Data I/O CHN Challenge

Looking into defining twitter verification of individuals

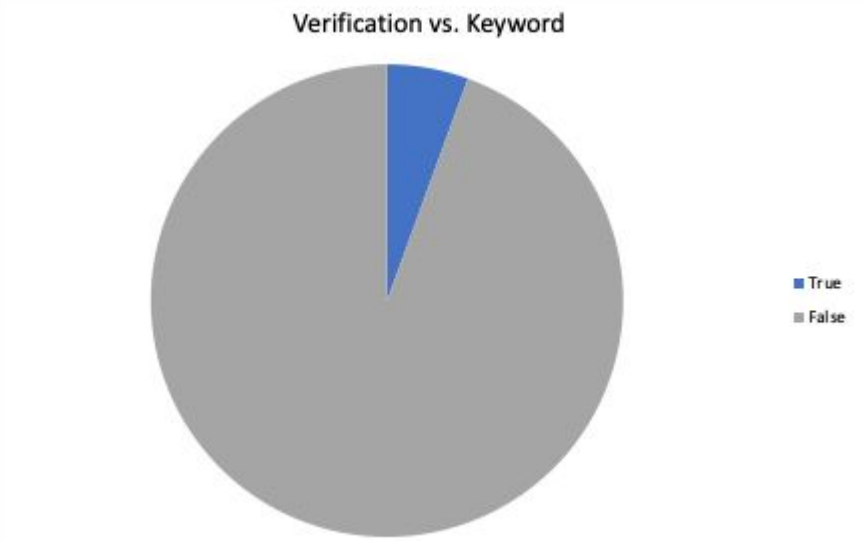
The task

Validating users on their connections and qualifications

The Data

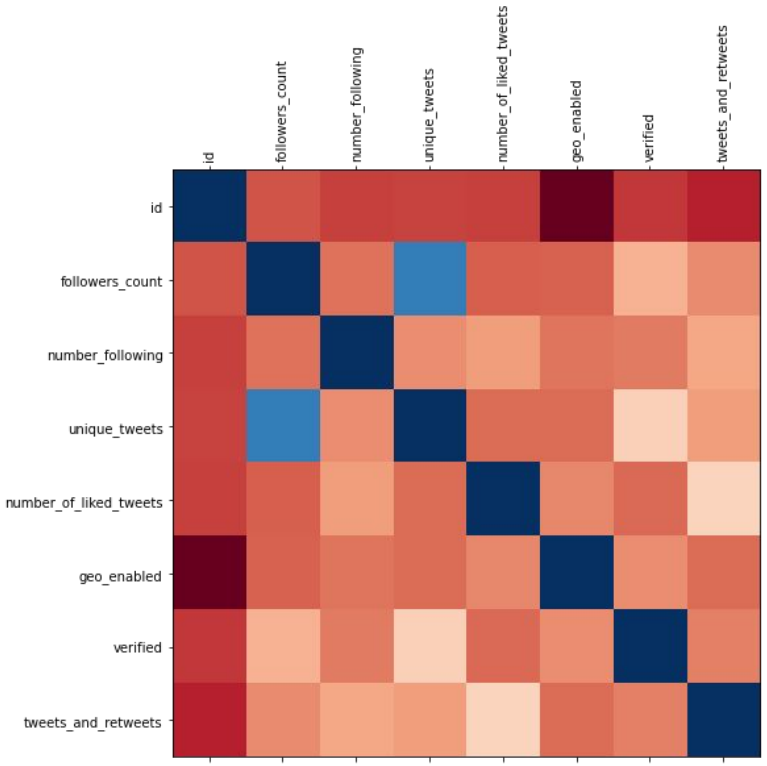
The data consists of twitter users and a subset of tweets that fall under specific categories. Some important data includes likes, retweets, unique tweets, followers, and mentions.

Verification vs. Keyword



- 94 % True; 6% False
- Keywords used: Phd, Ceo, Founder, Scientist
  Expert, Official, Leader, Owner, Executive, Director

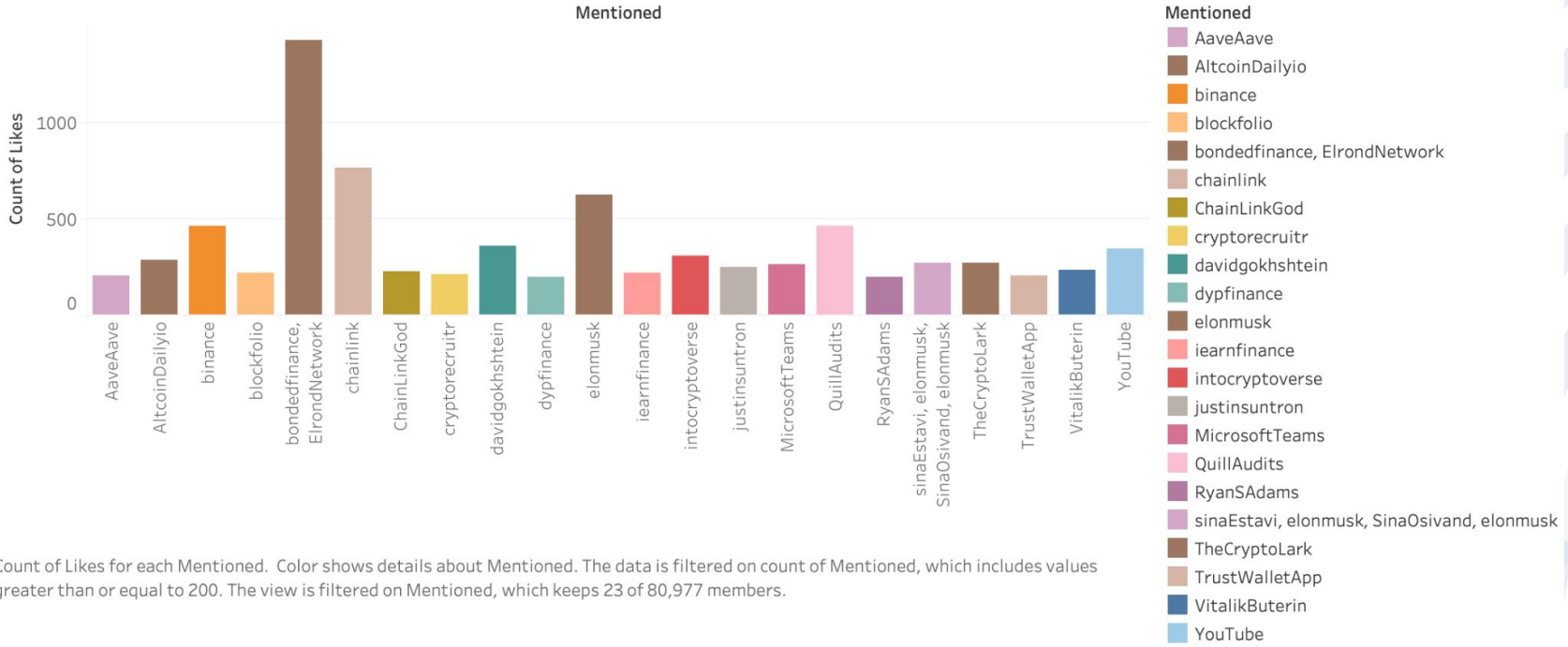
Correlation Matrix



Redder → Stronger Correlation
Bluer → Weaker Correlation

Likes vs. Top Mentions

Likes for tweets with top mentions

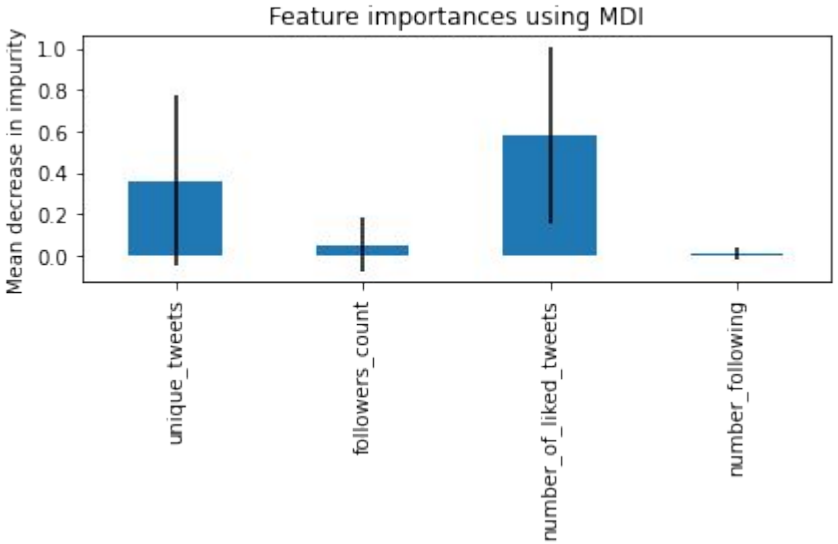


Count of Likes for each Mentioned. Color shows details about Mentioned. The data is filtered on count of Mentioned, which includes values greater than or equal to 200. The view is filtered on Mentioned, which keeps 23 of 80,977 members.

Random Forest Model

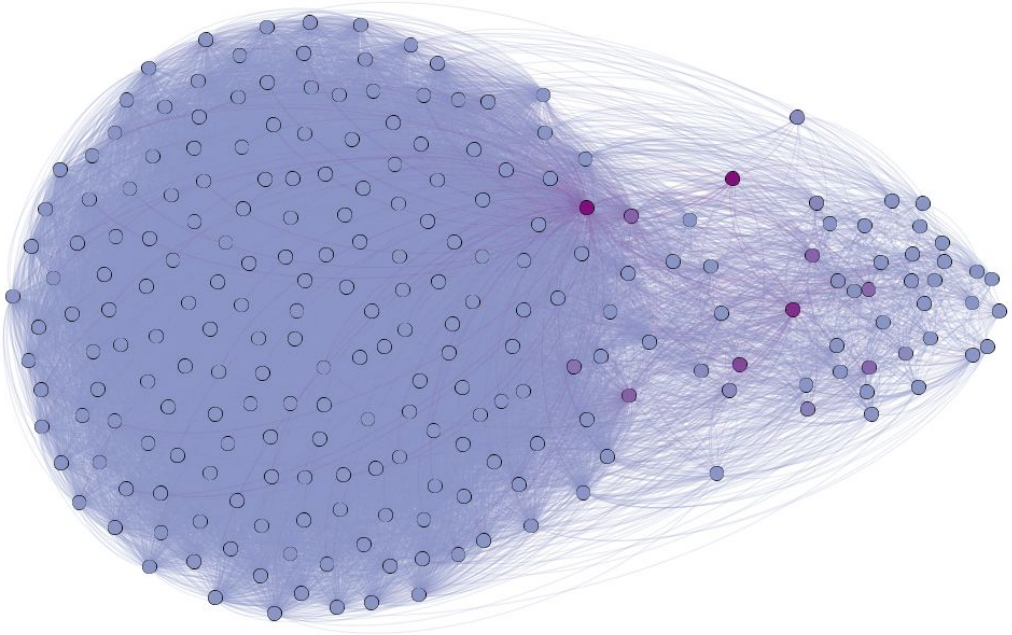
- Random Forest Model / Linear Regression
- Cleaning Methods
  - Balanced Data
  - Correct Data type
- Features
  - Unique Tweets
  - Number of Liked Tweets
  - Number of Followers
  - Amount of People Following
- Accuracy: 90.11%

Model Evaluation



	True Verified	False Verified
True Predict	181108	20657
False Predict	18876	179327

Future Network analysis



Can make use of mentions in tweets

Future Steps

- Add verified vs top mentions
- Add more user bio based data to the random forest model

Conclusions

- A machine learning approach for twitter verification is very effective
  - Model can be greatly improved by add connections to pre-verified users and keyword analysis of bio.
- Number of likes has a strong correlation with verification