

Stock price movement prediction based on tweets

Jorge Duran
March 2023

Definitions

U.S. Stock Market

- Refers to several exchanges in which shares of publicly held companies are bought and sold.
- Open between 9:30 am and 4:00 pm Eastern Time (ET) with extended hours trading:
 - Pre-market hours 4:00 am - 9:30 am ET (most transactions between 8:00 am and 9:30 am ET)
 - After-hours 4:00 pm - 8:00 pm ET (most transactions between 4:00 pm and 6:30 pm ET)

Retail Investor

- Non-professional investor who buys and sells securities or funds, such as stocks, mutual funds and exchange traded funds (ETFs).
- Purchase securities for their own personal accounts and often trade in dramatically smaller amounts as compared to institutional investors.

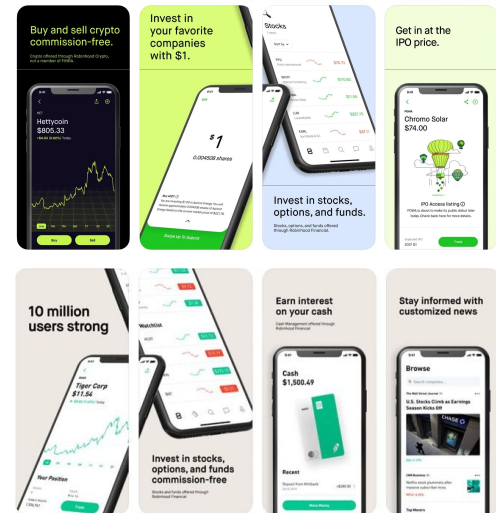
Introduction

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”***.

Increase in participation (easier access to tools, incentive programs, ability to buy fractions instead of whole shares, etc.)

Retail investors can have a significant influence on stock price changes (“power of the many”).

We want to use social sentiment to predict stock price changes/movements.



Project Objectives

- Analyze tweets related to 3 stocks from different market sectors and understand what features are good predictors for a stock price moving up. Google (GOOGL) - Communications, ExxonMobil (XOM) - Energy, and JPMorgan Chase (JPM) - Financials.



- Build a predictive model that retail investors can use as an additional tool to determine if the stock prices will move up in the next market open (9:30am ET), with precision $>60\%$.

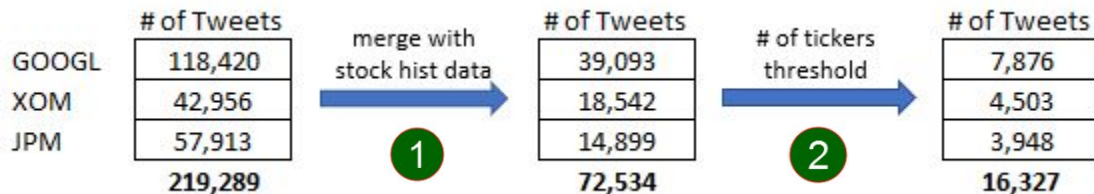
* We'll use tweets from market close (4:00pm) to next market opening (9:30am).

Data Acquisition

- Tweets were scraped for each company for 5 years (2018 - 2022) using SNScrape library
- Extracted the tweet's date, time, text, username, and count of likes, replies and retweets.
- Filters were applied to only scrape tweets in English with at least 1 like
 - During initial inspection we noticed a lot of “spam” tweets with no likes
- Number of tweets scraped:
 - GOOGL 118,240
 - XOM 42,956
 - JPM 57,913

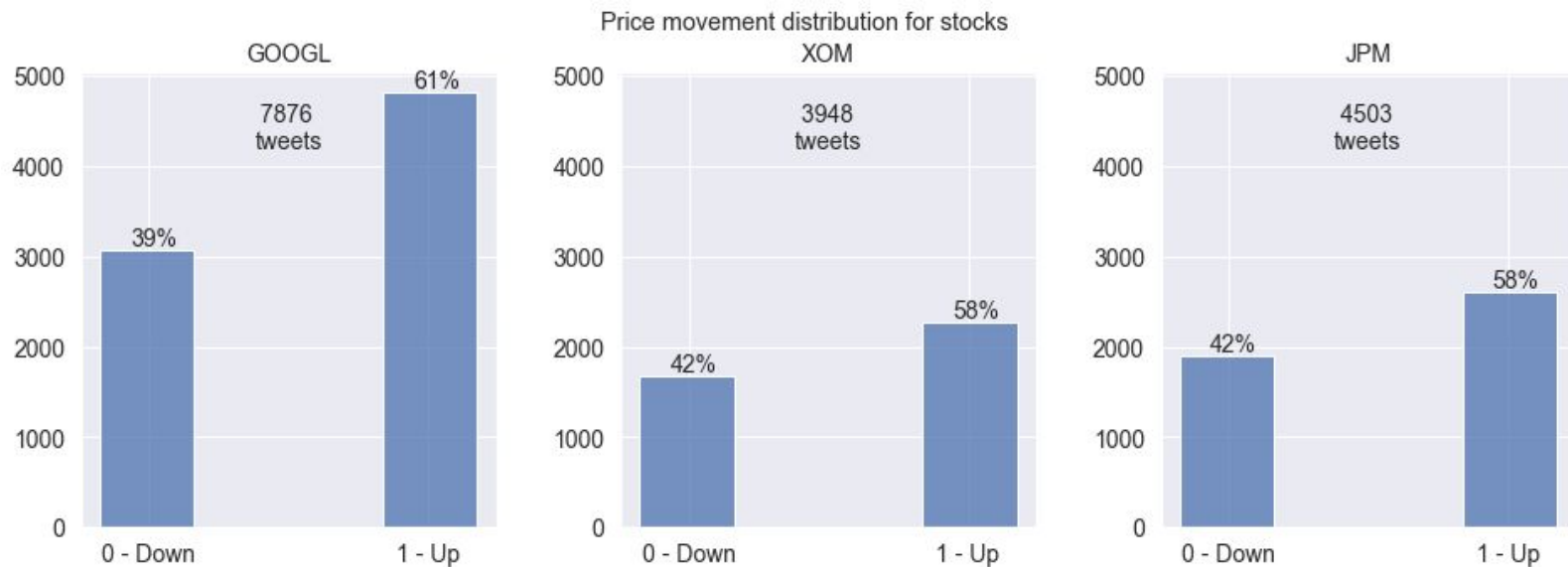
Data Wrangling

- Historical stock price data and added price movement (target variable)
 - Price Open - yesterday's Price Close (1 for an up movement, 0 for a down movement)
- Applied threshold to remove dates with small price change percentage (+/- 0.3%)
- Merged stock price data with tweets data (excl. market hours) ①
- Applied filter for tweets with only 1 ticker (our specific stock) ②
- Added sentiment label (pos/neutral/neg) from pre-trained roBERTa model from HuggingFace



Exploratory Data Analysis

Distribution of target variable “price_movement”, similar distribution across the 3 stocks close to a 60/40 ratio



Word predictability

- Created an identity matrix (each row has one word or feature)
- Basic Multinomial Naive Bayes model to predict each row
- Sorted rows by predicted probability and picked top/bottom words

[illegible]

Google - Top 3 good words

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

fire emoji: associated with good sentiment and indicates that the stock is on a good run in terms of performance.

margin: seems to be related to profit, operating and ebitda margins.

elliottwave: independent financial and social forecasting firm. 69% of these tweets come from their twitter account.



Google - Top 3 bad words

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

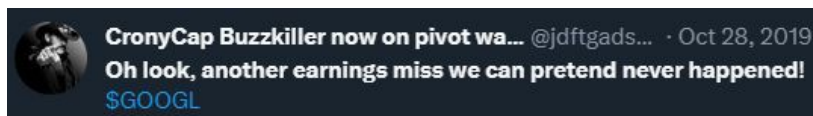
reuterstv: Around 70% of the tweets were on the period of May 10-12 2019 and are associated mainly with topics shown below.

*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*

miss: related to google missing revenue, earnings or other estimates.

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.



ExxonMobil - Top 3 good words

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

portfolio: Related to the stock being part of people's portfolio.

member: Refers to the company's board and changes in its members.

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, it does seem to appear more frequently in a positive light.



ExxonMobil - Top 3 bad words

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

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

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

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.

 **Tickeron** @Tickeron · Oct 23, 2021
\$XOM in Downtrend: its price expected to drop as it breaks its higher Bollinger Band on October 20, 2021. View odds for this and other indicators: srnk.us/go/3115007 #ExxonMobil #stockmarket #stock #technicalanalysis #money #trading #investing #daytrading #news #today

 **Arcady** @arcady_s · Dec 7, 2020
Replying to @andykarsner
I couldn't agree more. Having spent time with some of the other, more progressive, oil and gas majors, it's become abundantly clear how far behind \$XOM really is.

 **A Compounding Life** @a_compounding · Jun 23, 2020
Replying to @TheAlphaThought
This is how I view my \$XOM position

Dividends per year will soon be greater than paper capital loss

So not sweating it as cash flow is king

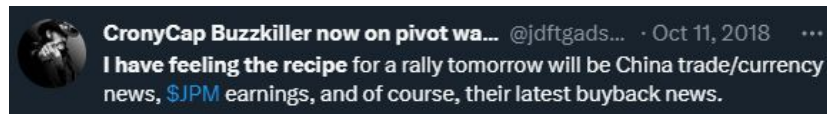
JPMorgan Chase - Top 3 good words

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

buyback: It appears to refer to the company's shares buyback.

quarterly: Related to changes in quarterly earnings or dividends and it seems to be positive.

dividend: similar to the last word, related to positive changes in dividends.



JPMorgan Chase - Top 3 bad words

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



release: seems to be talking about a reserve release.

reserve: same as above, seems to be talking about a reserve release.

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.

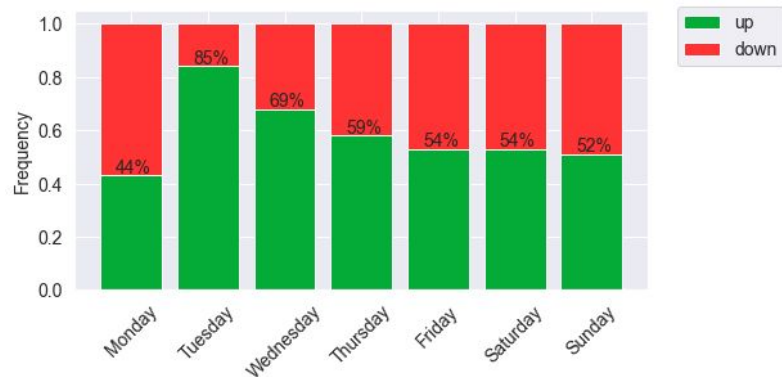
 **Wilfred Frost**  @WilfredFrost · Apr 14, 2021
\$JPM with a massive beat on the face of it - but drive by a far bigger reserve release than expected (\$5.2bn vs \$1.1bn expected). Dimon - "we do not consider [this] core or recurring profit". BUT - the release bigger than expected only because outlook is too.

 **Dogie Kass**  @DougKass · Oct 13, 2021
The EPS beat was materially from a \$2 billion reserve release. Ergo, a low quality beat at \$JPM.
[@tomkeene](#) [@lisaabramowicz1](#) [@ferrotrv](#)

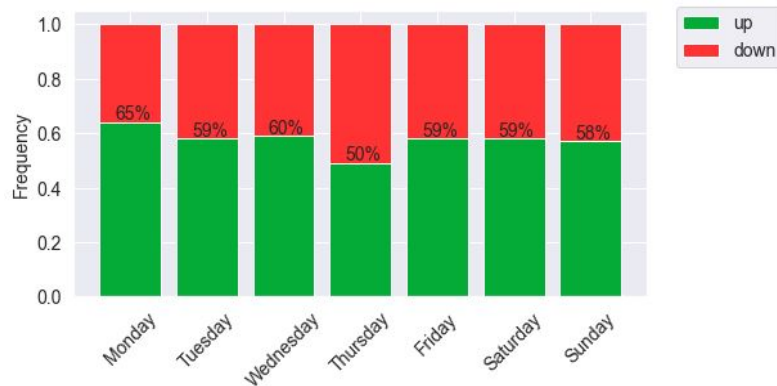
 **Dogie Kass**  @DougKass · Oct 13, 2021
[@tomkeene](#)
If \$JPM over reserved in the first half of 2020 (and under earned) and then release a large portion of those loan loss provisions in 3Q2021 it is a poor quality EPS beat.
[@lisaabramowicz1](#) [@ferrotrv](#)

Other tweet features - day of the week

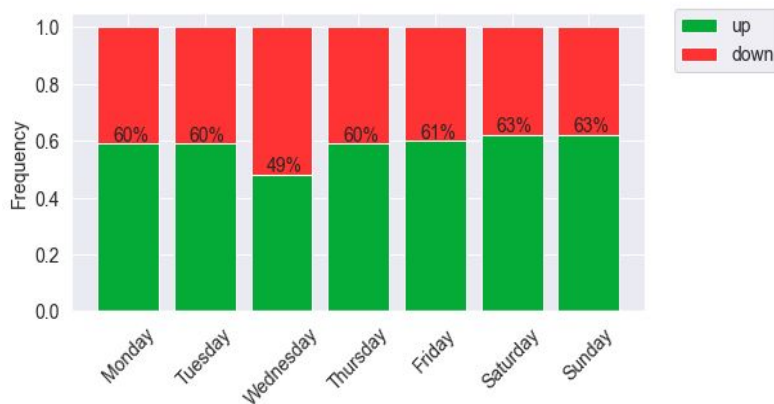
GOOGL - Price movement by day of the week



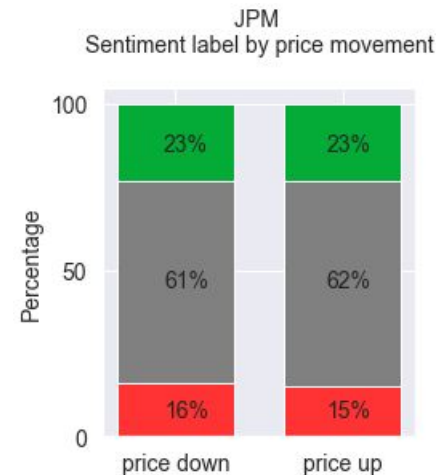
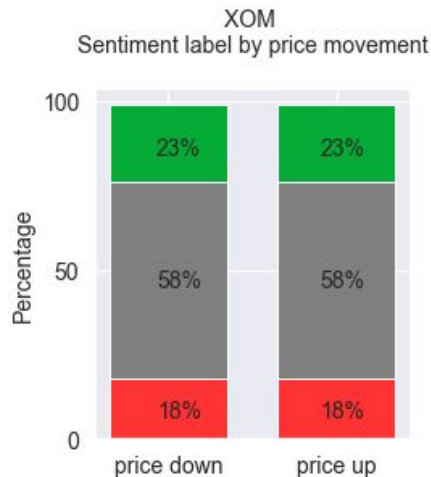
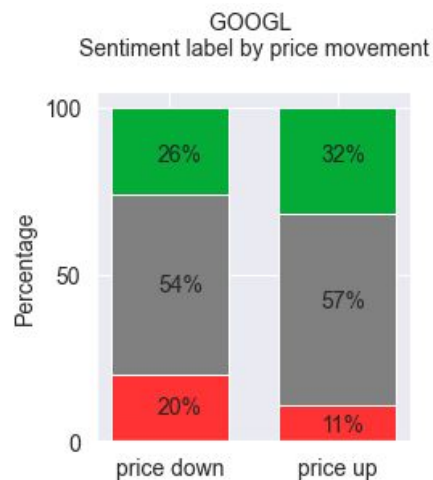
XOM - Price movement by day of the week



JPM - Price movement by day of the week



Other tweet features - sentiment label (roBERTa)



Preprocessing

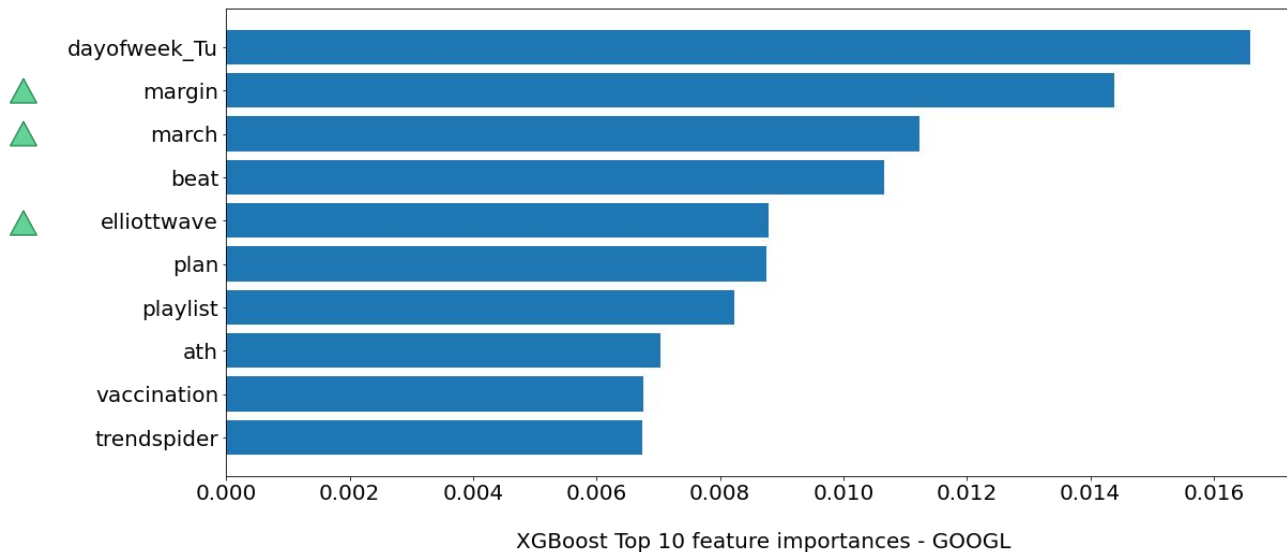
- One hot encoding for categorical features
- MinMax Scaler for numerical features
- Data was split in 70/30 ratio for train/test
- Count Vectorizer - token matrix (min_df of 10 and unigrams)

Modeling

- Individual models for each stock instead of one model for all stocks.

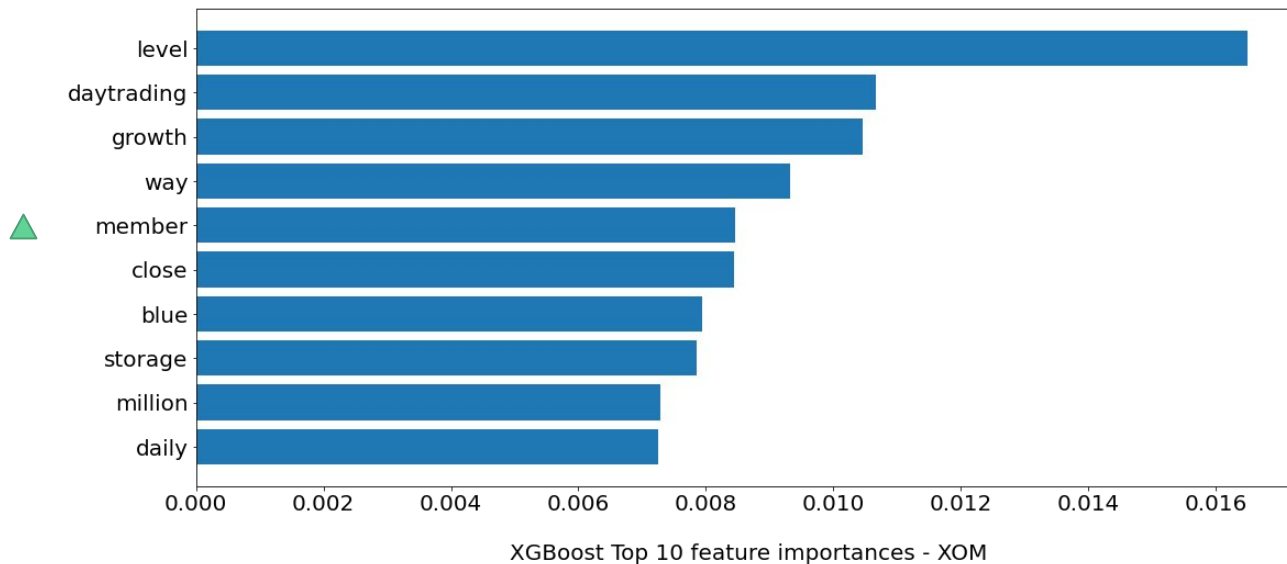
	ROC AUC Score		
	GOOGL	XOM	JPM
Multinomial NB	0.68	0.58	0.60
Random Forest	0.79	0.73	0.69
XGBoost	0.79	0.74	0.71

Google - XGBoost Top 10 feature importances



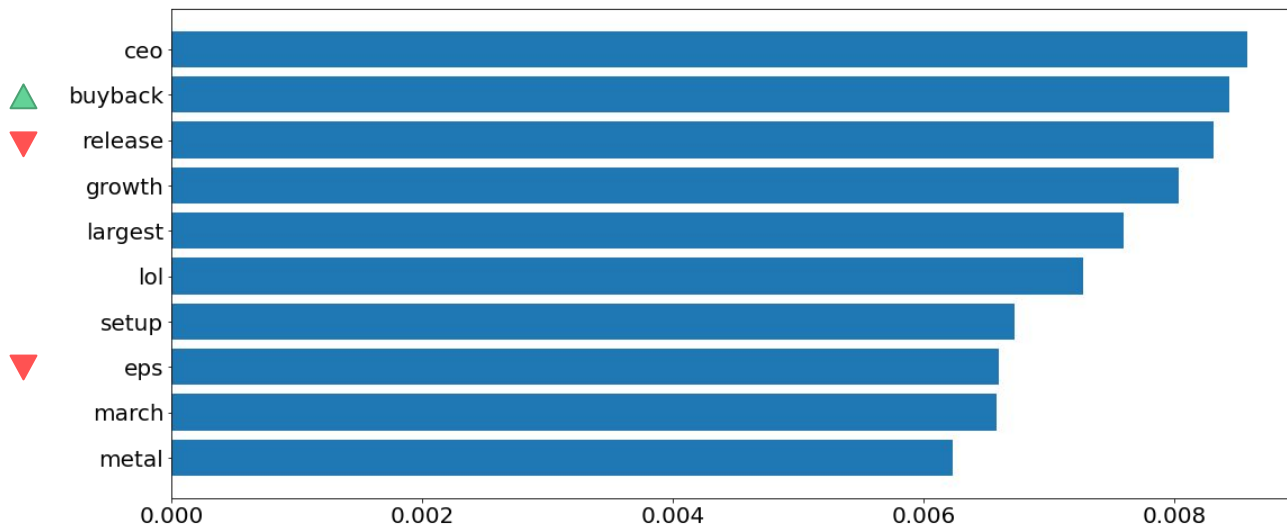
▲ Good predictability word

ExxonMobil - XGBoost Top 10 feature importances



▲ Good predictability word

JPMorgan Chase - XGBoost Top 10 feature importances



XGBoost Top 10 feature importances - JPM

- ▲ Good predictability word
- ▼ Bad predictability word

Metrics test set - after slight threshold adjustment

Google

Precision: 78%

Recall: 63%

Threshold adj: 0.52



ExxonMobil

Precision: 74%

Recall: 62%

Threshold adj: 0.54

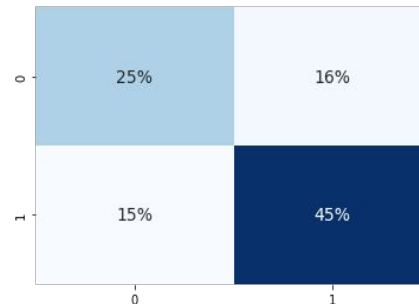


JPMorgan Chase

Precision: 74%

Recall: 61%

Threshold adj: 0.54



Conclusions

- Analysis showed there is good predictive potential from certain words used in tweets and also there seems to be a contrast across different stocks/sectors.
- Better performance on GOOGL model (more data available).
- GOOGL had a close similarity between the model's top features and the words identified during EDA as having good predictability.
- XGBoost models use more words from the tweets, but Random Forest models use other tweet features such as sentiment score, text length, hours and minutes.
- Final predictive models had precision 74% to 78% - models seem to be a bit overfitted (20% gap in precision between the train/test sets).
- Models can be used by retail investors as an additional metric/tool for their investment decisions.

Future work / Opportunities

- Expand to additional stocks within same or different sectors (i.e. Microsoft and Amazon very active on Twitter).
- Use different vectorizers (Tfidf) and lemmatizers (Spacy)
- Different price movement thresholds or no thresholds
- Not remove numbers to retain interpretability while reviewing tweets with important financial indicators (earnings, eps, dividends)
- Word predictability with XGBoost to get consistency between EDA and model's top features

Thank you!

Github Repo:

https://github.com/jduran3/Stock_price_movement_prediction_based_on_tweets