Language and Society
✨ Final Project Report ✨

- Advaith Malladi 2021114005
- Ayan Datta 2021114017

Introduction

Language is the primary means of communication. Communication is one of the primary uses of language. Therefore, any study of language in a society is essentially a study of how people communicate with each other. Before the digital revolution, mass communication was only possible by people talking to each other in real-time in the real world. After the rise of social networking platforms, the primary means of communication have shifted to these new platforms. Social networking platforms host the conversations of 4.4 billion people across the globe daily. Social media has also made one-to-many communications possible.

In lieu of all this, we felt that today, real human communications and conversations are held on these social networking platforms, contrary to what might have been true a couple of years ago. So, when all the people in the world are there and talking on these social media platforms, we feel that these online platforms would be the best place to analyze and study human language in general.

Research shows that the orthography of Computer-Mediated Communication (CMCs) has shown to exhibit some special pragmatic functions (Pragmatics of the Keyboard: An Analysis of Orthographic Conventions on Tumblr by Molly Ruhl). Hence we would like to emphasize the importance of studying these features.

As the internet becomes more widespread, we see it being used by many multilingual speakers, which leads to phenomena like code-switching/mixing. We would also like to study this phenomenon with special respect to its presence in CMCs.

Hypothesis

Given that most human communication takes place on social media platforms, we will try to identify distinct types of linguistic variables in online speech and try to correlate them with prevailing social variables.

These linguistic variables will be targets of special focus:

- Usage of code-switching/mixing by a user
- Usage of punctuation marks by a user
- Usage of emojis, emoticons by a user
- Usage of intense words

These online social variables will be targets of special focus:

- Like counts
- Retweet counts
- Follower/Friend counts

Along with these, we will also look at the other social variables wherever applicable.

We hypothesize that the mentioned social variables are correlated with the linguistic variables.

Method

The methodology for conducting the study is divided into two parts. One for the code-mixed/switched tweets and the other for the other linguistic variables. This was done due to the lack of automatic code-switching/mixing detection tools and the lack of tools that can calculate the degree of code-switching and mixing as compared to other linguistic variables, which are very easy to process, compute, and work with.

Method (Code-Switching/Mixing)

Data Collection

We have written code that can extract the latest 4000 tweets of any Twitter user and check for code-switching/mixing in those tweets. These tweets were differentiated based on the presence or absence of code-switching/mixing. Along with the tweet content, the retweet count and like count of each of these tweets were also extracted.

Accounts used for code switch/mix tweets:

- Gautam Adani: businessperson
- Amitabh Bachchan: actor
- Eenadu: Telugu Local news agency
- India news: National news agency
- KCR: leader/politician
- KTR: leader/politician
- Narendra Modi: leader/politician
- NDTV: National news agency
- S. S. Rajamouli: Filmmaker
- Sakshi News: Local Telugu News Agency
- TS Police: State-Level Police Agency

Tabulated Data

Gautam Adani: businessperson

English Tweets:

count	547
Retweet average	237.11334552102377
Like average	2217.5063985374773

Data: adani_english.csv

Code switch/mix Tweets:

count	10
Retweet average	874.1
Like average	6054.1

Data: adani_other.csv

Amitabh Bachchan: Actor

English Tweets:

count	2948
Retweet average	368.2174355495251
Like average	5973.6974219810045

Data: bachan_english.csv

Code switch/mix Tweets:

count	1053
Retweet average	929.5792972459639
Like average	16288.270655270655

Data: bachan_other.csv

Eenadu: Local news agency

English Tweets:

count	9
Retweet average	1.777777777777
Like average	8.666

Data: eenadu_english.csv

Code switch/mix Tweets:

count	3992
Retweet average	1.225200400
Like average	5.730210420

Data: eenadu_other.csv

KCR: Politician/Leader

English Tweets:

count	838
Retweet average	81.04773269689
Like average	548.9057279236276

Data: kcr_english.csv

Code switch/mix Tweets:

count	3163
Retweet average	43.788175782484984
Like average	345.8422383812836

Data: kcr_other.csv

KTR: Politician/Leader

English Tweets:

count	3856
Retweet average	207.19709543568464
Like average	1670.1130705394191

Data: ktr_english.csv

Code switch/mix Tweets:

count	145
Retweet average	345.3862068965517
Like average	2716.944827586207

Data: ktr_other.csv

Narendra Modi: Politician/Leader

English Tweets:

count	2762
Retweet average	4606.297610427227
Like average	24479.194424330195

Data: modi_english.csv

Code switch/mix Tweets:

count	1239
Retweet average	3963.0492332526233

Like average	18835.288942695723

Data: modi_other.csv

NDTV: National News Agency

English Tweets:

count	3998
Retweet average	18.384192096048025
Like average	123.86443221610806

Data: ndtv_english.csv

Code switch/mix Tweets:

count	3
Retweet average	1.6666666666666667
Like average	18.0

Data: ndtv_other.csv

S. S. Rajamouli: Filmmaker

English Tweets:

count	3967
Retweet average	272.6478447189312
Like average	1716.0569700025208

Data: rajamouli_english.csv

Code switch/mix Tweets:

count	34
Retweet average	726.9117647058823
Like average	4223.264705882353

Data: rajamouli_other.csv

Sakshi News: Local News Agency

English Tweets:

count	16
Retweet average	0.5625
Like average	14.125

Data: sakshinews_english.csv

Code switch/mix Tweets:

count	34
Retweet average	726.9117647058823
Like average	4223.264705882353

Data: sakshinews_other.csv

TS Police: State-Level Police Agency

English Tweets:

count	3162
Retweet average	25.547438330170777
Like average	74.8168880455408

Data: tspolice_english.csv

Code switch/mix Tweets:

count	839
Retweet average	54.07628128724672
Like average	135.8855780691299

Data: tspolice_other.csv

Data Analysis

Gautam Adani:

Out of the 647 tweets that Gautam Adani had on his Twitter account, 547 were completely in English, and 10 were code-switched.

Amitabh Bachchan:

Out of the 4000 tweets that we collected from Amitabh Bachchan’s Twitter account, 2948 tweets are in English, and 1053 are code-switched/mixed. We have noticed that the like and retweet count is significantly higher when he code-switches when compared to English.

Eenadu:

Out of the 4000 tweets we collected from Eenadu’s Twitter account, only 9 are in English; the rest are code-switched.

KCR:

Out of the 4000 tweets we collected from KCR’s Twitter account, 838 are in English, and 3163 tweets are code-switched. We realized that KCR’s tweets had more retweets and likes when he tweeted in English.

KTR:

Out of the 4000 tweets we collected from KTR’s account, 3856 were in English, and the rest were code-switched. We have noticed that the likes and retweets are higher in the case of code-switched tweets that, too, in Telugu.

Narendra Modi:

Out of the 4000 tweets we collected from Narendra Modi’s Twitter account, 2762 were in English, and the rest were code-switched in multiple languages. His tweets in English have more likes and retweets.

NDTV:

We have noticed that all NDTV’s tweets are in English.

S. S. Rajamouli:

Out of the 4000 tweets from Rajamouli’s account, 3967 are in English, and the rest are code-switched, that too in Telugu. His tweets in Telugu have more likes and retweets than those in English

TS Police:

Out of the 4000 tweets we looked at by TS police, 3162 are in English, and the rest are code-switched. The code-switched tweets are well-received as they have more likes and retweets.

Interpretation

Gautam Adani:

We feel that when he switches codes, the likes and retweets are more because when he tweets in his native language, Gujarati, there is a feeling of solidarity amongst Gujarati people, so they tend to like and retweet the tweet more than ever. **So, the code-switching variable has more reception in the native language due to solidarity in the case of Gautam Adani, a businessperson.**

Amitabh Bachchan:

As mentioned, we have noticed that the like and retweet count is significantly higher when he code-switches when compared to English. We feel this is because Amitabh Bachchan is a Bollywood actor, and his fan base on social media is a higher percentage of Hindi people. As Amitabh Bachchan's code-switched tweets consist mostly of Hindi, his Hindi-speaking fans are more likely to like and retweet the code-switched tweets. **So, the code-switching variable is positively received in the language of occupation in the case of Amitabh Bachchan, an actor.**

Eenadu:

We feel that the code-switch from English is because the majority of the readers of this account are Telugu people. **So, the Eenadu Twitter account mostly tweets in Telugu.**

KCR:

We realized that KCR's tweets had more retweets and likes when he tweeted in English. We feel that the more retweets and likes are because since it is the official account of the Chief Minister of Telangana, the tweets are more likely to be read by people across the nation when he tweets in English. **Thus, the code-switching variable is more accepted in the Lingua Franca, English in the case of KCR, a leader with nationwide fame.**

KTR:

As mentioned, we have noticed that the likes and retweets are higher in the case of code-switched tweets that, too, to Telugu. This is because KTR is a local state-level leader. His followers are mostly people who speak Telugu and are from Telangana. **Thus, his tweets were well received when the code switched because most of his followers were local Telugu citizens.**

Narendra Modi:

As mentioned, his tweets in English have more likes and retweets. This is because he is a global-level leader whose tweets are read by people across the world. **So, his tweets in English are more received because people across the world read his tweets.**

NDTV:

NDTV’s tweets are in English because it is a national channel and targets a wider audience.

S. S. Rajamouli:

As mentioned, his tweets in Telugu have more likes and retweets than those in English. This might be because most of his fans are from Telugu regions.

TS Police:

We feel the code-switched tweets are more well received because the code-switched tweets are more reachable to people across the state who might not be proficient in English.

Method (Other Linguistic Variables)

Data Collection

1,00,001 tweets, all from different users, were scraped from Twitter. A lower bound for likes of 50 likes was set while collecting tweets to filter out the bot spam tweets, which held no significance in this study.

Twitter’s language filter for English was also applied while scraping tweets.

The 1,00,001 tweets were stored in a .csv file named tweets.csv

Then the values for the linguistic variables were calculated for every tweet.

Calculation of Linguistic Variables

Total Punctuation Score

The total punctuation score is a linguistic variable that indicates the presence of punctuation marks and also to what degree they all have been used. It counts the following as punctuation:

- .
- ?
- ,
- “
- ‘
- *
- #

The total punctuation score was calculated as follows:

$$\text{Total Punctuation Score} = \frac{\text{Punctuation Count}}{\text{Total Character Count}}$$

Period Scores

The period score is a variable that indicates and measures the degree of usage of the punctuation mark (.). The period score was calculated as follows:

$$\text{Period Score} = \frac{\text{Period Count}}{\text{Total Character Count}}$$

The following were not considered to be periods:

- A contiguous sequence of “.”

  For Example:

  > *No…*

Following is the regex to detect periods:

```
[^.]?([.])[^.]?
```

Ellipsis Scores

The ellipsis score is a variable that indicates and measures the degree of usage of the punctuation mark (…). The ellipsis score was calculated as follows:

$$\text{Ellipsis Scores} = \frac{\text{Ellipsis Count}}{\text{Total Character Count}}$$

Any extension to the number of periods in the ellipsis does not change the score.

The following were not considered to be ellipses:

- A single period:

  For Example:

  > *No.*

Following is the regex to detect ellipses:

```
[^.…]?(\.|…)((\.|…)+)[^.…]?
```

Other Punctuation Scores

Scores for the following punctuation marks were calculated individually:

- Exclamation Score **!**
- Question Mark Score **?**
- Comma Score **,**
- Double Quote Score **“**
- Single Quote Score **‘**
- Asterisks Score *****
- Hashtag Score **#**

The scores were calculated as follows:

$$\text{Punctuation Score} = \frac{\text{Punctuation Count}}{\text{Total Character Count}}$$

Emoji Scores

The emoji score is a linguistic variable that indicates the presence of and also captures the degree of usage of emojis. The emoji score was calculated as follows:

$$\text{Emoji Score} = \frac{\text{Number of Emojis}}{\text{Total Character Count}}$$

Emoticon Scores

The emoticon score is a linguistic variable that indicates the presence of and also captures the degree of usage of emoticons, specifically the smiley face :) and the frowning face :(.

The emoticon score was calculated as follows:

$$\text{Emoticon Score} = \frac{\text{Number of Emoticons}}{\text{Total Character Count}}$$

The extension in the emoticon does not change the emoticon score.

For Example:

- *:)))))* will be counted as one emoticon

Following is the regex to detect emoticons:

```
:(\)+|\(+)
```

Capitalization Ratio

The capitalization ratio is a linguistic variable that indicates the presence of and also captures the degree of capitalization.

The capitalization ratio was calculated as follows:

$$\text{Capitalizaion Ratio} = \frac{\text{Number of Upper Case Letters}}{\text{Total Character Count}}$$

Intense Words Score and Unique Intense Words Score

The intense words score is a linguistic variable that indicates the presence of and also measures the degree of usage of intense words.

Intense words are words that have 2 or more repetitions of characters. "I loooooveeee youuuuuuu", and "I haaattteeee youuuuuu" are both intense and have intense words.

The intense word score can be calculated as follows:

$$\text{Intense Words Score} = \frac{\text{Number of Intense Words}}{\text{Total Number of Words}}$$

The Unique Intense Words score is the same as intense words, but it measures all unique occurrences instead.

Words were separated simply using a whitespace tokenizer.

Used Social Variables

The following social variables were extracted from tweets and were used:

- Like Counts
- Retweet Counts
- Twitter Age in days (Creation Days)
- User’s Follower Count

- User’s Friend Count

Data Analysis

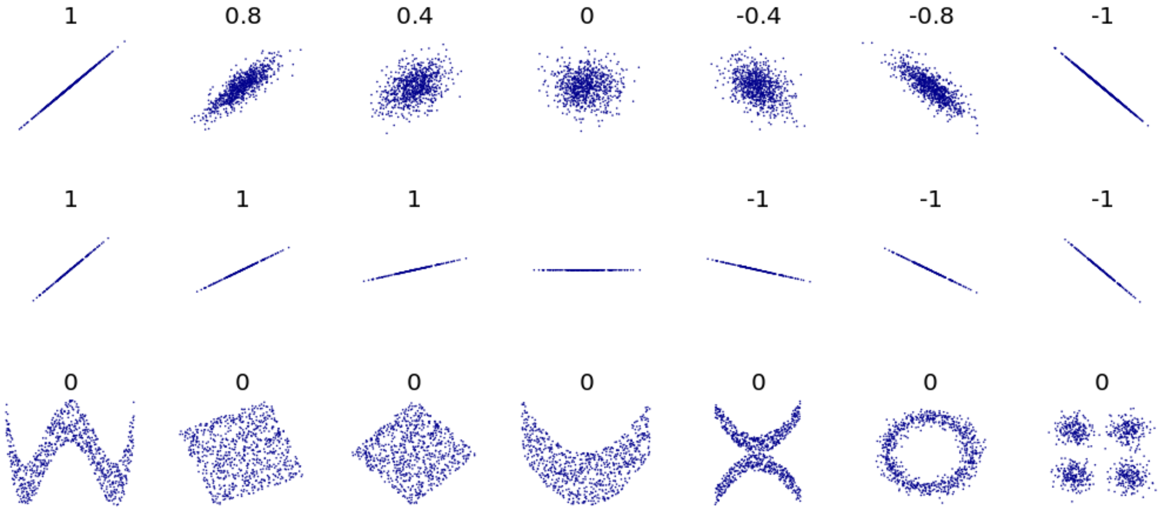
We used both plotting and statistical measures to find any correlation between any of the $75 = (15 * 5)$ social variable - linguistic variable pairs.

Statistical Method of Finding Correlation

The following measures were used to detect any sort of correlation in the data

Pearson’s Correlation Coefficient

The Pearsons Correlation Coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.



Values of the PCC for some data

A value of -1 indicates a strong negative linear correlation, and a value of +1 indicates a strong positive linear correlation. A value of 0 indicates that there exists no linear correlation between the two variables.

Spearman’s Rho

A Spearman correlation coefficient is also referred to as Spearman rank correlation or Spearman’s rho. It is typically denoted either with the Greek letter rho (ρ), or rs. Like all correlation coefficients, Spearman’s rho measures the strength of association between two variables. As such, the Spearman correlation coefficient is similar to the Pearson correlation coefficient. It also has a value always between -1 and 1.

A positive correlation coefficient indicates a positive relationship between the two variables (as values of one variable increase, values of the other variable also increase), while a negative correlation coefficient expresses a negative relationship (as values of one variable increase, values of the other variable decrease). A correlation coefficient of zero indicates that no relationship exists between the variables.

As compared to Pearson’s Correlation Coefficient, Spearman's correlation determines the strength and direction of the monotonic relationship between two variables rather than the strength and direction of the linear relationship between two variables

Kendall’s Tau

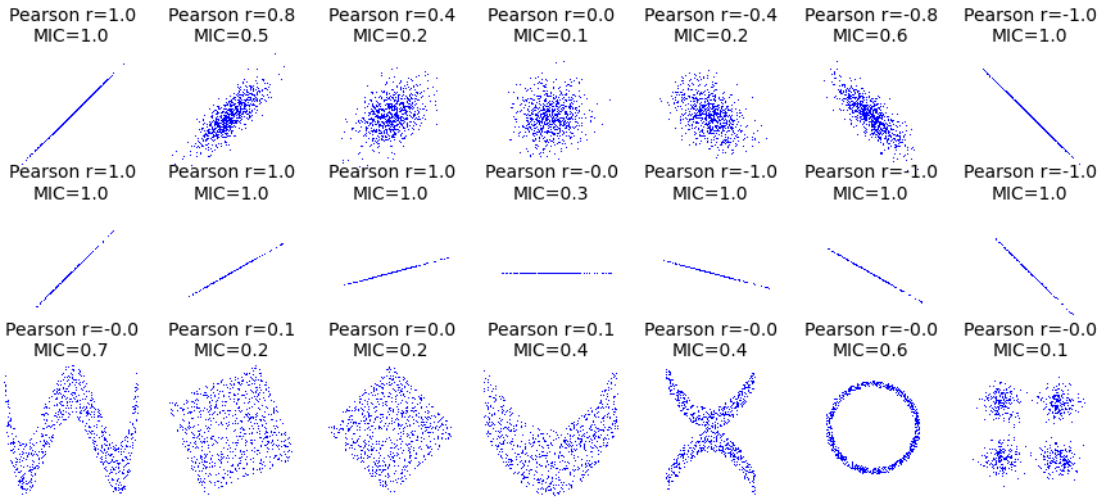
Kendall’s Tau is similar to Spearman’s Rho but usually has smaller values than Spearman’s rho correlation and is more insensitive to error. While the calculation of Spearman's Rho is based on deviations, Kendall’s Tau uses concordant and discordant pairs. Kendalls Tau is also non-parametric

The main advantages of using Kendall’s tau are as follows:

- The distribution of Kendall’s tau has better statistical properties.
- The interpretation of Kendall’s tau in terms of the probabilities of observing the agreeable (concordant) and non-agreeable (discordant) pairs is very direct.
- In most of situations, the interpretations of Kendall’s tau and Spearman’s rank correlation coefficient are very similar and thus invariably lead to the same inferences.

Maximal Information Coefficient

The maximal information coefficient (MIC) is a measure of the strength of the linear or non-linear association between two variables, X and Y. The maximal information coefficient has been described as a 21st-century correlation that has its roots in information theory. The MIC lies in a range [0,1], where 0 represents no relationship between variables and 1 represents a noise-free relationship of any form, not just linear. MIC will not give any indication of the type of relationship, though. It is possible with the MIC to find interesting relationships between variables in a way that simpler measures, such as the correlation coefficient, cannot.



Values of MIC and PCC for some data

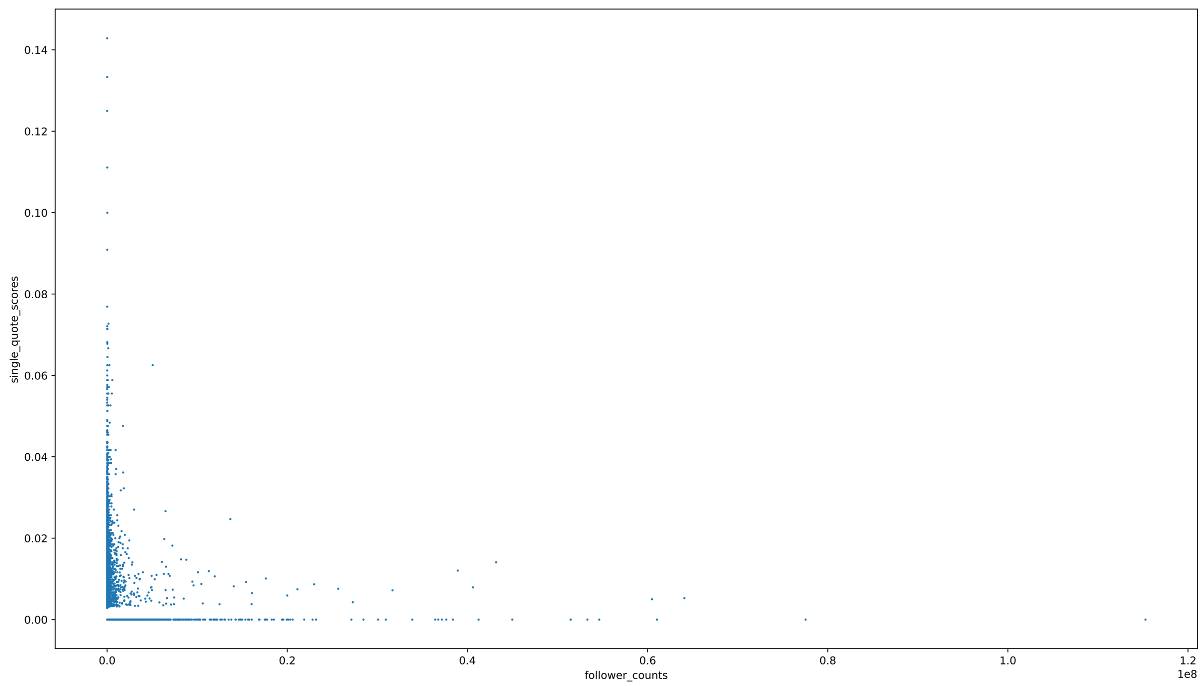
The Additional statistics that were calculated but not used in this study:

- Maximal Information Coefficient
- Maximum Asymmetry Score
- Maximum Edge Value
- Minimum Cell Number
- Minimum Cell Number
- Generalized Maximal Information Coefficient
- Total Information Coefficient

These statistics were calculated for each social variable - linguistic variable pair and stored as correlations/<social variable>-<linguistic variable>.csv

Scatter Plots

Scatter plots for all of the 75 social variable - linguistic variable pairs were made and are stored in the plots_png directory.



Interpretation

No Significant Correlation Found

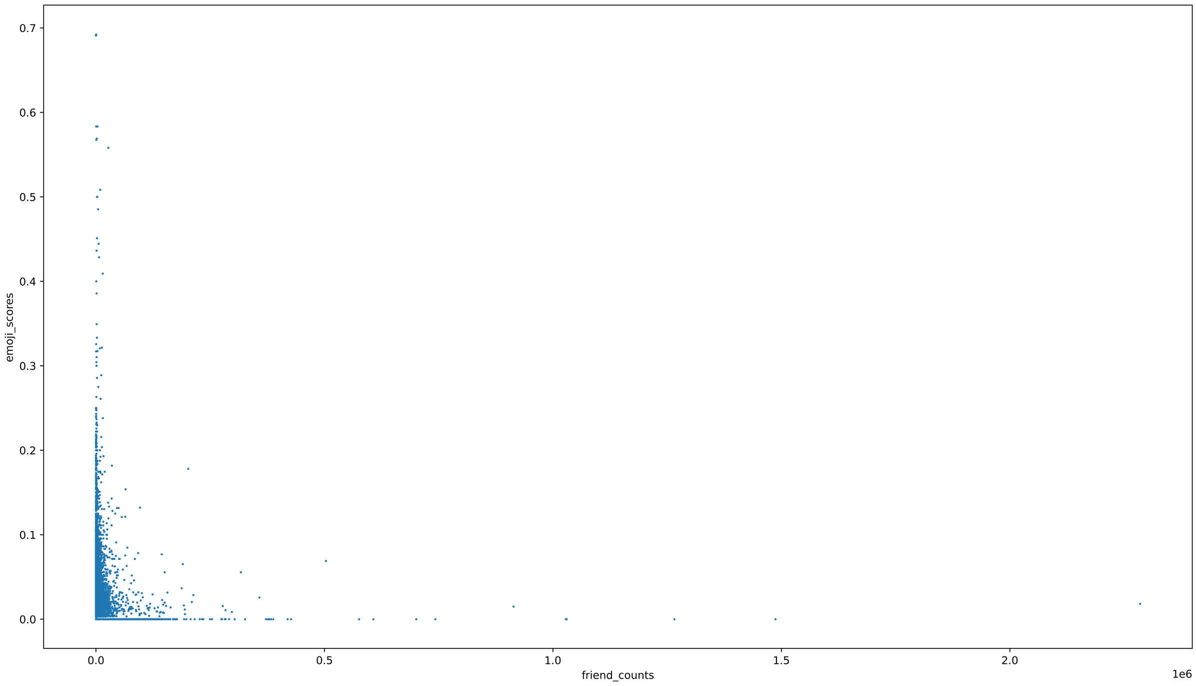
Among all the social variable - linguistic variable pairs, no significant correlation was detected. The MIC, PCC, Spearman's Rho, and Kendall's Tau for all pairs were strictly below 0.1. There were a few exceptions, though:

```
- retweet_counts-emoji_scores   Spearman's Rho 0.1296206365333962
                                Kendall's Tau 0.10077920233500265

- follower_counts-total_punctuation_scores  Spearman's Rho 0.12966867621992878

- retweet_counts-total_punctuation_scores Spearman's Rho 0.1977122594921993
                                          Kendall's Tau 0.13591717170980364

- retweet_counts-period_scores  Spearman's Rho 0.15034143569850592
                                Kendall's Tau 0.10242562687380016

- retweet_counts-capitalization_ratios  Spearman's Rho 0.16128512657589153
                                        Kendall's Tau 0.1105705237394032

- retweet_counts-hashtag_scores  Spearman's Rho 0.28209397616178983
                                 Kendall's Tau 0.22445376514536125

- follower_counts-period_scores  Spearman's Rho 0.1377609135299159
```

- These exceptions are very weakly positively correlated and are very weakly monotonic and non-linear
- For every correlation involving a specific punctuation score, there is a correlation in the total punctuation score and vice versa. Hence we can't make any conclusive inferences from the pairs involving the total punctuation score.

The Initial Spike Of Linguistic Variable Score In The Beginning Of The Plot

One observation for many of the plots is that we don't find many points in the regions of a high linguistic variable and high social variable score that is the top right region from all the points in the plot. Or we can also say that there is a spike in the linguistic variable in the lower social variable score region (the left side of the scatter)



One of the many plots that exhibit this property

Discussion

After looking at the code-switched/mixed data and having interpreted it, we would like to list some trends we noticed.

- If the person is a local celebrity, their code-switched tweets are likely to have more retweets and likes because of the local fan base they have.
- If the person is a global or national celebrity, their tweets in English are more likely to have more likes and retweets because more people speak English across the globe.
- Local news channels are more likely to tweet in code-switched varieties due to better reception.
- National/Global news channels are more likely to tweet in Hindi/English due to the above-mentioned reasons.

After looking at other linguistic variable data and having interpreted it, we would comment on the following:

- The study finds no correlation between the way people use punctuation, emojis, and intense words and popularity, which can be estimated using any of like counts, retweet counts, friend counts, and follower counts. This could mean the following:
  - There is an underlying distribution/ pattern that the correlation measure cannot detect, and it is through more careful study we can study them.
  - There really is no correlation between the degree of usage of these variables to the social variables mentioned. This means no matter how popular/experienced using twitter one gets, they all follow the same norms of using punctuation, emojis, and intense words.
  - There could be a correlation between a specific combination of linguistic variables and social variables.
- There is an irregularity in the lower social variable score region for the linguistic variable; for example, in the friend count - emoji score plot shown before, we can see a spike at the beginning of the graph. This could mean that Twitter users with lesser friends or, more generally, with lesser popularity tend to follow radically different linguistic norms than other users.

Conclusion

To conclude the work done, we have looked for a correlation between linguistic factors like usage of punctuation marks/emojis/emoticons/intense words, code-switching/mixing with various different social factors.

For Code Switching/Mixing, we have observed general trends which correlate factors like popularity to the use of a certain code.

For other linguistic variables, we have tried to look for correlation with social factors like likes, retweets, followers, friends, and Twitter age, but we didn't find any significant correlation, and we have listed down why this could have happened.

Use the below link to download related code, data, and the graphs used/made in this study:

final_report_attachment.zip