Stock price movement prediction based on tweets

Prepared by: Jorge Duran

March 2023

Disclaimer

The information contained on this report is not intended and shall not be understood as financial advice. This report is intended for informational or educational purposes only.

Introduction

This project is centered around the US stock market, so we'll start with some definitions.

U.S. Stock Market
The term stock market refers to several exchanges in which shares of publicly held companies are bought and sold. Such financial activities are conducted through formal exchanges and via over-the-counter (OTC) marketplaces that operate under a defined set of regulations. [1]

U.S. stock market exchanges are typically open between 9:30 am and 4:00 pm Eastern Time (ET). However, with the adoption of new technology and increased demand for trading, these hours have been extended to include what is known as pre-market and after-hours trading.

Pre-market hours take place from 4:00 am to 9:30 am ET, but these can vary depending on the online broker, most pre-market hours transactions take place between 8:00 am and 9:30 am ET. After-hours take place from 4:00 pm to 8:00 pm ET, but also can vary depending on the online broker, most after-hours transactions take place between 4:00 pm and 6:30 pm ET. [2]

Retail Investor
A retail investor, also known as an individual investor, is a non-professional investor who buys and sells securities or funds that contain a basket of securities such as mutual funds and exchange traded funds (ETFs). Retail investors execute their trades through traditional or online brokerage firms or other types of investment accounts. Retail investors purchase securities for their own personal accounts and often trade in dramatically smaller amounts as compared to institutional investors. Retail investors have a significant impact on market sentiment, which represents the overall tone in the financial markets. [3]

Project Description
According to a recent article in Yahoo Finance, "over the past month retail investors funneled an average of $1.51 billion each day into U.S. stocks, the highest amount ever recorded". [4]

One can speculate that this increase in participation can be related to expanded and better access to tools, easier to use interfaces, ability to buy fractions instead of whole shares, etc. We can also assume that the retail investor range can go from someone with zero or limited

[1] https://www.investopedia.com/terms/s/stockmarket.asp
[2] https://www.investopedia.com/ask/answers/06/preaftermarket.asp
[3] https://www.investopedia.com/terms/r/retailinvestor.asp
[4] https://news.yahoo.com/retail-investors-record-inflows-us-stock-market-193801422.html

knowledge on how the market works to people with knowledge on technical and fundamentals analysis.

As mentioned before, social sentiment from retail investors can have a significant influence on stock price changes and we want to use this to predict stock price changes or movements.

The objectives of this project are to analyze social sentiment from tweets related to 3 stocks and build a predictive model to determine if the stock prices will go up or down in the next market open (9:30am ET).

The analysis and predictive model will be based on tweets from market close (4:00 pm ET) to next day market open (9:30 am ET). The predictions from the model can be used to sell stocks on pre-market hours or at market opening.

The 3 stocks we chose are Google (GOOGL), Exxon Mobil (XOM) and JP Chase Morgan (JPM). The reason why we picked these is to get some contrast from different market sectors, since they could behave differently over time. As we can see below on Fig. 1, Exxon Mobil had a very good year in 2022 while Google and many other stocks had a very bad year.

Company	Stock ticker	Sector	5 yr perf. 2018 - 2022	1 yr perf. 2022
Google	GOOGL	Communication Services	68%	-61%
Exxon Mobil	XOM	Energy	32%	80%
JPMorgan Chase	JPM	Financials	25%	-16%

Fig. 1

Data Wrangling

Our data wrangling process started with data acquisition. We extracted tweets from 2018 to 2022 that were in english and had at least 1 like. Rationale behind the 1 like was to filter out many “spam” or “vague” tweets that just mention a bunch of tickers with no real information on the tweet. We extracted the tweet's date, time, text, username, and count of likes, replies and retweets. We also added the hour, minutes and day of the week the tweet was posted.

Stock price historical data was downloaded from Yahoo Finance for the period 2018 - 2022. With the price close from the day and the price open from next day we added price change percentage and our target variable price movement (1 for an up movement, 0 for a down movement). We also filtered to only dates that had a price change percentage of +/- 0.3%. This eliminated days where the price movement was small.

Finally the tweet data was merged with the stock price historical data on the prediction date column and the stock. As shown in Fig. 2, our original dataset had 219,289 tweets in total and after merging with the stock price historical data and excluding tweets during market hours and days with small price percentage change, we were left with 72,534 tweets.

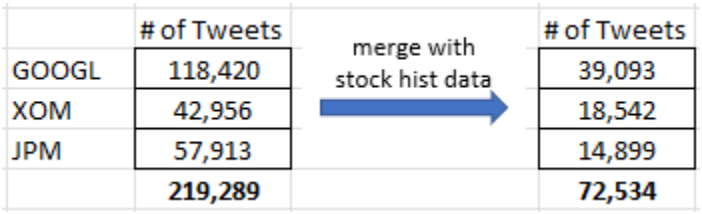


Fig. 2

After applying some common text processing practices, some additional features were added such as number of stock tickers/hashtags/mentions and text length. The number of stock tickers was used to filter tweets that only had 1 stock ticker, in this case our specific stock. We wanted to ensure the tweet was specific to the stock we were analyzing. This threshold reduced our total number of tweets from 72,534 to 16,327 as shown in Fig 3.

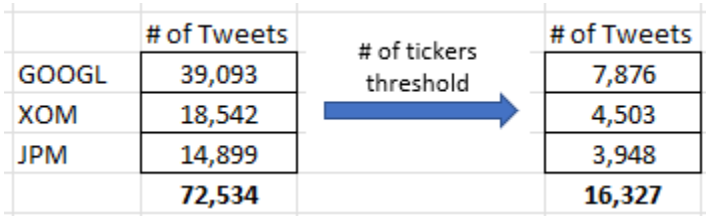


Fig.3

We then applied Lemmatization to our text, which seemed to work better than Stemming. And finally we used a sentiment analysis pre-trained model for tweets called roBERTa from Hugging Face (https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment). With this model we added a column for sentiment label (positive, neutral, negative) as well as the probability of the chosen sentiment label.

Exploratory Data Analysis

Price Movement (target variable)

The distribution of our target variable “price_movement” is shown below on Fig 3. We can see similar distribution across the 3 stocks close to a 60/40 ratio.

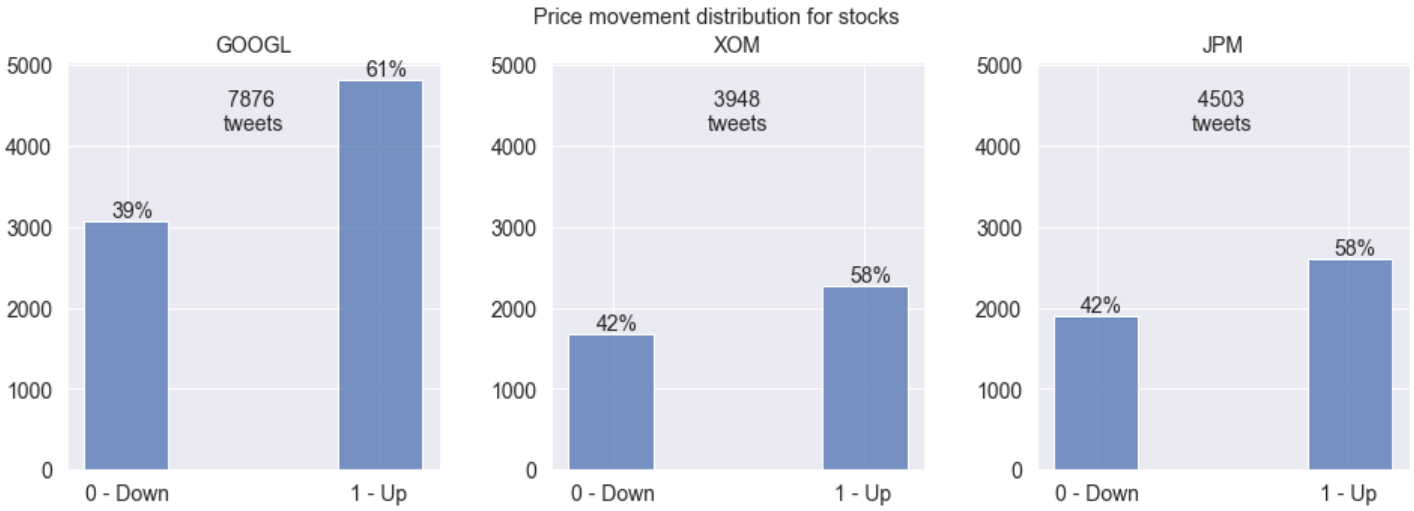


Fig. 3

Tweet text analysis

We’ll start by analyzing the text of the tweets, which seems to provide more value than the other features we captured.

Based on an exercise from Harvard’s CS109, we leveraged a technique to identify word’s predictability for an up movement.[5] This was done by creating a dataset such that each row has exactly one word or feature, represented by the identity matrix. We then used a basic Naive Bayes Multinomial classifier to make predictions on this matrix. Finally we sorted the rows by predicted probabilities and picked the top and bottom 10 words, these are our “good” and “bad” words for predicting an up movement.

As additional information we added the number of times those words appear in tweets (word frequency) and the percentage that it appears in up and down price movements.

[5] https://github.com/cs109/2015lab10/blob/master/TextAnalysis.ipynb

GOOGL - Google

On Fig. 4 we can see the top 10 “good” words for GOOGL.

word	probability	word_frequency	perc_price_up	perc_price_down
fire	0.96	45	0.87	0.13
margin	0.9	65	0.78	0.22
elliottwave	0.89	77	0.86	0.14
pullback	0.88	48	0.73	0.27
weekend	0.86	43	0.74	0.26
rocket	0.85	58	0.88	0.12
march	0.84	51	0.80	0.20
box	0.83	62	0.77	0.23
content	0.83	44	0.77	0.23
rev	0.81	725	0.63	0.37

Fig. 4 - Good words for GOOGL

Top 3 good words:

- **_fire_ (🔥):** seems to be associated with good sentiment and indicates that the stock is on a good run in terms of performance. Examples:
    - @swan the only trade that match my watchlist and ha the exact same trigger is $googl let’s get it 🔥
    - $googl absolutely crushed earnings 🔥 $b v estimate i have a position
- **margin:** seems to be related to profit, operating and ebitda margins. Examples:
    - @irbezek that’s a stretch but $googl is the best stock in the market insane profit margin
    - $googl is a $t company with ebitda margin growing top line at good lord
- **elliottwave:** according to their website, Elliott Wave International is the world’s largest independent financial and social forecasting firm. 69% of these tweets come from their twitter account. Examples:
    - $googl hr view from midday update managed to reach the blue box area amp offered a buying opportunity #elliottwave #trading #google httpstcoeegkqn
    - $googl hr view from weekend update pullback entered into a blue box area from where buyer were expected to appear #elliottwave #trading #google httpstcogwfangvle

On Fig. 5 we can see the top 10 “bad” words for GOOGL.

word	probability	word_frequency	perc_price_up	perc_price_down
reuterstv	0.15	42	0.14	0.86
miss	0.2	145	0.30	0.70
live	0.29	88	0.45	0.55
wait	0.36	69	0.51	0.49
covid	0.36	46	0.39	0.61
user	0.37	64	0.45	0.55
inside	0.37	108	0.44	0.56
bull	0.39	209	0.66	0.34
drop	0.4	93	0.43	0.57
cnbc	0.42	71	0.49	0.51

Fig. 5 - Bad words for GOOGL

Top 3 bad words:

- **reuterstv:** Reuters TV is a mobile video news service operated by the news organization Reuters. It’s worth mentioning around 70% of the tweets were on the period of May 10-12 2019 and are associated mainly with topics shown below. Examples:
    - facebook reject cofounder call for breakup senator urge u antitrust probe httpstcollbxohkzm via the @reuterstv tech playlist $googl httpstcokblmuzhmmu
    - watch google will allow user to choose how long the company keep their data more in this week tech playlist httpstcoutuqeaoej via @reuterstv $googl httpstcoynwnovm
- **miss:** related to google missing revenue, earnings or other estimates. Examples:
    - #google $googl big revenue miss stock down in aftermarket guidance disappointing too
    - oh look another earnings miss we can pretend never happened $googl
    - $googl ebit miss
- **live:** this word is almost associated 50/50 to up/down movement and it doesn’t seem to provide much context in terms of sentiment. Examples:
    - dozen of state are reportedly planning to sue google $googl alleging the company illegally abuse it power over developer on the google play store on mobile device according to bloomberg news ht @livenewsnick httpstcobebprbcpx

- $googl we nailed this both way today in the @rampeddaytrader room both trade alerted live before the move called for a long over for a point push and then a short on the break back under the same level for a point fade #cutfromadiffcloth

XOM - Exxon Mobil

On Fig. 6 we can see the top 10 “good” words for XOM.

word	probability	word_frequency	perc_price_up	perc_price_down
portfolio	0.78	48	0.73	0.27
member	0.77	72	0.74	0.26
debt	0.77	56	0.71	0.29
pay	0.76	137	0.64	0.36
say	0.74	135	0.73	0.27
bullish	0.74	110	0.74	0.26
end	0.73	817	0.57	0.43
work	0.72	83	0.63	0.37
dont	0.71	56	0.73	0.27
investor	0.71	133	0.66	0.34

Fig. 6 - Good words for XOM

Top 3 good words:

- **portfolio:** Related to the stock being part of people’s portfolio. Examples:
    - @dvdnddiplomats love seeing how my purchase are fairly in line with yours while i’ve built up a great portfolio it’s been mainly in bit amp piece too $xom ha been my # focus the past month
    - oil is dirt cheap right now bc the crisis today i added more exxonmobil to my portfolio $xom #stocks #stockmarket #investing #oil #coronavirus
    - @wealthbabylon great analysis will definitely consider adding some $xom to my portfolio next week i’ve been slowly adding position of aristocrat recently so this is awesome to read _thumbs_up:_medium-light_skin_tone_
- **member:** Refers to the company’s board and changes in its members. Examples:
    - @narutium @wholemarsblog they’re a small activist investor hedge fund that is attempting to force exxon to act responsibly on climate change impact by a ousting $xom board member amp replacing with environmentally friendly folk

    - @ovlegacyvp @radkapital can i just tell you how impressive that is compared to a board like $xom that i looked over yesterday where none of the member have oil and gas or refining experience

- **debt:** Although this word seems to be a good predictor for up movement, tweets seem to be divided in good/bad sentiment towards the company's debt. However, in aggregate it does seem to appear more frequently in a positive light. Examples:

    - @cmenergyintel funny $xom will have plenty of fcf to pay down debt and strengthen balance sheet going forward upgrade by qq
    - the oil giant drowning in debt $xom httpstcoxyynjbxyu
    - isn't $xom getting buried under debt

On Fig. 7 we can see the top 10 "bad" words for XOM.

word	probability	word_frequency	perc_price_up	perc_price_down
break	0.34	131	0.58	0.42
really	0.35	47	0.38	0.62
view	0.35	87	0.49	0.51
sector	0.4	78	0.40	0.60
interesting	0.41	48	0.40	0.60
report	0.42	66	0.52	0.48
stockmarket	0.43	67	0.51	0.49
higher	0.44	91	0.43	0.57
elliottwave	0.44	66	0.48	0.52
engine	0.45	80	0.60	0.40

Fig. 7 - Bad words for XOM

Top 3 bad words:

- **break:** Seems to be related more to technical analysis, when the stock is about to break some pattern. Examples:

    - beautiful cup and handle on $xom rise in gas price recently could cause this to break and send u to and httpstcoobppalzpmi
    - $xom in downtrend it price expected to drop a it break it higher bollinger band on january view odds for this and other indicator httpstcosjsktypi #exxonmobil #stockmarket #stock #technicalanalysis #money #trading #investing #daytrading #news

- **really:** used to add emphasis, it appears generally that emphasis is added when prospects are negative. Examples:

    - $xom #riskreversal for feb th action confirmed but due to the cheap nature of the call this is really about the sold put im going to label this instead simply #deepitmcallbuying no risk reversal
    - @andykarsner i couldn't agree more having spent time with some of the other more progressive oil and gas major it's become abundantly clear how far behind $xom really is

- **view:** this word is almost associated 50/50 to up/down movement and it doesn't seem to provide much context in terms of sentiment. It is strange therefore that it shows up as a highly predictive negative word, and this might highlight a potential limitation in this approach. Examples:

    - @thealphathought this is how i view my $xom position dividend per year will soon be greater than paper capital loss so not sweating it a cash flow is king
    - $xom enters an uptrend a momentum indicator exceeded the level on may view odds for this and other indicator httpstcodnjzzgo #exxonmobil #stockmarket #stock #technicalanalysis #money #trading #investing #daytrading #news #today httpstcomukfggfo

JPM - JP Morgan Chase

On Fig. 8 we can see the top 10 "good" words for JPM.

word	probability	word_frequency	perc_price_up	perc_price_down
buyback	0.86	78	0.86	0.14
quarterly	0.79	55	0.82	0.18
dividend	0.73	132	0.73	0.27
financials	0.71	50	0.68	0.32
make	0.71	123	0.63	0.37
result	0.71	74	0.78	0.22
investing	0.71	79	0.62	0.38
silver	0.71	111	0.67	0.33
trade	0.71	250	0.64	0.36
company	0.7	58	0.69	0.31

Fig. 8 - Good words for JPM

Top 3 good words:

- **buyback:** It appears to refer to the company's shares buyback. Examples:
    - @occupywisdom i have feeling the recipe for a rally tomorrow will be china tradecurrency news $jpm earnings and of course their latest buyback news
    - woohoo the buy back are back $jpm holder let make that money _money-mouth_face_ _money-mouth_face_ now the only thing were missing is a big fat #dividend boost #investing #investments #investors #stocks

- **quarterly:** Related to changes in quarterly earnings or dividends and it seems to be positive. Examples:
    - $jpm jpmorgan chase intends to increase it quarterly dividend by to $ per share in q following fed stress httpstcojesaqwr
    - $jpm jpmorgan boost quarterly dividend to $share
    - breaking news on $jpm gt jpmorgan chase to report quarterly earnings learn more httpstcoxhudfu #dow #stocks #swingtrading #daytrading #jpmorgan #jpm

- **dividend:** similar to the last word, related to positive changes in dividends. Examples:
    - $jpm jp morgan and ip partner to deliver taxefficient access to model portfolio httpstcooxheqwhfp great dividend stock buy jpm
    - jpmorgan $jpm announced today it will be increasing it quarterly dividend to $ per share up from their previous quarterly dividend of $
    - $jpm raise it dividend

On Fig. 9 we can see the top 10 "bad" words for JPM.

word	probability	word_frequency	perc_price_up	perc_price_down
release	0.24	71	0.24	0.76
reserve	0.32	110	0.34	0.66
eps	0.38	202	0.43	0.57
income	0.4	112	0.52	0.48
friday	0.41	57	0.47	0.53
sell	0.41	82	0.61	0.39
work	0.41	109	0.55	0.45
loan	0.42	124	0.49	0.51
growth	0.43	77	0.55	0.45
expected	0.44	71	0.39	0.61

Fig. 9 - Bad words for JPM

Top 3 bad words:

- **release:** seems to be talking about a reserve release. Examples:
    - jpmorgan smash q profit forecast with $ billion reserve release boost httpstcooutdarbda @mdbaccardax $jpm
    - $jpm with a massive beat on the face of it but drive by a far bigger reserve release than expected $bn v $bn expected dimon we do not consider this core or recurring profit but the release bigger than expected only because outlook is too
- **reserve:** same as above, seems to be talking about a reserve release. Examples:
    - jpmorgan smash q profit forecast with $ billion reserve release boost httpstcooutdarbda @mdbaccardax $jpm
    - $jpm with a massive beat on the face of it but drive by a far bigger reserve release than expected $bn v $bn expected dimon we do not consider this core or recurring profit but the release bigger than expected only because outlook is too
- **eps:** although this is a bad predictor, it's talking about earnings per share, which in JPM's case seem to be positive most of the time . Examples:
    - $jpm reported q this morning missing on eps @wilfredfrost break down the number httpstcovscapwbvj
    - @tomkeene if $jpm over reserved in the first half of and under earned and then release a large portion of those loan loss provision in q it is a poor quality eps beat @lisaabramowicz @ferrotv

Other tweet features

Day of week

Fig. 10 shows the distribution of up and down tweets by day of the week that it was posted. Tuesdays seem to be a good indicator of up movements for GOOGL, this could be related to a big event or release that happened on a Tuesday. In the case of XOM and JPM most of the days have a slight tendency for an up movement.


GOOGL - Price movement by day of the week
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
44%
85%
69%
59%
54%
54%
52%
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
up
down
XOM - Price movement by day of the week
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
65%
59%
60%
50%
59%
59%
58%
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
up
down
JPM - Price movement by day of the week
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
60%
60%
49%
60%
61%
63%
63%
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
up
down

Fig. 10

Sentiment

Fig. 11 shows the distribution of the sentiment classification done by the pre-trained roBERTa model across the 3 stocks. Distribution for sentiment seems to be about the same for the 3 stocks, with the majority having a neutral sentiment classification. We see a slight difference on price movement up distribution for GOOGL, having an increased number of positive sentiment and a decrease in negative sentiment.

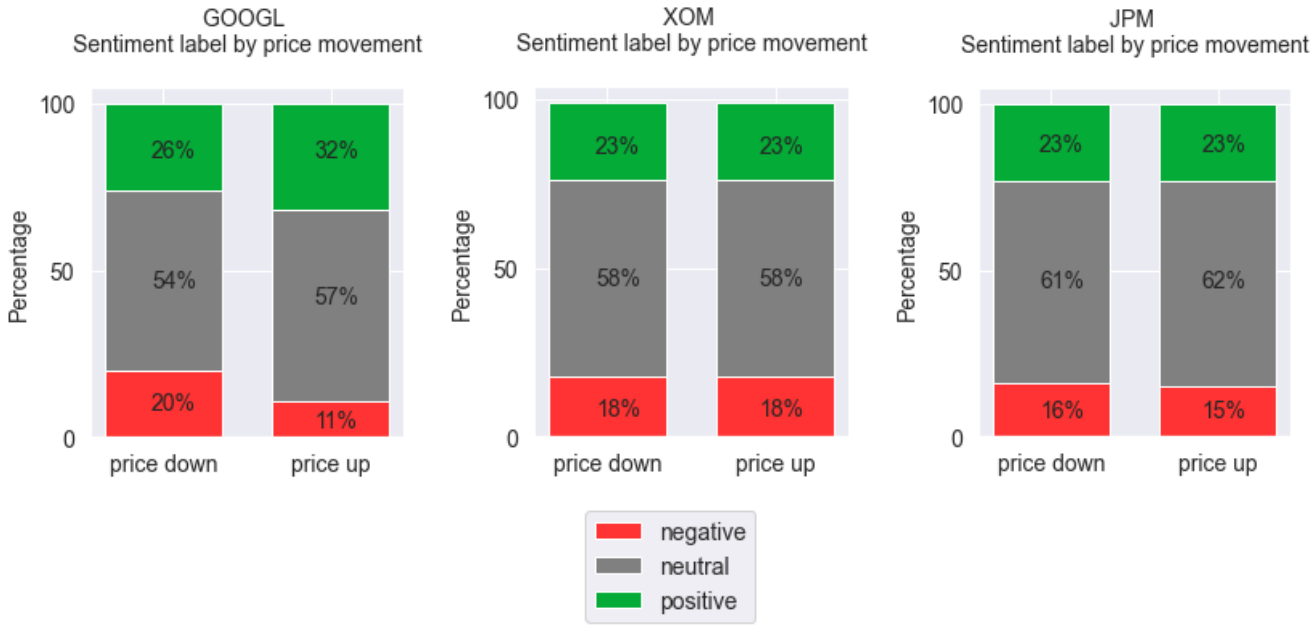


Fig. 11

Fig. 12 shows the correlation heatmap for the numerical features and we only see a low correlation between the text length and number of hashtags/mentions. We don't see any strong correlations with our target variable price_movement.

Fig. 12

Fig. 13 shows the correlation map for the categorical features. We see a slight to moderate correlation between username and our target variable price_movement. We also see moderate correlations between username and stock/sentiment_label.

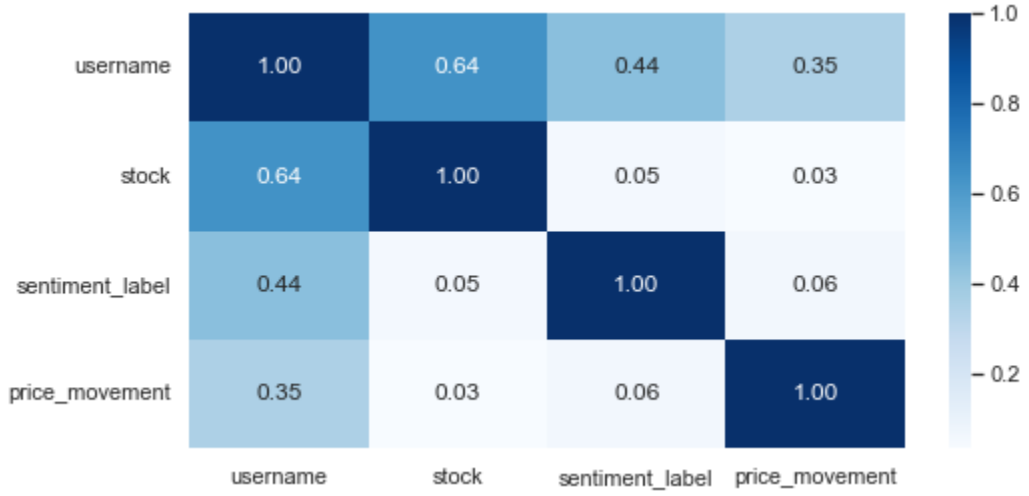


Fig. 13

Modeling

Before proceeding to modeling, we did some preprocessing of our data, such as one-hot encoding for categorical features and MixMaxScaler for numerical features since one of our algorithms (Multinomial Naive Bayes) doesn't work with negative values. The dataset was split in a 70/30 ratio for training and testing respectively. Finally Count Vectorizer was used to convert our tweets to a matrix of token counts with a minimum document frequency of 10 and only using unigrams.

For our initial model selection we built individual models for each stock instead of one model for all stocks. For each stock we built the following models: Multinomial Naive Bayes, Random Forest & XGBoost. The metric used to assess performance was the ROC AUC and for hyperparameter tuning we used Random Search with 5-fold cross validation.

Fig. 14 shows the ROC AUC scores for each of the stocks and models, as well as the best parameters after tuning.

	ROC AUC Score		
	GOOGL	**XOM**	**JPM**
Multinomial NB	0.68	0.58	0.60
Random Forest	0.79	0.73	0.69
XGBoost	**0.79**	**0.74**	**0.71**

Model Name	Best Hyperparameters
Multinomial Naïve Bayes	GOOGL - alpha: 10 XOM - alpha: 0.01 JPM - alpha: 1
Random Forest	*Same parameters for 3 stocks:* n_estimators: 300 min_samples_split: 4 min_samples_leaf: 4 max_features: 0.3 max_depth: 30
XGBoost	*Same parameters for 3 stocks:* n_estimators: 150 max_depth: 10 eta: 0.3 colsample_bytree: 0.6 booster: gbtree

Fig. 14

XGBoost was the model with the best performance for XOM and JPM, in the case of GOOGL both Random Forest and XGBoost obtained similar results. We chose to use XGBoost as the final model for the 3 stocks.

Fig. 15 shows the top 10 feature importances from XGBoost for GOOGL. As we had observed before during our EDA, Tuesdays were a very good indicator for an up movement and this shows as our top feature. We can also observe that our second top feature "margin" was the second word in our list of good predictability words. The words "march" and "elliottwave" were also part of our good predictability words list.

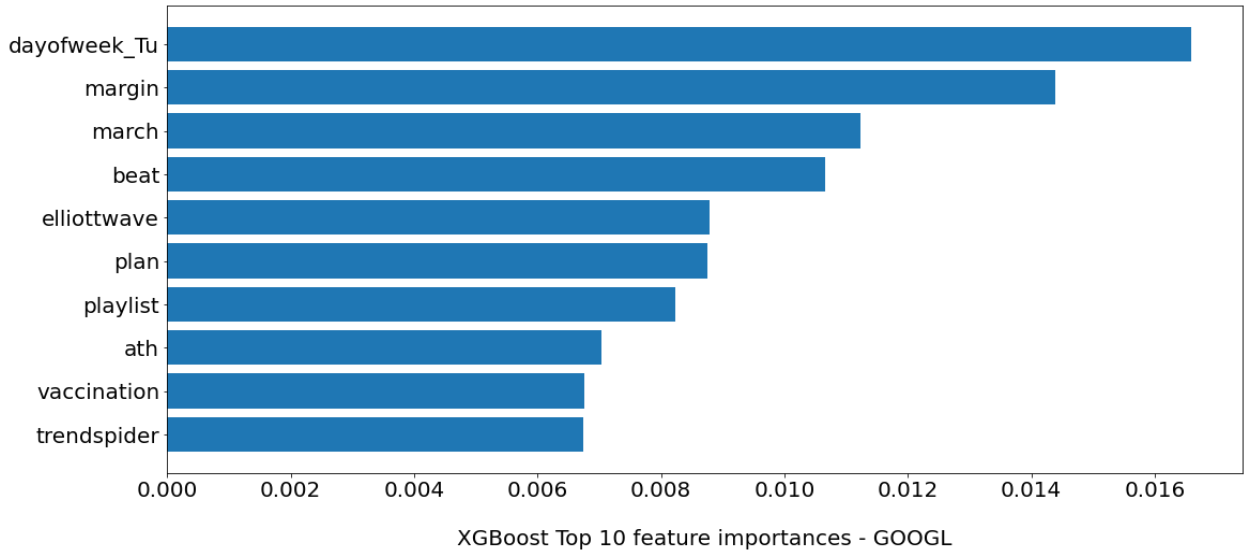


Fig. 15

Fig. 16 shows the top 10 feature importances from XGBoost for XOM. The word “member” was the second in our good predictability words and it shows as the top 5th feature. Other than this word, we don’t see any other words from our good predictability words list.

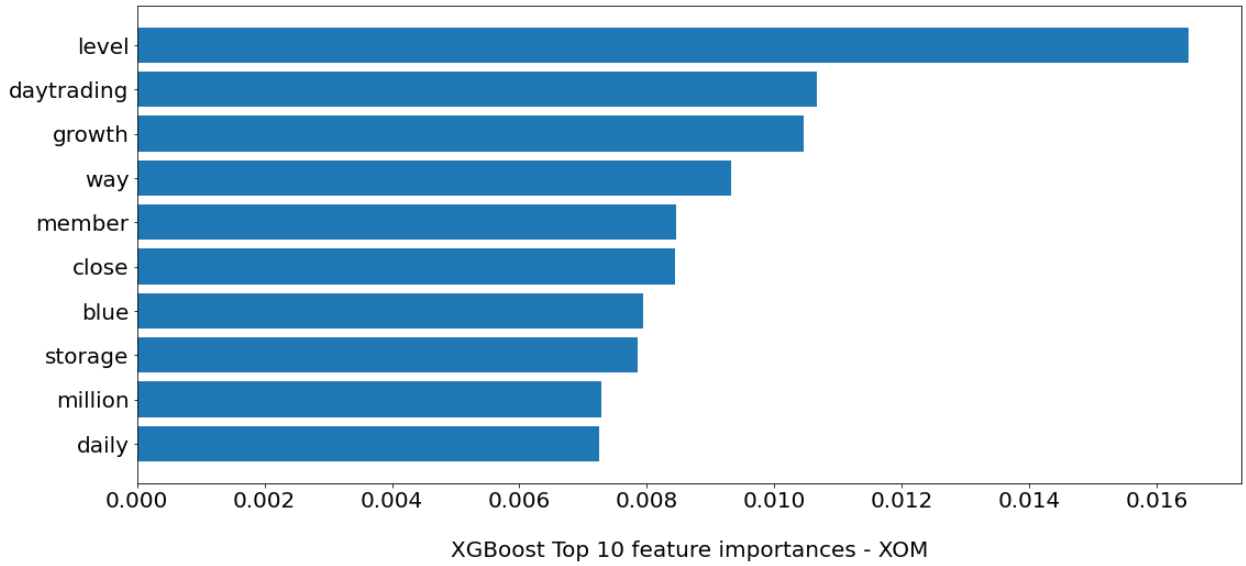


Fig. 16

Fig. 17 shows the top 10 feature importances from XGBoost for JPM. The word “buyback” was the top word in our good predictability words list and it shows here as the second top feature. Words “release” and “eps” were in the top 3 bad words for predictability and they show here in the top 10 feature importances.

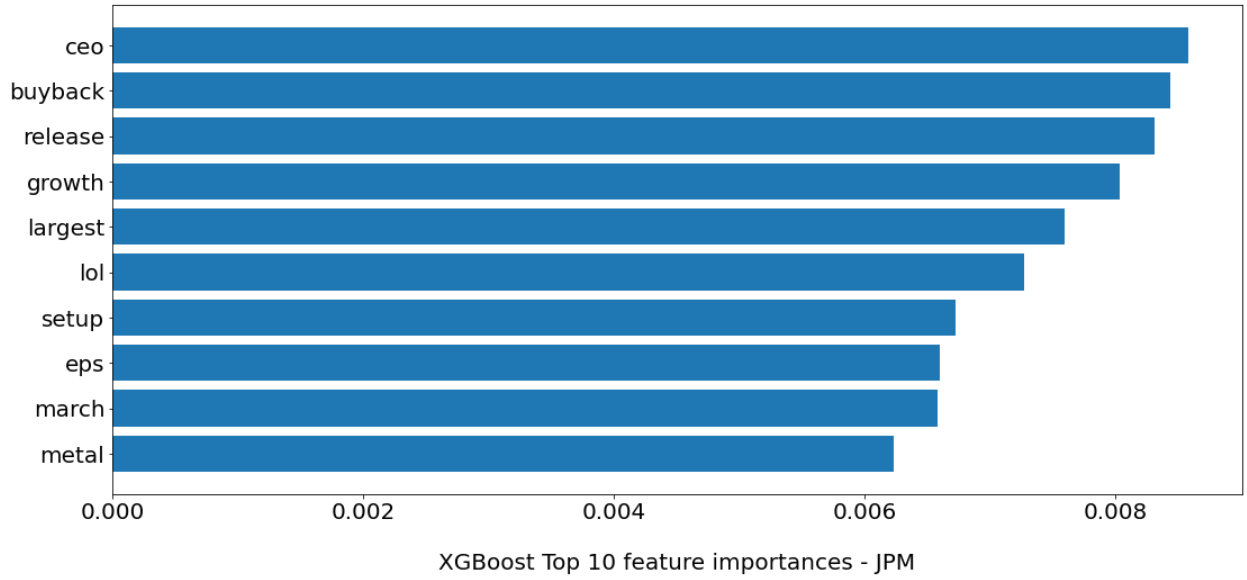
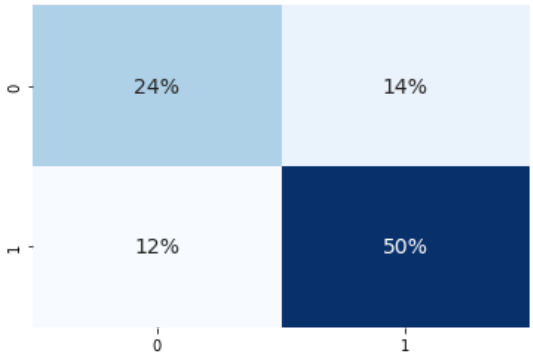


Fig.17

We then proceeded to adjust the threshold based on our objective of increasing precision (true positive up movements). We are interested in predicting up movements since the strategy for the retail investor would be to buy low and sell at a higher price for a profit. The 3 stocks threshold adjustment was very similar and not too far from the default threshold (0.5).

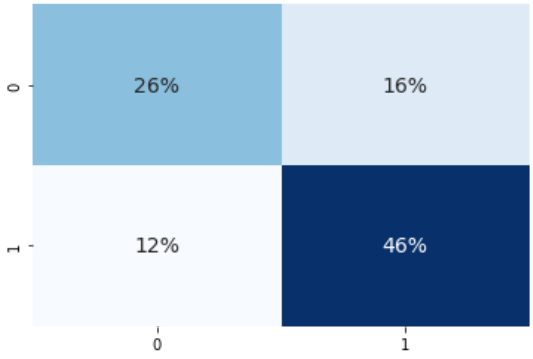
For GOOGL, the threshold was adjusted to 0.52 and a final validation was done on the test set to obtain a final precision of 0.78 and recall of 0.63. Fig. 18 shows the confusion matrix and classification report.



	precision	recall	f1-score	support
0	0.67	0.63	0.65	907
1	0.78	0.81	0.79	1456
accuracy			0.74	2363
macro avg	0.72	0.72	0.72	2363
weighted avg	0.74	0.74	0.74	2363

Fig. 18

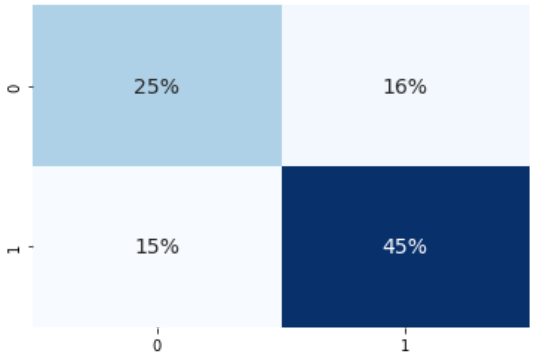
For XOM, the threshold was adjusted to 0.54 and a final validation was done on the test set to obtain a final precision of 0.74 and recall of 0.62. Fig. 19 shows the confusion matrix and classification report.



	precision	recall	f1-score	support
0	0.69	0.62	0.65	503
1	0.74	0.79	0.76	682
accuracy			0.72	1185
macro avg	0.71	0.71	0.71	1185
weighted avg	0.72	0.72	0.72	1185

Fig. 19

For JPM, the threshold was adjusted to 0.54 and a final validation was done on the test set to obtain a final precision of 0.74 and recall of 0.61. Fig. 20 shows the confusion matrix and classification report.



	precision	recall	f1-score	support
0	0.62	0.61	0.62	545
1	0.74	0.75	0.74	806
accuracy			0.69	1351
macro avg	0.68	0.68	0.68	1351
weighted avg	0.69	0.69	0.69	1351

Fig. 20

Conclusions

Although this was a proof of concept with limited scope to 3 specific stocks, our analysis and modeling showed that there is good predictive potential from certain words used in tweets and also there seems to be a contrast across different stocks/sectors. We saw better performance in our GOOGL model and this could be related to having more data than XOM or JPM.

GOOGL had a close similarity between the top features used for the model and the words we had identified during EDA as having good predictability. In XOM we didn't find that similarity and for JPM we found both good and bad words for predictability as top features in the models. We noticed also that the XGBoost models would use more words from the tweets for predictions, but Random Forest models would use other tweet features such as sentiment score, text length, hours and minutes.

Our final predictive models had precision ranging from 74% to 78% which is better than we expected. It's worth mentioning that the models seem to be a bit overfitted since we saw a 20% gap in precision between the train and test sets, this might need some additional tuning in the future or additional data. As such, it does seem that the models can be used by our retail investors as an additional metric/tool for their investment decisions.

For future work this can be expanded to additional stocks within the same or different sectors. Models and analysis could also be adjusted by years or seasons so we could see different trends in sentiment. An example of other stocks that could be analyzed are Microsoft and Amazon as they seem to have a very active participation on Twitter.

In terms of technical adjustments or tuning of the models we could try different vectorizers such as TfIdf or different lemmatizers such as Spacy. We could also try different price movement percentage thresholds, no thresholds at all or even thresholds for big moves (>10%) which only occur limited times. Another opportunity would be to not remove numbers during text processing as we noticed we lost some interpretability while reviewing tweets that mentioned important financial indicators (earnings, eps, dividends, etc).