Analysis Model

Given: Dalit politics corpus, feminism corpus, in intermixed Devanagari and Roman script.

We can perform the following analyses:-

1. Analysis of the occurrence of a list of ideological words, appearing in their Hindi and English equivalents, to determine the trends in ideological discussion.
2. Sentence-based analysis of relative Hindi presence.
3. Word-based analysis of relative Hindi presence
4. Identify tweets with CS.
5. Identify run length of CS when it exists: estimate whether it is phrasal or at the sentence level.

Methodologies:

1. The number of tweets we get for each is a general indicator of the relative strength and popularity of these two conversations on Twitter in India.
2. This is pretty straightforword. Prepare the list, run a counting search on non-LID’d data.

LID:

For the next four, we will LID the corpora. We will also make certain assumptions(?):

Assumption 1:-

There are very few English sentences written entirely in the Devanagri script, (although we might have long segments of Roman transliterated Hindi). This seems true from observation in the corpora – it is also fairly intuitive, since the English – Hindi Roman – Devanagri baseline for Twitter is 70-15-15; it is probably rare that people typing in the Devanagri script are expressing themselves in English. In any case, we can observe our own corpus to validate this assumption.

Assumption 2 :-

There is very little *script*-mixing in one sentence. Again, this is true from observation. People do insert English acronyms (SC, BJP, etc.) into their Devanagri tweets, but they do not switch scripts halfway through a sentence, or for particular phrases.

(Possibility: We can remove the sentence which have script mixing. Will this matter?)

The necessity for the above two assumptions arise from the limitation that we do not have an en-hi LID software for Devanagri script. However, both the assumptions are in fact reasonable, and will probably result in only marginal error.

**In general, we assume that sentences are either in Roman or in Devanagri script, with perhaps some lexical level insertions in the other script, which we can ignore. Furthermore, a sentence in Devanagri script is a Hindi sentence, although it may contain some transliterated English words.**

Now, we LID the corpus, in which: a) the beginning of each new tweet is marked b) The beginning of each new sentence is marked, assuming that a newline indicates end of sentence, as well as question marks, full stops, exclamation marks.

1. We count the number of Devanagari sentences in the corpus, which are simply ignored by the LID, and count them. These are Hindi tweet-sentences. They contain some CS, of course, but we do not at present have the resources or the scope to measure this.

(One method: Run each word by a Hindi dictionary. If no match is found, it must be an English word. This approach has several holes – abbreviations, acronyms, slang, etc. will not show up in a dictionary.)

We LID the Roman script in the corpus.

Assumption 3:-

The language of a sentence is usuallythat in which the larger number of words is present. This tends to be true in general, but not, of course, always – e.g. in the sentence “She told me to come but I said that **Hindi phrase HP**” the sentence is English, no matter how long HP is. However, in most cases, this assumption holds.

Given the LID’d corpus, we do a hi-en comparison for each sentence, and assume its language is that which has greater word presence. We append Hindi count (HR\_count) to our previous Hindi Devanagari HD\_count. This gives us a sentence-based analysis, i.e. analyzing at the lowest unit of a conversation, and ignoring CS for the moment. The reason we are choosing the sentence as a unit and not a tweet is because a multiline tweet may contain two lines, say, in different languages, in which case the dominant language cannot be considered at that containing the most words (in fact, we may say that there is no such language). That is, we are trying to analyse, simplistically, the language in which this conversation is occurring at sentence level.

1. This analysis can only be performed on the LID corpus, since we cannot identify English Devanagri.
2. Once again, we cannot determine this heuristic exactly, for the same reason. If a (multiline) tweet contains en *and* Devanagari script, we may safely assume it contains code-mixing. We assume that purely Devanagari tweets are all-Hindi.
3. Given a line, we can find the run length of CS, if present. We have labelled the language of the line in (2). Typically, a run-length of 3 or more in the opposite language indicates phrasal CS. We can count, in fact, EP type lines and HP type lines: this figure gives us greater indication than (4) as to how fluid the parlance is between Hindi and English i.e. how comfortable are people switching back and forth mid-sentence in this discourse, as compared to between sentences. Of course, this count is a subset of (4). Similarly, we can count CS fragments of different lengths: a 1-length fragment, for example, indicates lexical mixing.

In detail:

1. **Relative popularity:**

We have run 22,307 two-term combinations of words, generated by 271 keywords associated with a conversation around Dalit issues (D\_corpus).

For the feminism corpus (F\_corpus), we had 293 keywords, and generated 23,844 combinations.

Each combination collected tweets from Twitter with a cap of 30,000 items, which we may safely assume suspect was far over the number of relevant tweets required, because there was no significant difference in time or number of tweets collected per search combination upon experimentally reducing this number, down even to 10,000.

The fact that D\_corpus exceeds F\_corpus in massive proportions, therefore, soundly indicates that the former conversation is in fact more discussed or more agitated than the latter. The recent visibility of caste is theorized by Vivek Dhareshwar (); by ‘recent’ here, we refer to a post-Mandal era, when the caste becomes a part of the nation’s consciousness, and not simply a ‘backward conversation’ to have. (Elaborate)

1. **WordSearch analysis:**

We have created a list of Hindi-English equivalent pairs. These have been segregated by a rough semantic or functional categorization: emotion, ideological abstractions and technical terms.

This list is obviously amenable to expansion and refinement, of course. At the moment, we have included only the broadest of terms, and its size is as following:

(i) Emotion: 8  
(ii) Ideological abstractions: 14  
(iii)Technical terms: 12

These terms are not specific to either corpus or discussion (except ‘feminism’ and ‘untouchability’, belong to the last list). In any case, we are looking for a ratio of Hindi: English from each corpus, rather than the absolute count of occurrences. This tells us about the language that people prefer to use within these arenas, and which language the ideological, technical, and emotional thrust of the conversation lies, no matter the ‘base’ language or the surrounding words.

Search list:

**(i)Emotion:**

1. Anger, rage, राग, गुस्सा, raag, gussa
2. Happy, Happiness, खुश, खुशी, khushi, khush
3. Humiliation, शर्म, नीचा, sharm, neecha
4. Sad, sadness, दुखी, दुख, dukhi, dukh
5. Love, affection, romance, प्यार, प्रेम, मुहब्बत, इश्क, pyaar, prem, muhabbat, ishk
6. Empathy, sympathy, solidarity, दया, हमदर्दी, daya, hamdardi
7. Pride, गर्व, मर्यादा , garv, maryada, maryaada
8. Endurance, सहनशीलता, sahanshilta

**(ii) Ideological abstractions:**

1. Humanity, इंसानियत, insaaniyat
2. Equality, बराबरी, समानता, baraabari, samaanta
3. Inequality, असामता, भिन्नता, asamaanta, bhinnta
4. Freedom, liberty, स्वतंत्रता, swatantrata
5. Upliftment, alleviation, उत्थान, utthaan, uthan
6. Justice, न्याय, nyay
7. Injustice, अन्याय, बेइंसाफी
8. Social, society, समाजिक, समाज, samajik, samaj, samaaj
9. Oppression, repression, दबाव, dabav
10. Abuse, atrocity, atrocities, अत्यचार, हमला,
11. Discrimination, repression, भेदभाव, पक्षपात, bhedbhav, bhedbhaav, pakshpat, pakshpaat
12. Empowerment, सशक्तिकरण, sashaktikaran
13. Privilege, विशेषाधिकार, visheshaadhikar, visheshaadhikaar,
14. Violence, violent, हिंसा, हिंसक
15. Strength, शक्ति, ताकत
16. Progress, advancement, विकास, vikas

**(iii) Technical Terms:**

1. Democracy, लोकशाही
2. Election, चुनाव
3. Government, सरकार
4. Activism, सक्रियता, sakriyata
5. Representative, प्रतिनिधि, pratinidhi
6. Revolution, protest, आंदोलन, विरोध, संघर्ष
7. Campaign, मोर्चा, morcha
8. Religion, धर्म, dharma
9. Rights, अधिकार, adhikar, adhikaar
10. Rape, बलात्कार, balaatkar, balatkar
11. Casteism, untouchability, छुआछूत, मनुवाद, जातीवाद, ब्राह्मनवाद, chuachut, manuvad, jaativad, brahmanvad
12. Feminism, नारीवाद, naarivad, narivad
13. Law, constitution, कानून, संविधान

**Word-count analysis**

This give us an idea of the landscape before we move into code-mixing: i.e. how much percentage of Roman English, Roman Hindi, Devanagari Hindi words we’re looking at in each corpus. It also helps us understand the significance of our results in (1).

**Sentence-based analysis**

Now that we have identified, rather crudely, the language of certain words, we will move on the sentence; this is different from analysis a tweet, because tweets can be and often are multiline, where the user switches languages or scripts for, say, emphasis, or convenience, or any other reason.

In this case, it is invalid to try and identify the language of the tweet: the tweet may not have a ‘base’ or primary language at all.

What we are trying to analyze via this study is the nature of the conversation of these discourses; for the purpose of this sub-analysis, we treat each sentence as a unit of this conversation, regardless of the tweet it belongs to. Then we find the language of this sentence, which is a more plausible thing to do, and tally up the totals for Hindi and English, for both corpora. Note that we are not saying that these results indicate that *users tend more to speak in Hindi than in English (or vice versa)*, we are only saying that the *conversation* heuristics are henceforth.

**Code-mixed tweets**

This is a user or tweet-based analysis. We want to say comparatively how many people code-mix on Twitter when participating in each discourse. The motivations and implications of this have been discussed elsewhere.

**Code-mixed fragments**

We can also count the level (or granularity()) at which code-mixing occurs in each corpus. A run-length of 3 or more indicates that the fragment is larger than a tag.

Most tags: ‘Hello’, ‘good morning’, ‘dekho toh’, ‘vaise bhi’, ‘kya batau’, ‘kuch nahi’, ‘bas’, ‘anyway’, ‘okay’, ‘cool’, ‘haan’, ‘haina’, ‘for example’ are not more than two words. We might have a few 3-word tags: ‘by the way’, for example, but we have chosen to keep the threshold run length as 3, *in any case* (voila 3-word tag), because we want to identify 3-word phrases. We simply check for these few cases to avoid error.

We are avoiding naming lexical CS, because it is not within the scope of this project to be able to distinguish between CS and borrowing. We may, however, name 1-word CS fragments as ‘insertions’.

CS of higher granularity indicates greater ease between the two languages: for example, one need not be fluent in a language to be able to insert tags from it – in our situation: a conversation need not be immersed in a language to contain tags.