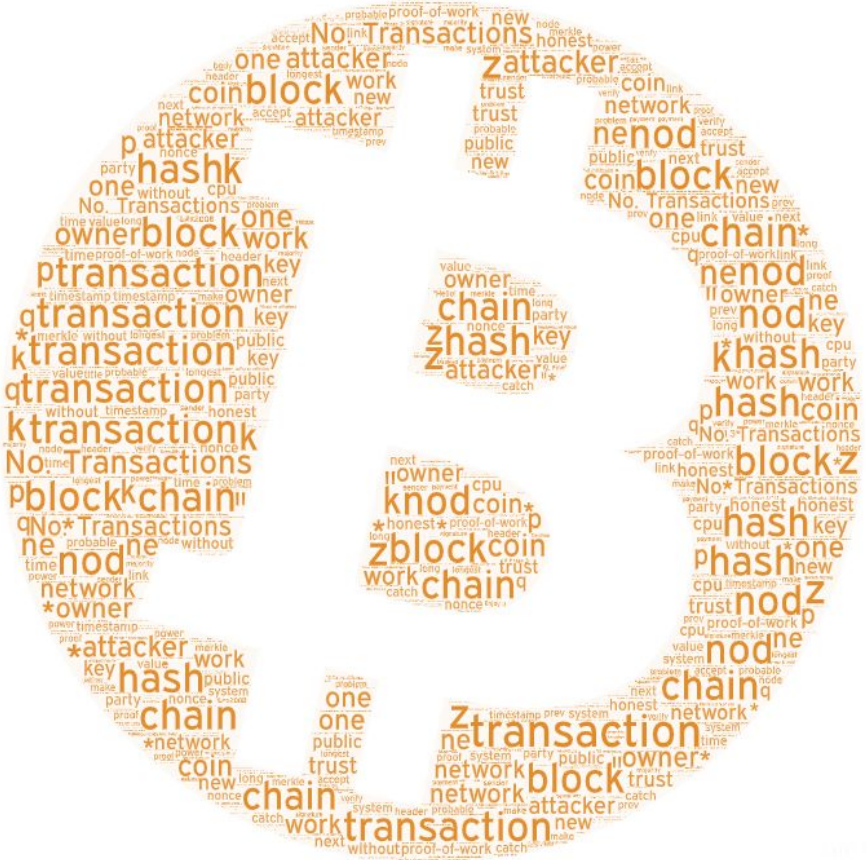
What's the subreddit?

/r/bitcoin vs. /r/wallstreetbets

Justin Fischer
GA-DSI-11-DEN

Goal

- Attempt to classify submission titles, bodies or comments to which subreddit they are from using NLP and different classification models



No. Transactions
one attacker
zattacker
coin block work
network
trust
p attacker
nenod
hash k
coin block
owner block work
chain
p transaction
nenod
q transaction key
owner
chain
zhash key
k transaction
khash
q transaction
hash
k transaction k
No. Transactions
block chain
pblock chain
knod coin
No. Transactions
hash
zblock coin
nod
nod
chain
network
*owner
*attacker
hash
nod
chain
chain
one
ztransaction
network block
chain
transaction

Msft
Out
More
Look
Nice
Shit
Back
Re
Good
Time
Make
Fuck
Day
Post
Put
Go
Up
See
One
Here
Still
Stock
Now
Year
Buy
Sell
Money
Bruh
Call
Sub
Know
Guy
Trade
Market
Think
People

Setup

- Used the PRAW API to scrape the top 1,000ish posts from each subreddit
  - Captured the title, author, date, number of comments, url and the body of the submission
    - Body is the main part of a “self-post”
- Iterated over each submission to get all the top-level comments for each post
  - Removed all deleted comments
-

	/r/bitcoin	/r/wallstreetbets
Posts	990	980
Comments	580,871	660,912
Self-Posts	118	168

Vectorizer

- Used the Tfidf Vectorizer to process each input
- Gridsearched a number of parameters:
  - Max Features: 2500, 5000, No max
  - NGram Range: (1,1), (1,2)
  - Max Documents: 90%, 95%
  - Min Documents: 5%, no minimum
  - Stop words: None, English
  - CV: 5
  - Total: 240

Models

- Logistic Regression
  - Default parameters
  - 240 models
- Bernoulli Naive Bayes
  - Default parameters
  - 240 models
- Support Vector Classifier
  - Kernel: RBF, Polynomial
  - Degree: 2, 3
  - 960 models
- Total: 1,440 per test set (4,320 in total)

Tests

- Only Title
- Only Comments
- Both Comments and self-posts
- For the second two, took a sample of 20,000 rows from each subreddit to model from since it was taking too long

Title

	Training	Testing
Logistic Regression	68.79%	64.91%
Bernoulli NB	68.79%	64.91%
SVC	75.15%	64.91%

Comments

	Training	Testing
Logistic Regression	60.73%	59.95%
Bernoulli NB	58.82%	58.5%
SVC	58.85%	58.5%

Comments and Self-Text

	Training	Testing
Logistic Regression	60.7%	60%
Bernoulli NB	58.53%	58.53%
SVC	58.53%	58.38%

Conclusions

- Getting the exact same accuracy score for multiple models seems wrong
  - Even with the same data
- When just working with titles, the SVC performed the best
  - Parameters: RBF Kernel, 3 degrees
  - max_df=0.9, max_features=2500, min_df=0.05
- With the comments, Logistic Regression outperformed the others
  - Comments + Body: max_df=0.9, max_features=2500, min_df=0.05, ngram_range=(1, 2)
  - Comments: max_df=0.9, max_features=2500, min_df=0.05, ngram_range=(1, 1)

Next Steps

- Try with subreddits that will have less similar language
- Increase the size of the dataset
- Not use ‘all-time’ top posts
- Likely any of these would need to be done via AWS