# Project 3: Web APIs & NLP

By: Jayson Villena
01/18/2023

# Problem

A coworker was tasked with gathering data of individual investors and current thoughts about the market. He gathered the data from two subreddits, WallStreetBets and StockMarket. The senior data scientist likes the content but he doesn't want to utilize data from wsb, however the worker didn't label where the subreddits came from.

## Problem Statement:

Collect data from the two subreddits and create a model that can predict where the data came from and match it with the original data.

# What do we know?

- Original data
- Equal Amount

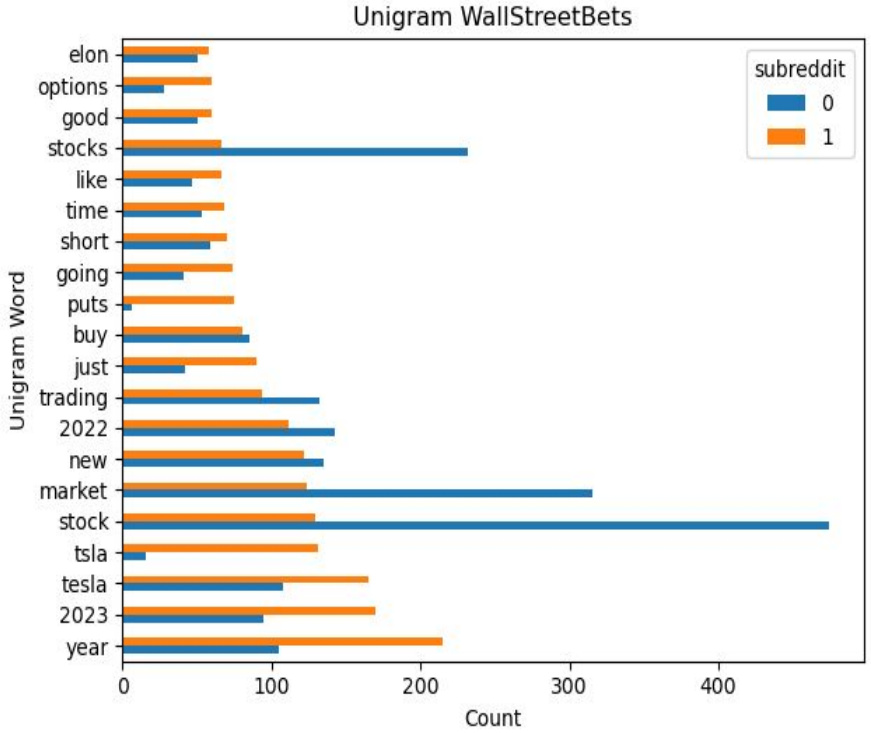
What do we need to do?

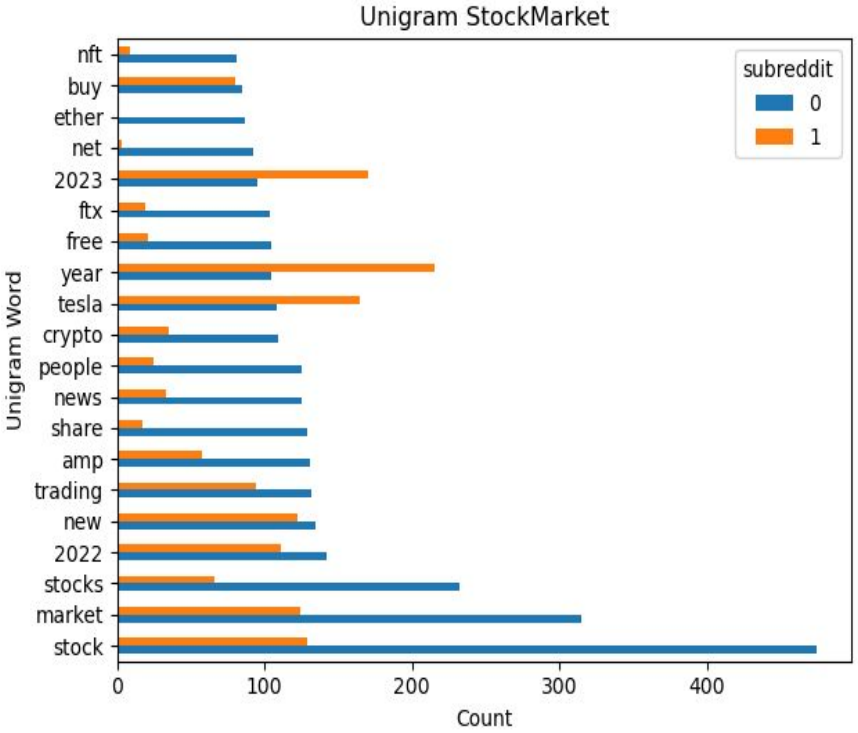
- Gather new data
- Create a predictive model
- Use model on original data

# How?

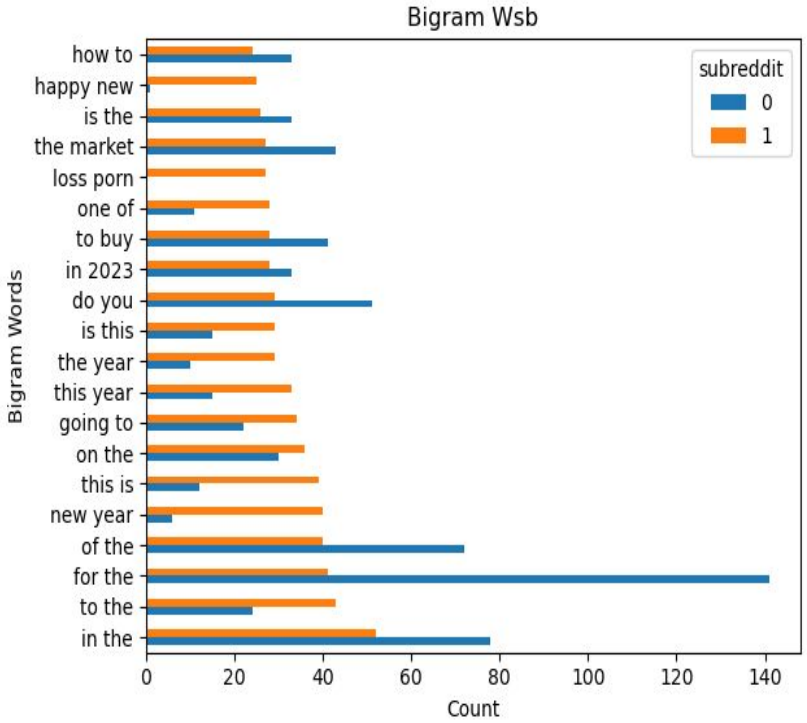
1. API
2. Preprocess and EDA
3. Model Optimization

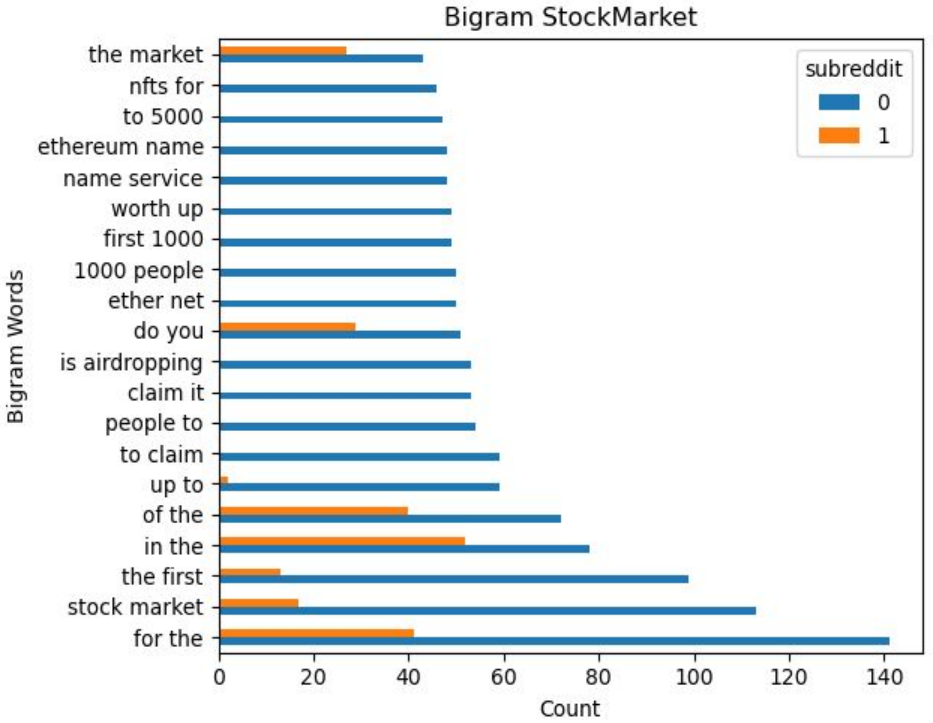
# What are we looking for?



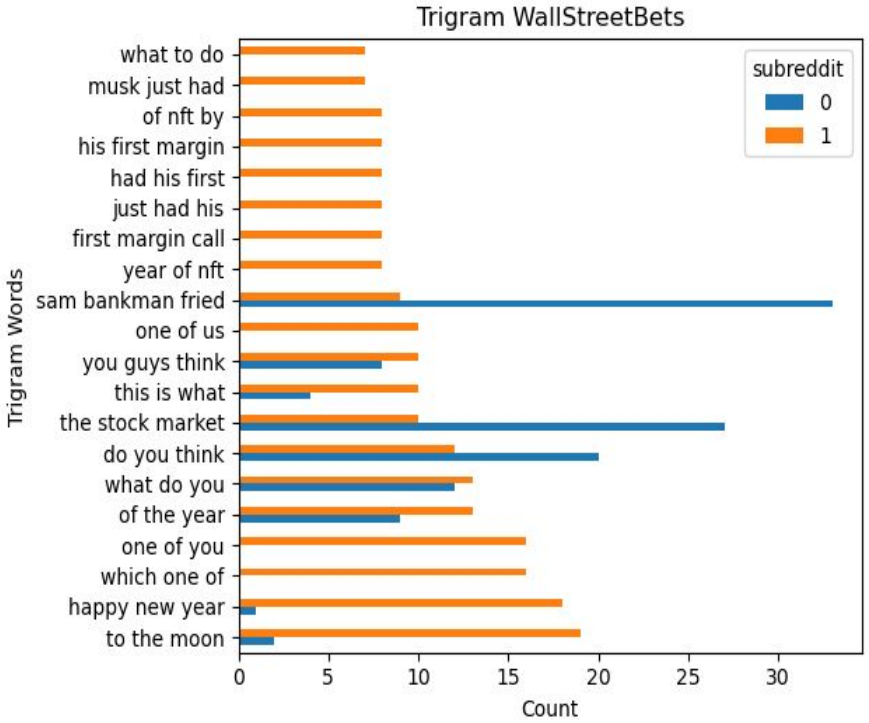



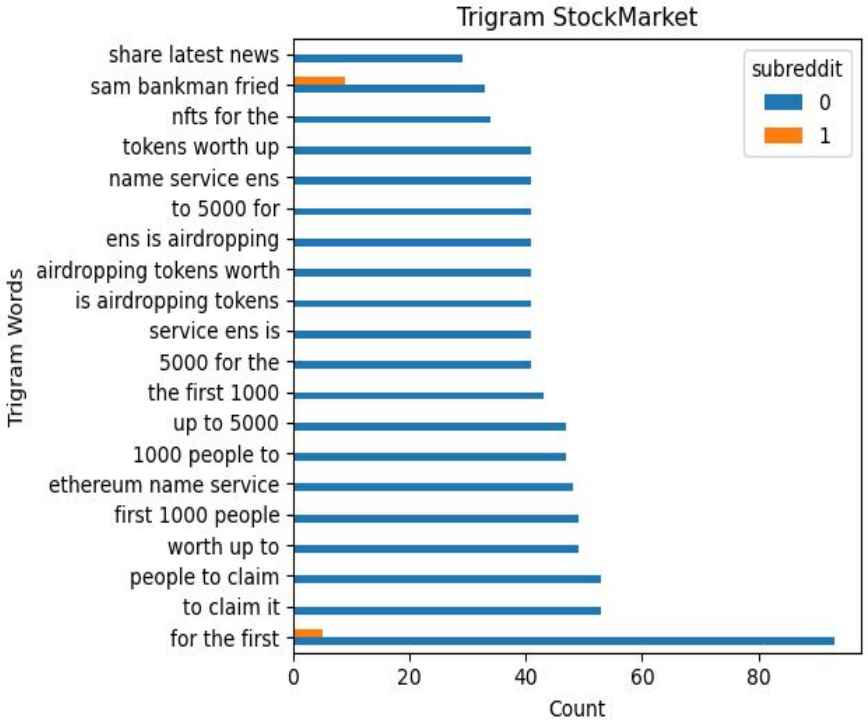
# Bigrams


Bigram Wsb
how to
happy new
is the
the market
loss porn
one of
to buy
in 2023
do you
is this
the year
this year
going to
on the
this is
new year
of the
for the
to the
in the
Bigram Words
0
20
40
60
80
100
120
140
Count
subreddit
0
1

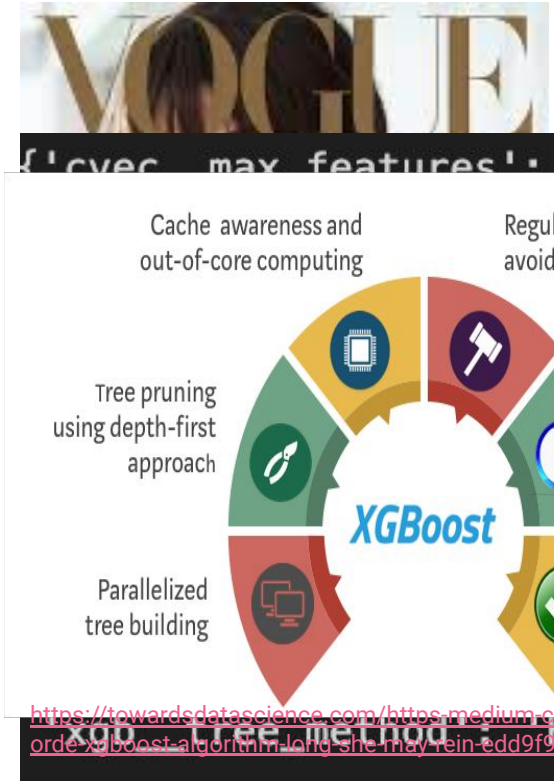

Bigram StockMarket
the market
nfts for
to 5000
ethereum name
name service
worth up
first 1000
1000 people
ether net
do you
is airdropping
claim it
people to
to claim
up to
of the
in the
the first
stock market
for the
Bigram Words
0
20
40
60
80
100
120
140
Count
subreddit
0
1

# Trigrams


Trigram WallStreetBets
Trigram Words
what to do
musk just had
of nft by
his first margin
had his first
just had his
first margin call
year of nft
sam bankman fried
one of us
you guys think
this is what
the stock market
do you think
what do you
of the year
one of you
which one of
happy new year
to the moon
subreddit
0
1
0
5
10
15
20
25
30
Count

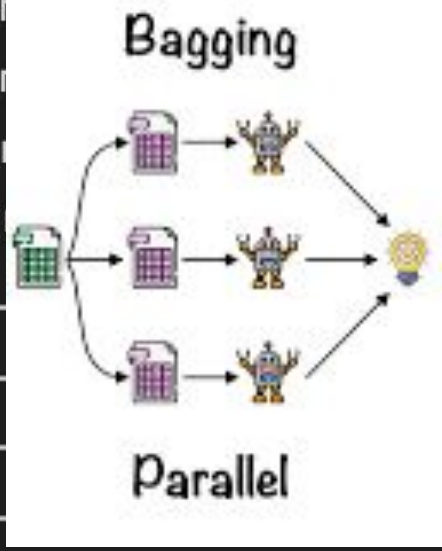

Trigram StockMarket
Trigram Words
share latest news
sam bankman fried
nfts for the
tokens worth up
name service ens
to 5000 for
ens is airdropping
airdropping tokens worth
is airdropping tokens
service ens is
5000 for the
the first 1000
up to 5000
1000 people to
ethereum name service
first 1000 people
worth up to
people to claim
to claim it
for the first
subreddit
0
1
0
20
40
60
80
Count

# Modelling




https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

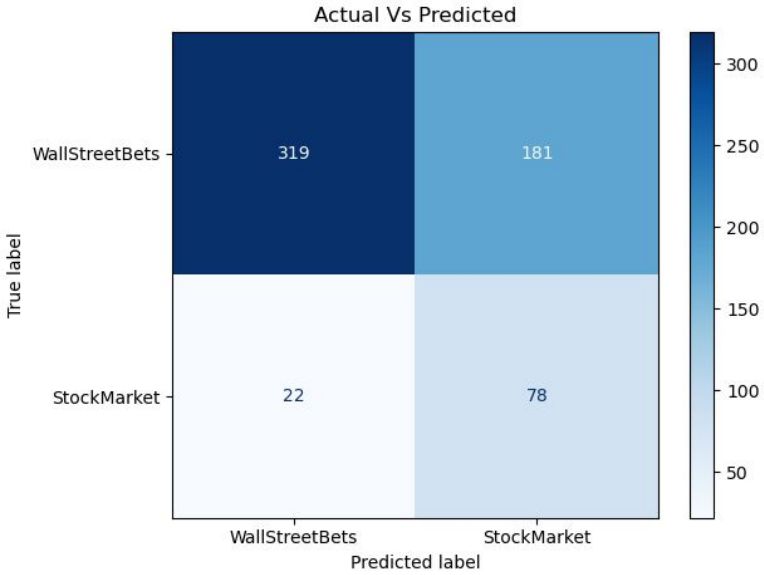
Bagging




https://towardsdatascience.com/ensemble-learning-bagging-boosting-8079e5422

# Results

XgBoost: 0.71  Bagging: 0.72




# Future:

- Better understanding of the  API
- Optimization

# Project 3: Web APIs & NLP

By: Jayson Villena
01/18/2023