Data Mining Final Project Report

Daniel, Stanley, Jeff,other guy

April 2023

1 Objective

The objective of this project is two fold:

1. If Twitter could be used as a reliable predictor of market movement (for a the day or in the future).
2.

2 Data collection

To achieve this, we need to get two data sets, and engineer them to fit the format we expect. Namely want our data points to represent individual tweets with all relevant data like handle, key words used, sentiment score, date, and direction the market when that day.

2.1 Tweets

Twitter is not a monolith, it a mixed bag of nonsense, and information, genuine and satire,

So when we talk about Twitter, we actually mean a subset of tweets that were tweeted by reputable handles/accounts, those being:

- @DeItaone
- @PriapusIQ
- @financialjuice
- @business
- @Reuters
- @WSJ
- @YahooFinance
- @FirstSquawk
- @unusual_whales
- @zerohedge
- Date: an exact time stamp of the tweet.
- Context: the text context of the tweet.
- Account: the twitter handle of the account that posted said tweet.

After merging the tweets from every account, the data-set had $\tilde{1}$16 thousand tweets/ data points.

2.2 NASDAQ

The other data set that we looked at was the NASDAQ index which will be our representative for the "market". The raw data set represented a the opening, closing, high and low of the NASDAQ for a trading day. This data set notably lacked a "movement" attribute, which would be our target. Some engineering was required. First we need to define what it means for a market to to go "UP" or "Down" as either a positive or negative difference in a particular days closing and the closing of the previous day. This calculation was done on Excel rather then on the source code.

2.3 Data cleaning

We realized early on that there were a number of issue that would need to be rectified before we could merge the two data sets into our actual dataset.

1. The different date format between the tweets and the NASDAQ index.
2. The NASDAQ index is not open on weekend or the various holidays.
3. The disparity in the range of dates of the tweets in the data set.

2.3.1 Solving 1

The Twitter date attribute was a very perceive time stamp with time zone information and other irrelevant informational. Additionally, the NASDAQ index data attribute had a different calendar standard of MM/DD/YY whereas the Twitter used the DD/MM/YY standard. Standardizing the data formats was easy in excel, as such, this work was done there rather then on in the source code

2.3.2 Solving 2

We contemplated two solutions, predict or randomly assign a direction for these missing dates, or, drop these data points. We decided on the later as trying to intelligently predict the direction of these tweets is a goal of this project. We thus decided to instead drop these data points as introducing randomness to the data set would be messing with the already fragile data set we were working with.

2.3.3 Solving 3

Some data set stretched back to around middle of 2022, where as some could didn't make it to 2022 at all. To to prevent there being over-representation in the future, we found that January 3rd, 2023 was a good cut off point point and truncated all of the individual twitter handles data sets. After this step, we were left with $\tilde{9}8$ thousand Tweets.

2.4 Scoring Tweets

We decided to use a Bag-Of-Words approach to determine the sentient analysis of the tweet. However, as the vernacular of Twitter is a more "unique" then that found among spoken dialects, the particular vernacular of financial twitter is even more niche (especially when you go into the various sub-communities). So we decided that we would greater our own collection of words and manually classify them. Rather then looking for words that give positive or negative sentiment, we look for words that have a "bearish" or "bullish" sentiment (denoting downwards or upwards respectfully). We also have a superlative category for each. From this we assigned each data point a Score, that is initially zero, and goes up as more bullish words were found in the Content attribute of the data point, and down as bearish words are discovered. We tried a few approaches to the actual bag-of-words. Our initial collection was built by Stanley from his experience within the community, and was further expanded when we tokenized the tweet and looked at the most frequent words. We also tried a version that used the SentimentIntensityAnalyzer modular from the VaterSentiment library, and also one that had a separate bearish and bullish score.

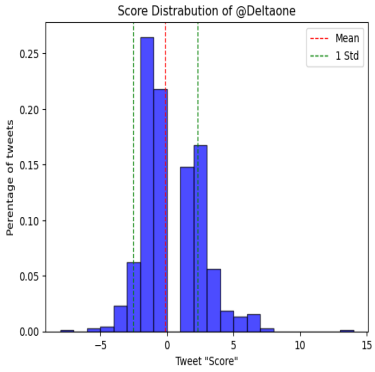
2.5 Merging Data sets

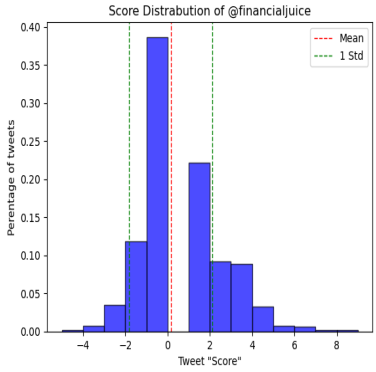
The way in we merged was effectively the same way we could assign a target to the data points. However this task was not as trivial as a .merge() function call. We wanted to match tweets of a specific date, to a target of "Up" or "Down" depending on if the direction of the NASDAQ index data set. To do this awkward comparison and assignment, we first grouped the tweet data set by data. Then turned the NASDAQ index into a map/ hash table where they key was the date and the value was the market direction. From then we could apply a lambda function to assign the a new column value for a grouped set of tweets to be what the the NASDAQ map. The result of which was our final data set that we would be working with.

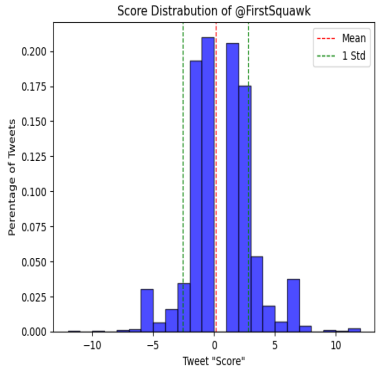
3 EDA

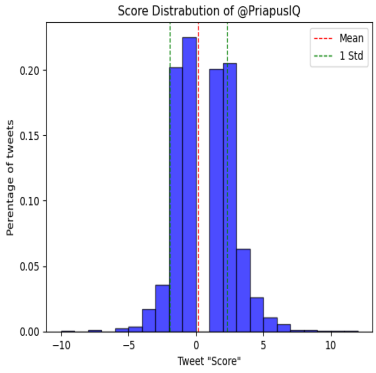
3.1 Distribution of Tweet Sentiment Across all Handles

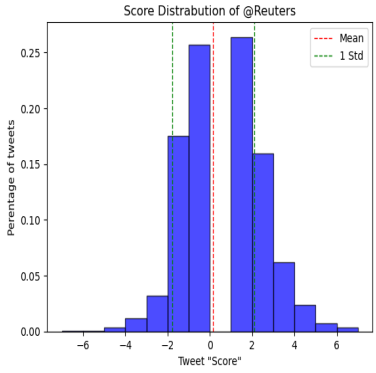
One attempt to improving the model's improvement was to turn the binary encoding to be a weight rather then a one. This weight would be determined by analyzing the performance of predictive performance of each individual handles data set, or if they demonstrated a certain propensity for a negative or positive score, etc. However, it seemed that overall, there is a slight negative sentiment with a few being more towards neutral.

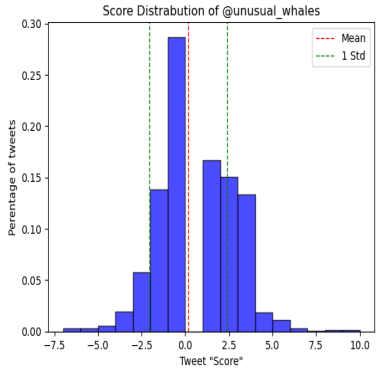

Score Distrabution of @Deltaone
Mean
1 Std
Perentage of tweets
Tweet "Score"

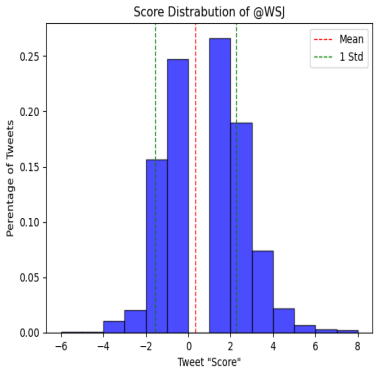

Score Distrabution of @financialjuice
Mean
1 Std
Perentage of tweets
Tweet "Score"

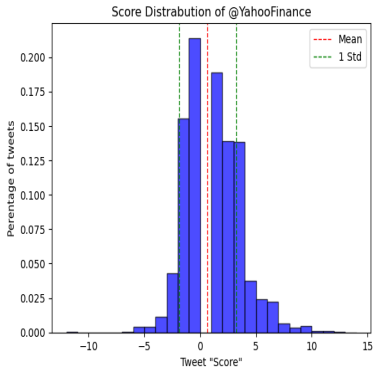

Score Distrabution of @FirstSquawk
Mean
1 Std
Perentage of Tweets
Tweet "Score"

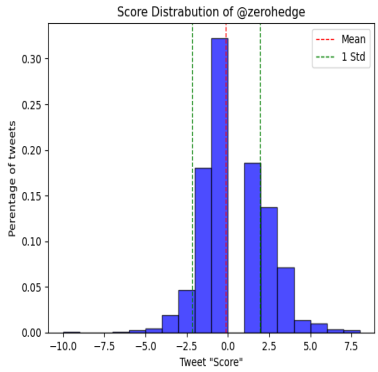

Score Distrabution of @PriapusIQ
Mean
1 Std
Perentage of tweets
Tweet "Score"

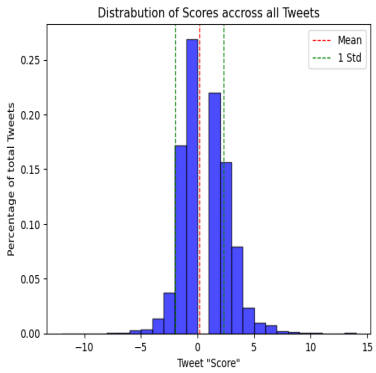

Score Distrabution of @Reuters
Mean
1 Std
Perentage of tweets
Tweet "Score"


Score Distrabution of @unusual_whales
Mean
1 Std
Perentage of tweets
Tweet "Score"


Score Distrabution of @WSJ
Mean
1 Std
Perentage of Tweets
Tweet "Score"


Score Distrabution of @YahooFinance
Mean
1 Std
Perentage of tweets
Tweet "Score"


Score Distrabution of @zerohedge
Mean
1 Std
Perentage of tweets
Tweet "Score"


Distrabution of Scores accross all Tweets
Mean
1 Std
Percentage of total Tweets
Tweet "Score"

4 Bag-Of-Words-1

	accuracy
GausianNB	.5277
Random Forest	.5306
Decision Tree	.5283
Logistic Regression	.5350

5 Improvements

A clear flaw of the model was that it, for models that only accepted numeric attributes, most data was either binary, or had a small range. This stems from the largest flaw in our data set, that being that there are not many features to work with (for non-time series models). This was the motive for trying to add features by adding another bag of words, or splitting the sentiment into two scores rather then one.

6 Bag-Of-Words-2

	accuracy
GausianNB	.5234
Random Forest	,5315
Decision Tree	.5389
Logistic Regression	.5259

7 Vater-Bag-Of-Words

	accuracy
GausianNB	.5354
Random Forest	,5262
Decision Tree	.5337
Logistic Regression	.5242

8 Time-series (sort of)

If the sentiment of the tweets was a specific way, that number have a tie with the market index in the near future rather then immediate day or day after. Also, all of the models before worked off a single tweets, what if the answer lies in an aggregate analysis of twitter. To do this we aggregated the score for the day to see if it correctly predicted up or down by sifting over the data. The results where more or less the same for the single tweet model, however, we found that the greater the difference, the worst the prediction was, with the worst prediction being for 5 days after at 40 percent accurate.

9 Conclusions

Theoretically, an accuracy of 50 percent was not bad, but the fact that we were also some what above 50 percent told us that something about this was right. However, for results that were lower then 50 percent did not mean that the model was bad, rather that it the score had an inverse relation to the target. In any case, both can be considered a good thing, that is, we don't really expect our accuracy to be about 50 percent, plus or minus 3 percent. Anything past this is desirable as one could simply do the opposite of the bad model to make money.

But As for aggregate data, we concluded that our data set simply did not span enough time to be reliable. Regardless, this project, though far from perfect, is something that

Interestingly, on the day of our presentation, Stanley found a report released by the Federal Reserve titled, "More than Words: Twitter Chatter and Financial Market Sentiment."

The bottom line is, In the rapidly changing world of technological advancement, one must either get with the times, or get left behind as the creation of programs similar ours or the aforementioned government is inevitable. One can only begin to wonder how or if AI will play a role in this, especially when there are billions of dollars to be made in this still developing market. Coupled with a rising interest in the stock market, it will only be a matter of time before someone makes a breath through. To return back to the scope of this project, what we learned from it was that, as much of a casino as it is, the market be leveraged to give the everyday people a chance.