Final

Observations

Details

We have used the following key to annotate the features

Annotation

Aa Key	Full Form	Values annotated
CS	code switch	languages change from sentence to sentence
AH	all Hindi	Only Hindi sentence, exception: word swaps
AE	all English	Only English sentence (no RWS observed here)
HKIN	Hindi kinship	yaar/bhai/bhaiya/mummy/papa/babu/bhaijaan/bhaiyya/dada
EKIN	English kinship	mom/dad/darling
WS	word swap	Sentence is in Hindi, but some word(s) are swapped for English words
RWS	reversed word swap	Expected English word, got Hindi word
NP	noun phrase + compound words	In Hindi sentence, instead of word being swapped, we see noun-phrases or compound words being swapped
RC	relative clause	jiska/jitne/that clause marking Hindi RC in English sentence (this is the only format observed)
GREET	greeting	Greeting in English, but rest of the sentence is Hindi
QUOTE	quotation	The sentence is in English, but it quotes something in Hindi. (Only cases of this configuration observed, not the reverse)
RESP	respect	sir/maam/maadam
IMP	important	Shows something extra-ordinary happening
CM	case markers	Usage of Hindi Case markers
JJ	Adjective	Usage of English Adjectives in Hindi sentence
R	Romance	Usage of English in romantic words/romantic clauses

Key	Full Form	Values annotated
HI	Higher Institution	Following trends set by a higher institution

We have manually tagged 130+ sentences with the above features, and deeply analysed another 100 sentences.

Why Manual Annotation?

Simply because there did not exist any reliable tool which could do the task we wanted to do. Some of the common problems faced were due to

- Tools not able to recognise latinised language: Example: 'hey ram!', here though 'hey' is a valid English word, and we could probably make a case for greeting a person called ram, but Hindi speakers would know that this is an exclamation.
- Identifying Junctures: Since they are not able to aptly identify language, it becomes even more difficult for any tool to recognise the juncture or point where the language switch happens.
- Marking Features: We did not know *what* to look for, hence it would have been impossible to tell any tool to look for anything. We had to learn by going through all the sentences, and this allowed us to observe things we wouldn't have been able to otherwise.

Introduction to our Work

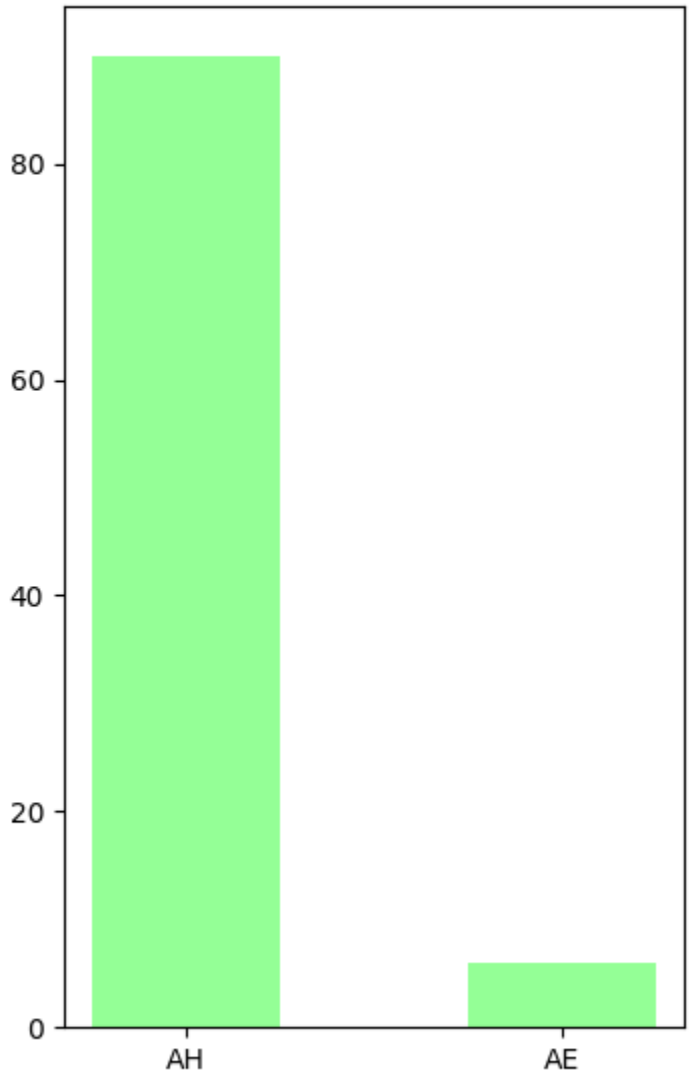
We have taken ~600 sentences of Hindi-English code mixed twitter data. We took care of not to clean it, since cleaning might remove or edit the context which allows us to make a decision whether the word/clause is in Hindi or English. Each sentence was divided into 3 base categories

- All English (AE): Tweets which were written in pure English
- All Hindi (AH): Tweets which were written in pure Hindi
- Code Switch (CS): Tweets which had words of both Hindi and English

After this, we went through and deeply analysed 100 CS sentences to identify some key features which gave us insights towards the rules used in code-mixed languages. Once we got these rules, we marked all the sentences with those features wherever we could observe them. This gave us both a qualitative and quantitative bases about code-mixed languages and the features which exist in it.

Matrix and Embedings

Matrix is the primary language of the sentence (tweet) on which the secondary language is embedded. In our dataset, we observed that Hindi was the matrix in most sentences, with English being embedded in it. Another point that leads to this conclusion is the sheer number of AH sentences as compared to AE tweets.



Features

Word Swap

Word Swap refers to the sentences mostly in a single language but has 2-3 words in the embedded language. This was one of the most observed features, and we have found a lot of insights using this.

- In most cases we observed that one or two English words were being used in Hindi Language
- In some cases, a specific pattern was seen with Word Swap, where adjectives were often used in English in Hindi sentences
- A few word swaps were used as a case of using borrowed words from English in Hindi sentences.

Noun Phrase: We also observed some number of instances where the compound word, or a noun phrase was swapped. This is when we have the matrix language as Hindi.

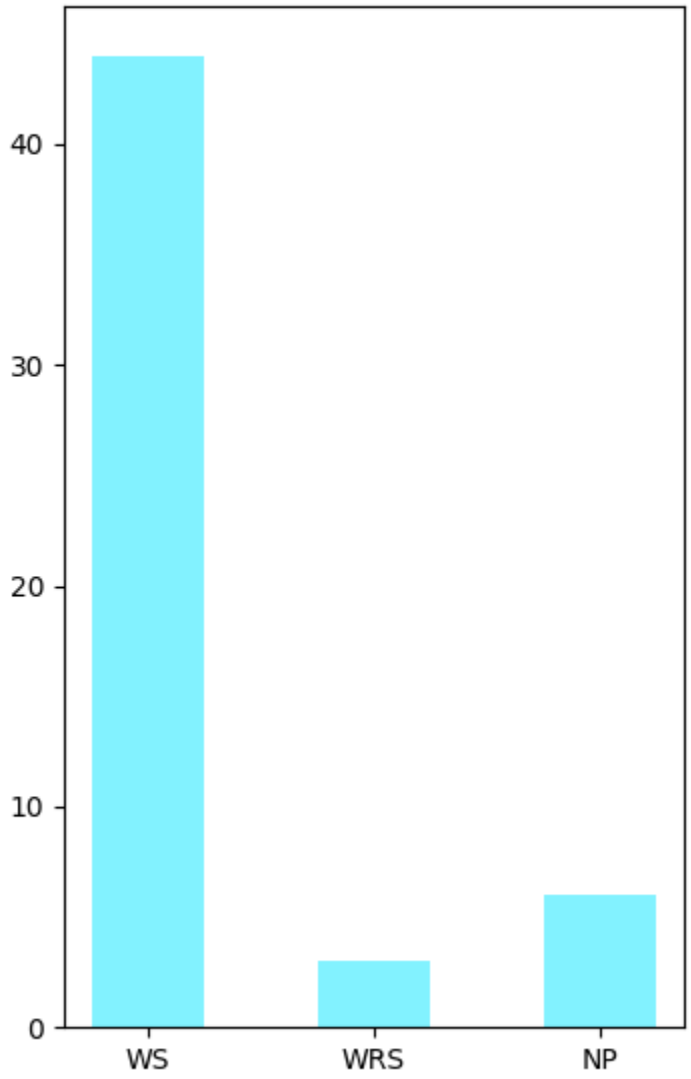
Examples:

- Sab anti national hai saale.
- @lack_a_daisy @Being_Humor bhai minus one point tujhe
- Chalo ab koi sting operation karke kisi sarkar ko nanga nahi kar sakti ... maare jaoge #RightToPrivacy
- @kamaalrkhan sir ye tweet unko bhej do. Itni Hi-Fi English padh ke wo waise hi suicide kar lega.

Reverse Word Swap: An instance of Word Swap, where we expect a word to be used in a particular language, but the speaker uses it in a different language. We observe that for Hindi as matrix language, we would see someone using more common English words, but the opposite happens.

Examples:

- Tumhare chashme ka number hai kya?
- Chalo ab koi sting operation karke kisi sarkar ko nanga nahi kar sakti ... maare jaoge #RightToPrivacy
- Desh ka yuva jaag chuka hai, yuvika ka kya?



Code Switching

There were numerous instances of code switching that took place, which had complete all Hindi and all English sentences, bundled together.

- @MeetUunngLee bhai tu mil raha hai weekend. Thats it
- @himanshujainon @DrGarekar @ashvasant are bhai uski biwi ne mana kia tha . He is innocent otherwise like # kanhaiya , poor student of #JNU

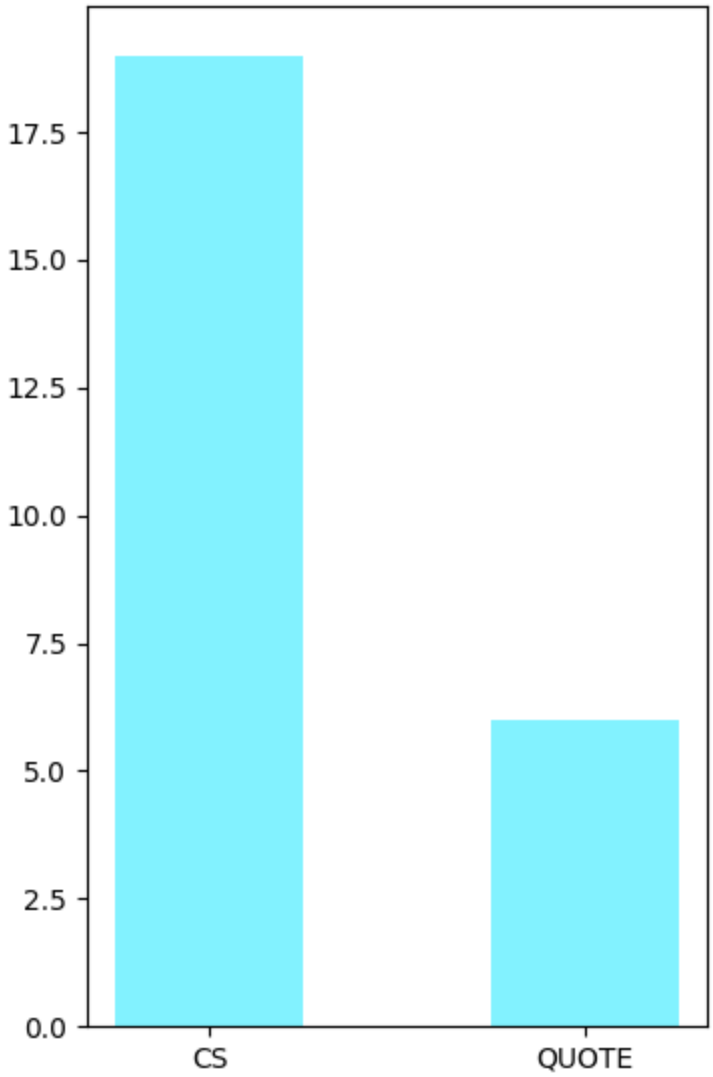
- #GST is The new Duckworth Lewis method. Jo Aaj Tak Kisi ko samjh nhi aaya
- Grow up Bharkha ... It's too much now . Bharat mai kahi shaktiman hai joo aaj bhi problem mai hai ... Whaha Jaa kar aao
- Jeet ka jashn aur shubah ki shuruat by eating bread pakoda at Tandon's Cottage Vaishali Damoh https://t.co/ix2dB9b1lv

Quotes

In cases, we observed that the sentence would be in English or Hindi, as the matrix language, and there would be a Hindi sentence in quotes. That hindi sentence itself can show other phenomena like word swapping, which is explained above.

- #RailBudget2015 #PrabhuKiRail "Prabhu der aaye par durust aaye". This budget is a revolution in Modern Indian Railways.
- " munh khol ke 20 rupay mang liye ..."oh i love this line .
- Thank God they didn't end up with "Dekhna na bhooliye Ae Dil Hai Mushkil apni nasdeeki cinema gharon mein". #MadeByGoogle
- From "azaadi hi meri Dulhan hai" to "Mera toh kaam hi mera valentine hai"
- @shubhansh1504 jab koi bahana nahi hota to bolte hai 'server down hai'

Also note in the above, we have "server down" as a NP in a Hindi quotation in a Hindi sentence. These rules can be thus combined.

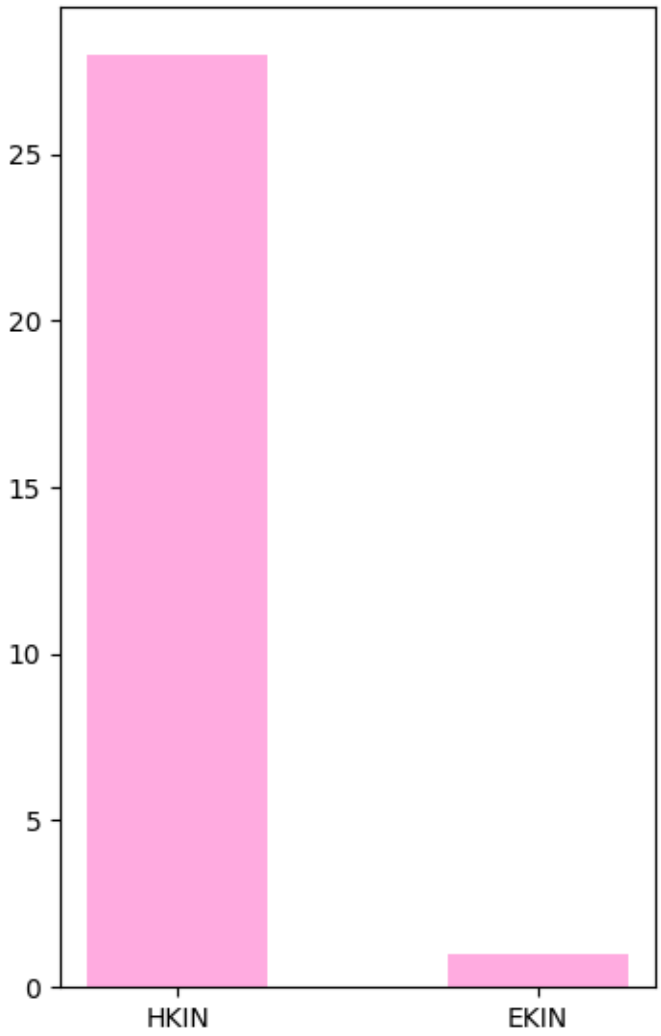


Kinship Terms

Kinship terms refer to the words with which a speaker refers to their kin or family. It was observed in codemixed sentences that speakers used kinship terms in the matrix language (which was Hindi in most cases)

Examples:

- Happy Birthday Dhoni & Hello to Sakshi bhabhi. Aur baaki sab khairiyat?: Here bhabhi is used to refer to Sakshi, which is a Hindi kinship term. The
- Bhai kuch kal ke liye bhi chhod de.. I hope that India wins the World Cup: Here bhai is used which is a Hindi Kinship term
- Everyone on TL is thanking their Moms.. Bhai aiyaashi karne ke paise to Papa hi dete hai na.. #ThankYouPapa: Here both English and Hindi kinship terms are used in respective clauses. This may seem contasting to the first sentence, but it is not so since it is observed that Greetings are often said in English (which is the case in the first sentence)
- RG : Mom mujhe budget budget khelna hai Mom : Beta , tumhare women empowerment khelneke din hai RG : yaay ! ringa ringa roses #RahulOnLeave : This is an interesting example since both Hindi and English kinship terms are being used. Here, since we know who the sentence is about, we can draw more inferences. Here RG uses English kinship to signify that he is more comfortable using that words, which can be seen with 'ringa ringa roses' as well. Contrastingly, the mother uses 'beta', hindi kinship term, since the sentence she speaks is in Hindi.



Romance

When trying to express romantic feelings, oftentimes people end up embedding English words in Hindi sentences, or use purely english sentences

Examples:

- ultimate yarrr . . . . just love it

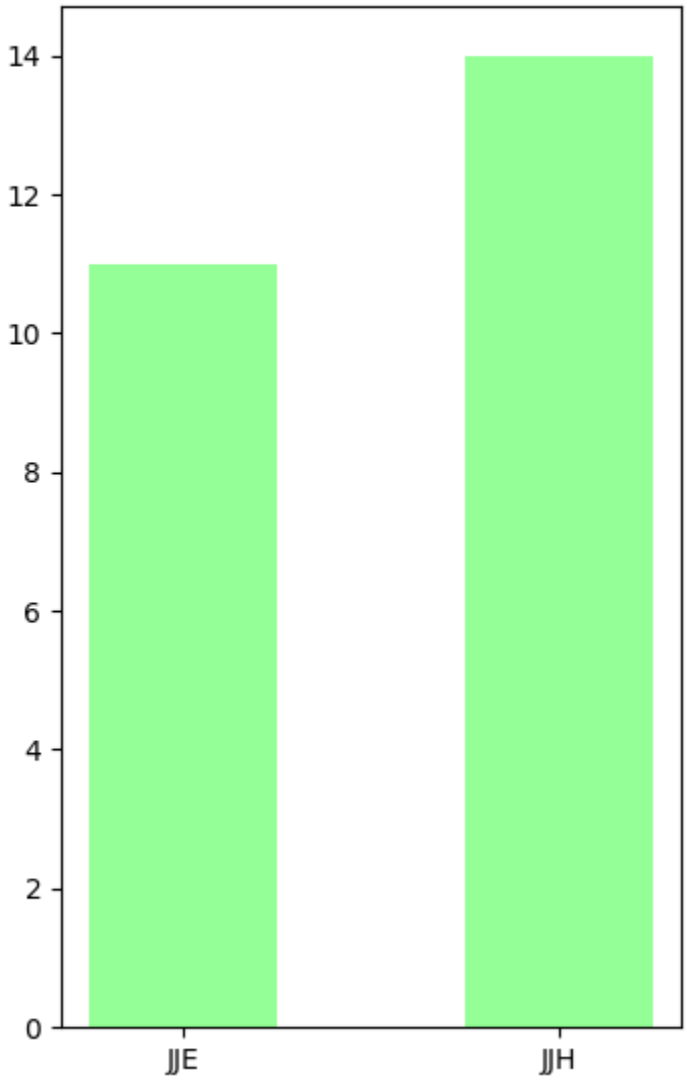
- Bahut log hai jo aapse baat karna chahte hai..plzzzzzz dn't brk thr hearts sir plzzzz
- @huh_watnow darling tumhara kuch ni ho sakta

Adjectives

Oftentimes in Hindi adjectives of English are embedded to show stronger appreciation or disgust. Sometimes they appear as clauses as well.

Examples:

- ultimate yarrr . . . . just love it
- jabatdast dost . . . . 1 page me rula diya yaar . . . . . kamaal hai . . . . superb nd hatss off to ur crearivity . . .
- @Dipti1104 bahut bold hain aap



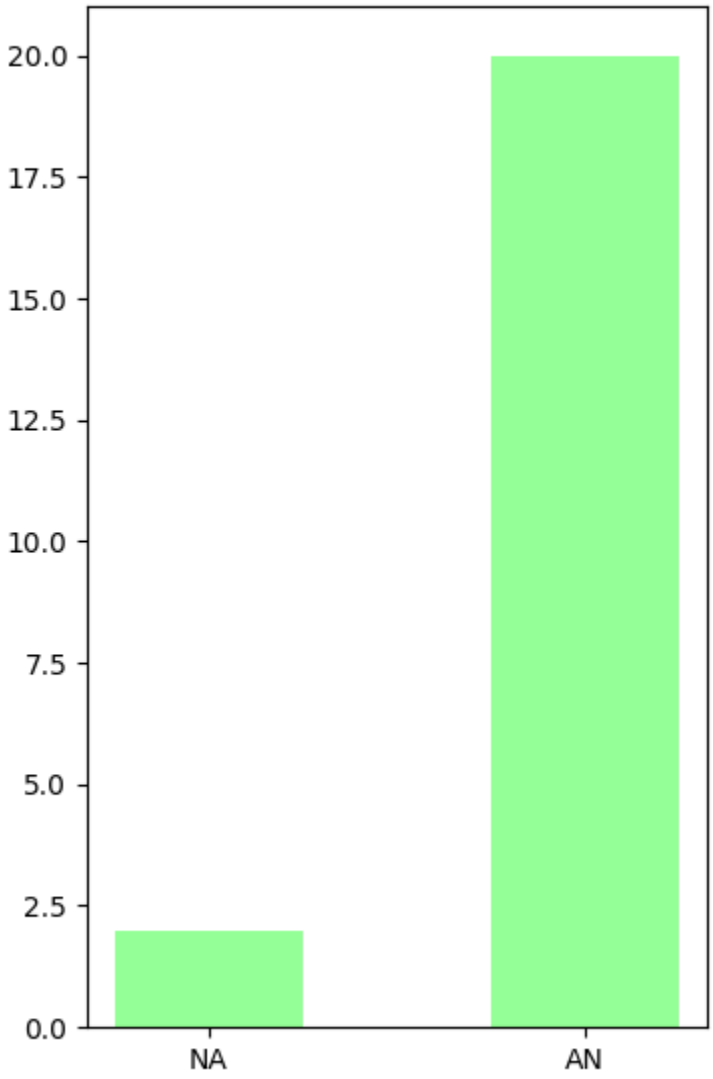
Noun-Adjective order

We observed that the data was overwhelmingly Adjective-Noun order, and the instances of Noun-Adjective are rare, for example in our manual annotation of 130+ sentences, we observed only two such instances:

- Tu jo dariya mein utre saara paani gulaabi
- @ashutosh83B to phir rishta pakka samjhen?

Most of the cases had Adjective Noun ordering

- it is one of the best gigs of grbg bin i hv evr encontrd .... smply owsm
- har cheez ka galat matlab nikal kar ye, galat shalat smmileys bhejti hai.
- Mujhe @Olacabs ne nahi bataya! Ye sarasar na insaafi hai chikna bhai



Relative Clause

In some sentences with a main clause and relative clauses we observe that the transition is marked by Hindi words like ‘jinke’, ‘jisey’ or English words like ‘that’.

Examples:

- RT @ruchikokcha : @Atheist_Krishna Isiliye (sadly) I am perhaps the only woman jisey Kohli ko khelta dekh bhai wali feeling aati hai .
- #saynotocorruption #timetochange say no to dummies *jinke* pass khud ka dimaag nahi dusro k kahne par chalte hai #MakeInIndia
- @razonater @QararaRasha Sindh is not a property of Pakistan . The philosophy is that aap apna Sindh kahin bhi abaad kar sakte hen .

Higher Institutions

People often end up using the words in the same way that a higher institution has used it since it has become a norm.

Examples:

- Please follow @SaleemChikna & @sayshardik, ye log free mein daaru pilaate hai. #AaiShappath
- aur free wifi bhi wahi hai. Lol
- @narendramodi @arunjaitley "Sir #GST rate cut ka fayda Consumer ko nahi mil raha "hai trader base price up karke usi price par bechta hai

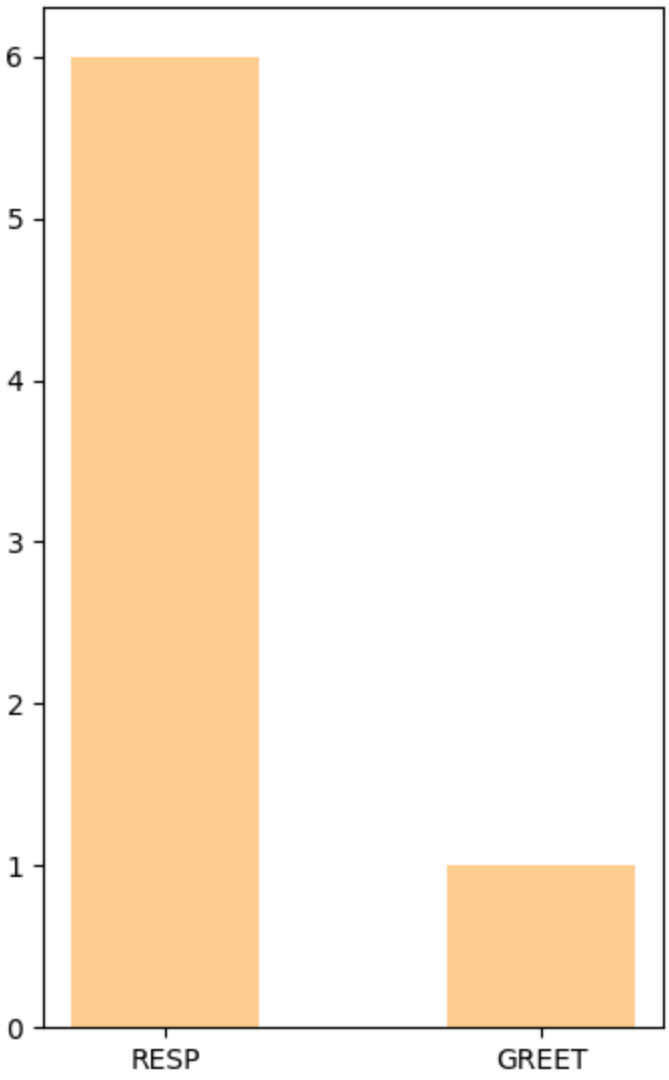
Respect and Greetings

We also observed some amount of respect and greeting words being used. Most of them were in English, and would be sir/maam or cases where only the first word, the greeting would be in English and the rest of the sentence would be all Hindi.

Examples:

- okay sir. hamare bhi din aayenge
- @kamaalrkhan sir ye tweet unko bhej do. Itni Hi-Fi English padh ke wo waise hi suicide kar lega.

Note how the second tweet here has only the first word English, which is respect word, and then



Case Markers

English does not have explicit case markers, hence it is observed that in sentences which start in English, Hindi case markers are used which often then continues as a Hindi sentence

Examples:

- note ye application google playstore se download hogi , toh fake toh ho hi nahi sakti
- Phone ka wallpaper dekhte dekhte zindagi kat rahi hai.

Rules

Find the junctures of where the language switches can be difficult, so we instead re-frame the problem. What are the rules for generating code-mixed data? Once we have solved this problem we could move on to the original problem.

1. For a tweet, select some sentences to be pure Hindi and some to be pure English.
2. Hindi: Now, we swap out low frequency noun-phrases - could be 2 to 4 words, and we also swap low frequency Hindi words. This covers most of the cases observed. We use the following features:
    - Kinship terms
    - Adjectives
    - Higher Institutions
    - Respect
3. English: If the sentence includes some form of quotation, swap the quotation with pure Hindi. Then, use the rules for Hindi, taking the quotation as a sentence in itself. This covers:
    - Quotes
    - Relative clause
4. Append any of "pls/plz/please" at the end, where the sentence is a request.
5. If at the end we have a Hindi adjective and noun, in that order, then with a very low probability flip the order. Hindi will allow this (at least in cases that we

observed).