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.

Some more 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. These keywords were divided in equal proportions between Hindi and English for either corpus – the entire lists are available in WordSearchResults.xlsx. We have tried, in short, to minimize any possible bias on our part in the collection of data by making both sets of keywords equally ‘strong’.

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 by a considerable proportion, 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-denoting**, **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: 16  
(iii)Technical terms: 11

Each of the words in the above list is searched in a cluster, i.e. along with its close synonyms, possible spelling variants, etc. so as to gather as accurate an estimate of how many times the concept appears in the corpus as possible.

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:[[1]](#footnote-1)

**(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, campaign, आंदोलन, विरोध, संघर्ष, मोर्चा, morcha
7. Religion, धर्म, dharma
8. Rights, अधिकार, adhikar, adhikaar
9. Casteism, untouchability, छुआछूत, मनुवाद, जातीवाद, ब्राह्मनवाद, chuachut, manuvad, jaativad, brahmanvad
10. Feminism, नारीवाद, naarivad, narivad
11. Law, constitution, कानून, संविधान

**RESULTS**

We observe that the percentage of Hindi used in Dalit\_corpus to represent these word clusters outdoes that of Feminism\_corpus in the majority of words in each category.

|  |  |  |  |
| --- | --- | --- | --- |
| **Categories** | **Number of clusters in which D\_corpus Hindi ratio is greater/Total number of clusters** | **Outdoes by a landslide (at least 20% margin) / Total** | **More or less equal/ Total**  **(within 10% margin in either way)** |
| Emotion | 7/8 | 6/8 | **-** |
| Ideological abstractions | 9/16 | 5/16 | 3/16 |
| Technical terms | 10/11 | 5/11 | 4/11 |

A more detailed representation of the results, with the figures for each word cluster and its constituents and percentages is available in WordSearchResults.xlsx.

This observation aligns with our hypothesis: indeed, it is commoner for people in the Dalit discourse to express key concepts in Hindi than for people in the Feminism corpus, no matter what the surrounding language of the tweet is.

**A closer look at the figures, and some sub-results**

The absolute values of the number of occurrences of each constituent in the above clusters are also telling, as well as the individuals percentages.

Absolute distribution of negative emotions in F\_corpus

The Hindi/Total ratio for **Emotion** keywords were **consistently higher than 50% for D\_corpus**, whereas they varied for F\_corpus. Interestingly, F\_corpus hits 98.8% for (3), i.e. humiliation-शर्म. While happiness, sadness, love and pride appear about 80% of the times in English, humiliation and anger are almost always expressed in Hindi in F\_corpus.

The case of ‘pride’ is a little different: while both the Hindi figures are high (97.2% and 65.5% for D\_corpus and F\_corpus respectively), this is for different reasons. In D\_corpus, the Hindi sense ‘गर्व’ populates the figure, with 1407 hits, whereas in F\_corpus, ‘मर्यादा’ populates it. ‘मर्यादा’ signifies, actually, *male*-pride, from an Indian patriarchal tradition; it not only does not possess an exact English equivalent, possibly promoting Hindi usage, but also contains highly negative connotations in this conversation, fitting in, therefore, with the above Hindi trope of humiliation and anger in F\_corpus.

Ideological Abstractions

Here, we are observing ideological words from an essentially Western discourse of humanity, discrimination, equality, and empowerment. (This is not to say that these concepts do not exist in India or Hindi, only that the global conversation around them solidified with the likes of the UNHRC and therefore a Westernized discourse stylistic.) We see that in F\_corpus, where we had expected the conversation to align more with the global discourse, the Hindi percentages are accordingly **consistently low, under 10% for 7/16 words, and hit 0% for ‘inequality’.** The figure spikes for ‘शक्ति’, probably because of the slogan of नारी शक्ति (literally, women strength) in Hindi. The word ‘privilege’ is a prime example of an intensively Western discourse-point: the discussion of privilege politics is ubiquitous in any ‘educated’ discussion these days. The figures, again, corroborate (only 0.41%, 3.33% Hindi respectively).

Some other caveats are that certain words, like ‘society’ and ‘discrimination’, show markedly different values for D\_corpus and F\_corpus (83.8%, 24.6% and 56.6%, 5.77%); this is to say that this ideological conversation of community and rights is present in Hindi, but that while the Dalit conversation chooses to occupy the space of ‘भेदभाव’ and ‘समाज’, the feminism conversation is still inspired mainly by the global human rights registers.

There are some exceptions to the trend, of course: surprisingly, cluster (16): progress-विकास showed near-complete Hindi percentages: 95.3%, 82.9%, respectively – it is difficult to pinpoint a reason.

And finally, Technical Terms

The Hindi percentages here are high-ish, both for D\_corpus and F\_corpus, which is surpising given that the language of legality and jargon is shifting to English. However, there is an observable trend: we note that more functional words such as ‘government-सरकार’, ‘representative-प्रतिनिधि’ and ‘election-चुनाव’ are dominated significantly by Hindi in both corpora, while the figure drops to below 20% for ‘democracy-लोकशाही’ and ‘activism-सक्रियता’, which are more conceptual and less likely to appear in, say, a discussion about the coming elections.

We achieve a rough 50-50 for both corpora for ‘religion-धर्म’, which is suggestive that religion is still based in the native tongue, even while the surrounding ideological discourse may be tilting towards English.

(Of course, while the above observations have been generalized roughly for both corpora, the Hindi figures are still consistently higher for D\_corpus, as we primarily wished to demonstrate.)

Intersectionality

The terms ‘feminism-नारीवाद’ and ‘casteism-जातीवाद’ were included in the list: a look at these figures gives us a snapshot indication of the state of intersectionality in each discourse. Unfortunately, it is clear that intersectionality is near-absent: the feminism cluster appears 11 times in D\_corpus as compared to 3083 appearances of the casteism cluster, and the casterism cluster appears 24 times in F\_corpus as compared to 470 appearances of the feminism cluster.

The following four sub-analyses are interrelated: I will discuss the results collectively below.

1. **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).

**RESULTS**

|  |  |  |
| --- | --- | --- |
|  | **Total words** | **Hindi percentage (including words in either Roman or Devanagari script)** |
| **Dalit Corpus** | 1798364 | 78.88 |
| **Feminism Corpus** | 543338 | 48.51 |

Although comparing word percentages is not a very significant thing to do, it is indicative of the results that follow. Furthermore, we now have a baseline to juxtapose our results of the WordSearch Analysis i.e. if roughly 80% of words in the corpus are Hindi, then it is all the more important that the word cluster for ‘justice’ is expressed only 24.88% times in Hindi, and so on. Similarly, given that only 48.51% of words are in Hindi in F\_Corpus, we must be very surprised at the 98% figure we get for ‘humiliation-शर्म’. Each figure that we arrive at in our Word Search Analysis, therefore, can be put into perspective.

The ratio of the total words in the corpora (about 3:1 in favour of Dalit\_Corpus) also puts the absolute values of the occurrences of our key terms in Word Search Analysis into perspective. For example, the word clusters for ‘justice’, ‘empowerment’ and ‘rights’ appear *in absolute terms* more in F\_Corpus than in D\_Corpus, despitethe 3:1 ratio in the opposite bias. This ties in, once more, with our hypothesis above, about the globalized Englishized discourse we can observe in F\_Corpus.

1. **Sentence-based analysis**

Now that we have a crude estimate of the language of the individual words, we move on to the sentence; this is different from the analysis of 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.

**RESULTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Total sentences** | **(Romanized) Hindi sentences %** | **Devanagari Hindi sentences %** | **English sentences %** |
| **Dalit Corpus** | 55168 | 7.47 | 56.96 | 35.56 |
| **Feminism Corpus** | 35828 | 6.64 | 15.79 | 77.56 |

**The skew is clearly visible between the Hindi figures of the Dalit and Feminism corpora (64.43% - 22.43% respectively)**. This indicates strongly that, in fact, **the subject matter of a discourse affects the distributions of language within the discourse.** This was the presupposition of our hypothesis, that the Dalit Corpus and Feminism Corpus will diverge in the Hindi-English proportions that they exhibit because of their content and sociolinguistic situations in the current Indian scenario.

In more detail:

The Dalit corpus, in fact, is showing us a **35 – 65 English-Hindi division**, which is **counter to the general Twitter baseline of 70 – 30 English-Hindi**.[[2]](#footnote-2) A brief look at the Dalit Corpus shows a good deal of charged, highly political discussion around national parties, policy and on-ground issues; it is not surprising that this conversation is embedded in the regional language, i.e. Hindi, especially since there is also campaigning and opinion-recruiting that is happening here. For example,

“priyankagandhi BhimArmyChief प्रियंका जी में केवल इतना जानना चाहूंगा कि आपकी सरकार ने आजतक हम दलितों के उद्धार के लिए क्या कदम उठाये है ##जात की राजनीति छोड़ अपने दम पर लड़ो”

(Priyanka Ji, all I want to know is what is your government doing on the behalf of us Dalits #forget caste politics fight on your own strength)

This entire genre of tweets, thanks to the growing Hindi trend/imposition on political propaganda, is likely to occur in Hindi.

The feminism corpus, on the other hand, with its **78-22 English Hindi proportion**, **slightly overshoots the Twitter baseline**. Again, this aligns with our expectations: the conversation around feminism is still perceived to inhabit an elite upper-class arena, occupied majorly by the English-speaking educated class who will tend to choose English as the language of ideological discourse, perhaps slightly more than they would do given informal settings (as would be given a ‘normal’ discussion on Twitter). As we have also already discussed in (1), the key terms of discussion appear in English in F\_Corpus far more than they do in D\_Corpus – these results corroborate each other.

1. **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.

1. **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), we have counted and displayed the number of sentences that exhibit CS fragments of all lengths from 1- 20 in the more extensive results file.

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.[[3]](#footnote-3) 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.

**RESULTS**

|  |  |  |
| --- | --- | --- |
|  | Code-switched sentences %[[4]](#footnote-4) | Phrasal CS % |
| **Dalit Corpus** | 15.95 | 45.25 |
| **Feminism Corpus** | 15.14 | 38.48 |

Here, we find that the figures match, indicating that people code-switch with roughly equal frequency in both corpora. Phrasal CS, as we have specified it, occurs more in the Dalit\_corpus, indicating a slightly higher fluidity between Hindi and English in this conversation. This, again, is in line with all our previous results.

**Thus, the above 5 analyses demonstrate that relative Hindi presence as measured at various levels is consistently higher in the discourse on Dalit issues than it is in the discourse on women’s issues on Twitter in India.**

1. Some transliterations/variations, especially Roman transliterations of Hindi words, have not been included here; they are present in the comprehensive search list, available in WordSearch.xlsx [↑](#footnote-ref-1)
2. Although there the unit is the tweet. Still, we can safely say that although our figures may not exactly compare, there is certainly marked divergence from the general English-Hindi baseline on Twitter. [↑](#footnote-ref-2)
3. This is also to reduce error due to the LID, which is occasionally eccentric. [↑](#footnote-ref-3)
4. For similar reasons as the above, we have counted CS sentences as those with at least a 2-word fragment. This is also to reduce error due to Named Entities: for example, ‘Priyanka’ is tagged as \HI, but if it appears in an all-English sentence, we do not want to mark this sentence as code-switched. [↑](#footnote-ref-4)