Thane
2. Patna, Lucknow and Jaipur

Again, Mumbai and Thane also have high positive correlation but the other members of those groups do not. Compared to our average (India), Chennai has the highest positive correlation and Pune has highest negative correlation.

A.2.5 Emotions

Emotions matrix shows positive correlation between cities of (taking a cut off of 0.81)

1. Bengaluru, Hyderabad, Kolkata, Pune and Thane

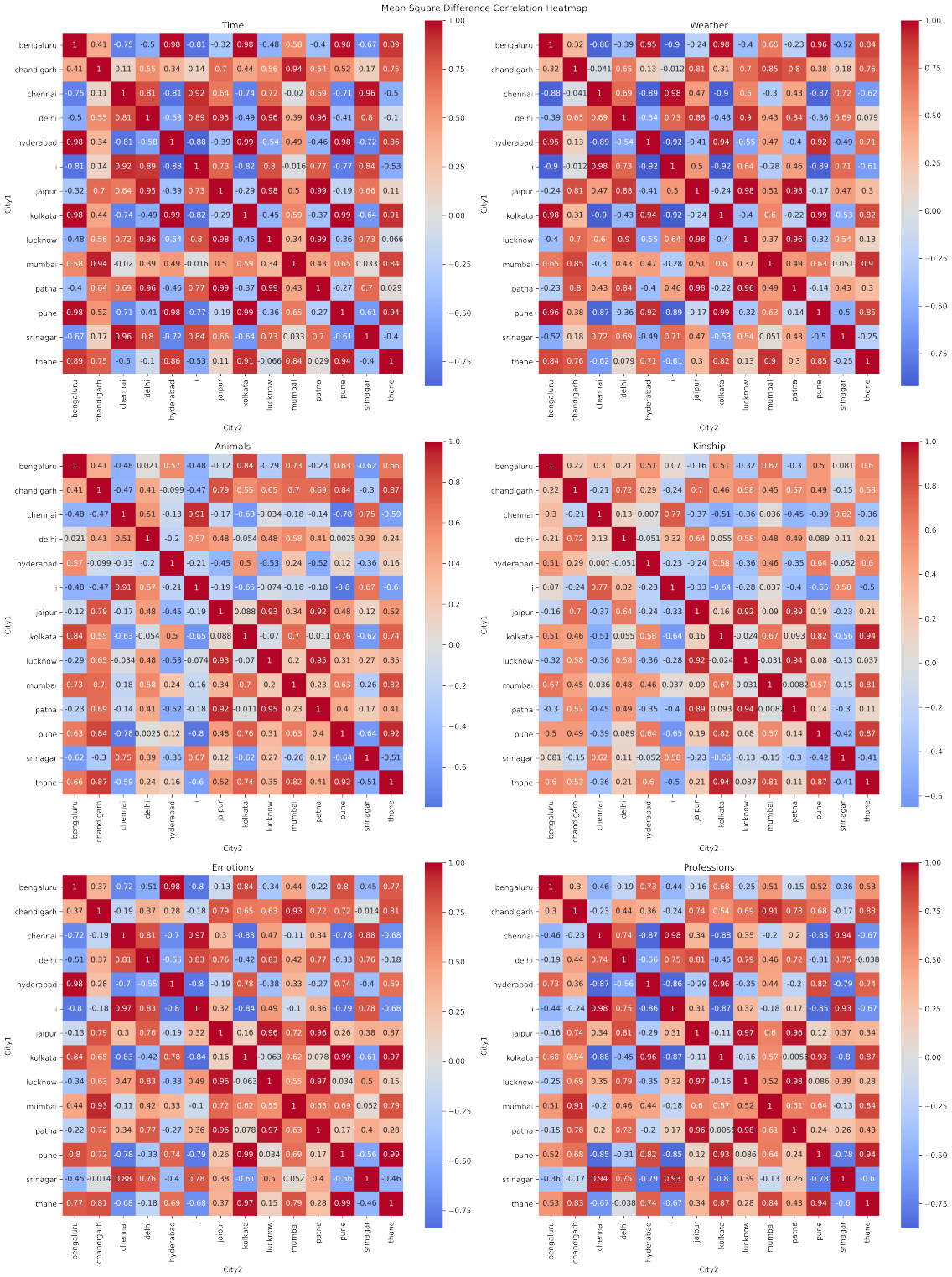

Figure 5: Correlation matrices of mean square differences for all domains. Here a value close to -1 suggests negative correlation between those cities (very different mappings) and a value close to +1 suggests positive correlation between those cities (very similar mappings)

2. Chandigarh and Mumbai, Thane

3. Chennai and Srinagar, Delhi

4. Delhi, Jaipur, Lucknow and Patna

Compared to our average (India), Chennai and Delhi have high positive correlation and group (a) have high negative correlation.

A.2.6 Professions

Professions matrix shows positive correlation between cities of (taking a cut off of 0.81 with atleast two cities of the group)

1. Hyderabad, Kolkata, Pune and **Thane**

2. Chandigarh, Mumbai and **Thane**

3. Chennai and Srinagar

4. Delhi, Jaipur, Lucknow and Patna

Bengaluru does not have any high correlation with other cities. Compared to our average (India), Chennai and Srinagar have high positive correlation and Pune, Kolkata and Hyderabad have high negative correlation.