PREDICTING STOCK PRICE CHANGES OF HEALTHCARE COMPANIES BASED ON NEWS HEADLINES

By Lee Shi Min
DSI-23 | An NLP project

01 OBJECTIVES

Background
Focus on Healthcare
Problem Statement

02 METHODOLOGY

Data

03 RESULTS ANALYSIS

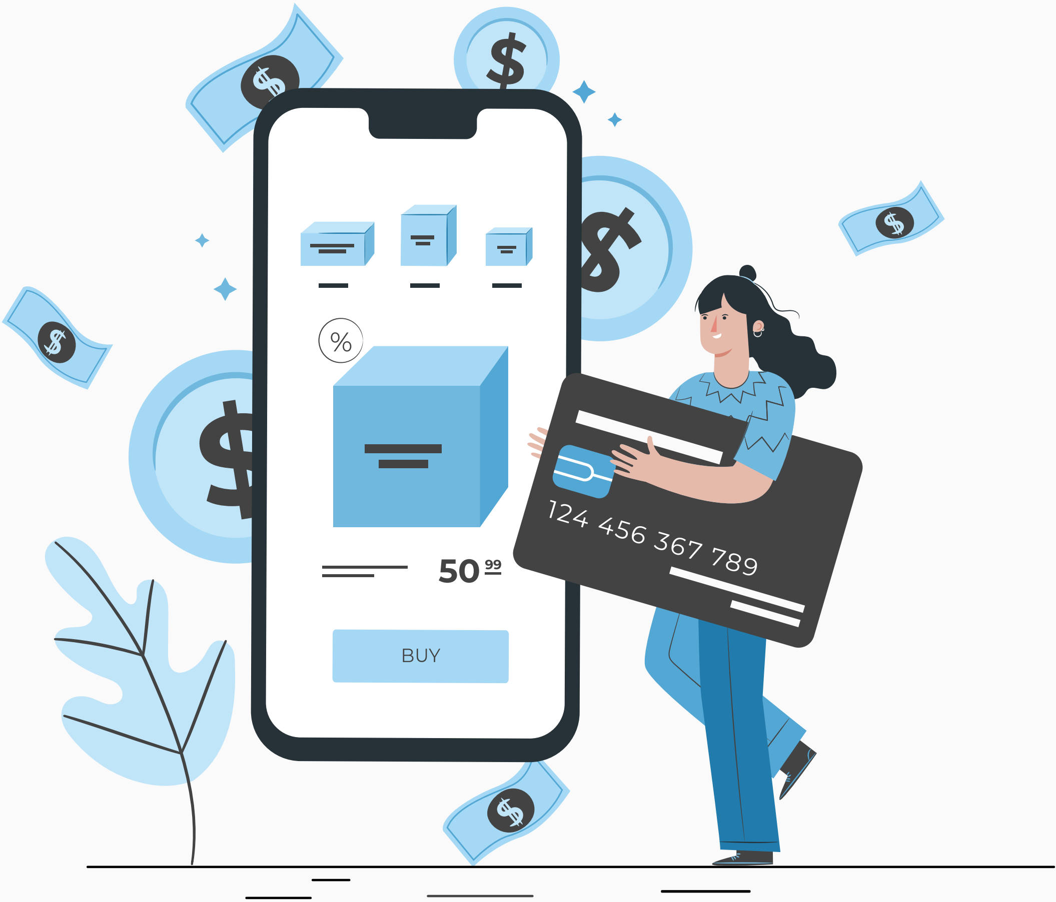
Model Performance
Other Findings

04 CONCLUSIONS

Learning Points
Recommendations for Future Project

PREDICTING STOCK PRICES IS HARD...

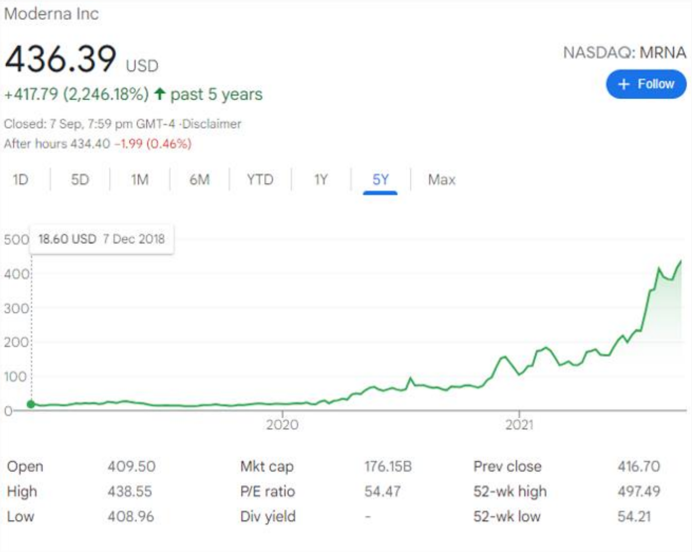
- Technical versus fundamental analysis
- Evidence of post-news drift
- "Buy the rumor, sell the news"

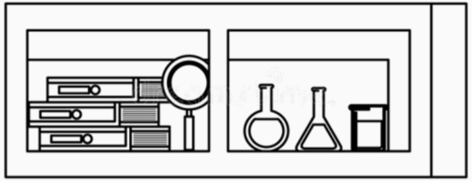


KEY QUESTIONS

1. Can news headlines be relied on to predict stock price movements?
2. Does the release of positive/negative news result in a respective increase/ decrease in stock price?

FOCUS ON HEALTHCARE COMPANIES





Tremendous Growth & Potential

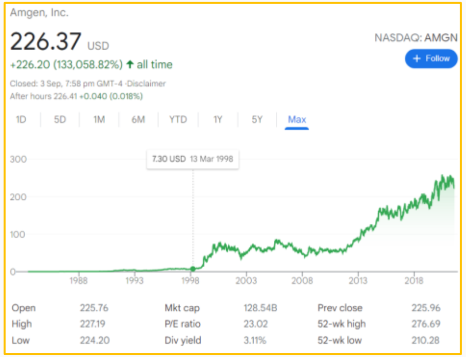
- Development of vaccines and therapeutics highly valued by society (Moderna stock increased by > 2,000% in the last 5 years)
- Outperforming S&P500 in some instances

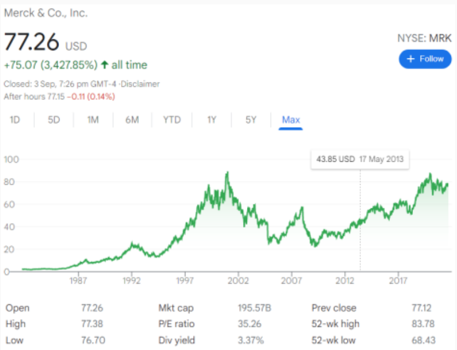
Knowledge Gap

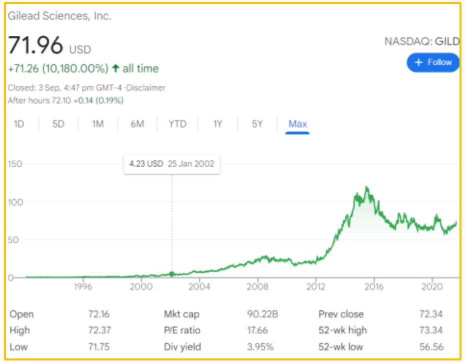
- Finance knowledge is not enough
- Understanding of drug development process, commercial implications of scientific technologies and regulatory frameworks needed

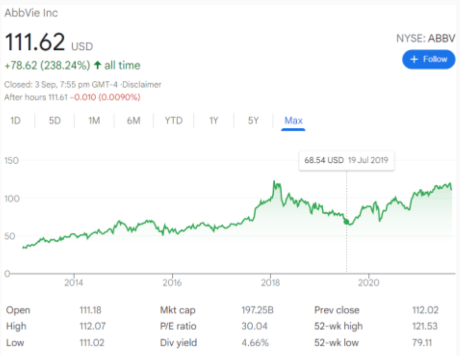
FOCUS ON HEALTHCARE COMPANIES

6 out of 8 companies in list outperformed S&P 500 (~3,616% all time increase)









PROJECT FOCUS

Using NLP/Data Science to help layperson or amateur investors learn about healthcare investing

1. NLP Project to Predict Stock Price Jumps given News Headlines	
1. Use past data for prediction	**ROC AUC**
2. Accurate identification of price jumps on news	**Accuracy**
3. Precise prediction of price jumps	**Precision**

2. Deep dive into the headlines that are indicative of price jumps
Study feature importances

DATA

REUTERS®

- **Date**
- News headlines

CSV

ALPHA VANTAGE

- **Date**
- Adjusted Closing Price (Dividends, Stock Splits, S&P500)
- Trading Volume

Feature	Description
Date	Date during which the stock and news information was pulled.
adjusted_abs	**Absolute percentage change** for the adjusted closing price of the stock with respect to the previous trading day.
adj_direction	Denote the directionality of the stock price changes.
cleaned	Daily news headlines from Reuters pertaining to the particular ticker concatenated into one single string.
headline_word	Word count of news headlines
dict_score	Score to denote the news sentiment for healthcare domain. Negative values are assigned to bad news (e.g. failed trial) and positive values are assigned to good news (e.g. fda approval).
target_var	Classification label: **1** denotes when stock prices changed by more than 1.31% (P75), **0** for otherwise.

WORKFLOW

Data Cleaning

- Removal of outliers
- Text cleaning – define new stopwords, removal of stopwords, lemmatization

Feature Engg & EDA

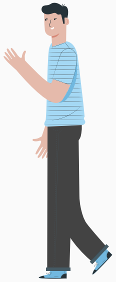
- Word Count for News Headlines
- Polarity Scores
- Vectorization – Count & Tf-idf

Choice of Model

- Using PyCaret to compare models
- Choose model based on precision, accuracy & AUC ROC scores

Model Tuning & Evaluation

- GridSearchCV
- Confusion Matrix, AUC ROC Curve



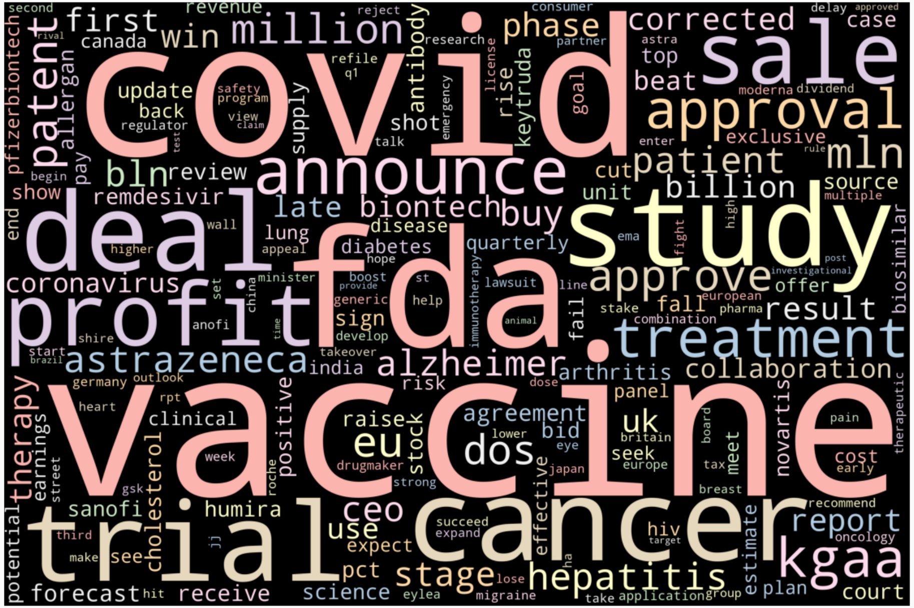
TEXT CLEANING - AMENDING THE LIST OF STOPWORDS

```
#add words that aren't in the NLTK stopwords list
new_stopwords = ['reuters', 'country', 'population', 'government', 'united', 'states', 'company', 'economy',
                 "said", 'say', 'inc', 'data', 'business', 'one', 'two', 'three', 'four', 'five',
                 'january', 'february', 'march', 'april', 'may', 'june', 'july', 'august', 'nyse',
                 'september', 'october', 'november', 'december', 'weekly', 'monthly', 'quarter', 'year',
                 'abbvie', 'pfizer', 'gilead', 'merck', 'eli', 'lilly', 'amgen', 'biogen', 'regeneron',
                 'bristol', 'myers', 'squibb', 'nasdaq', 'new', 'york', 'exchange', 'earning', 'price',
                 'medicine', 'healthcare', 'dow', 'jones', 'index', 'dji', 'investor', 'percent', 'market',
                 'drug', 'share', 'health', 'biotech', 'per', 'cent', 'co', 'sp']

new_stopwords_list = stop_words.union(new_stopwords)

#remove words that are in NLTK stopwords list
not_stopwords = {'no', 'not', 'up', 'further', 'above', 'down', 'under', 'over', 'through', 'off', 'below'}
final_stop_words = set([word for word in new_stopwords_list if word not in not_stopwords])
```

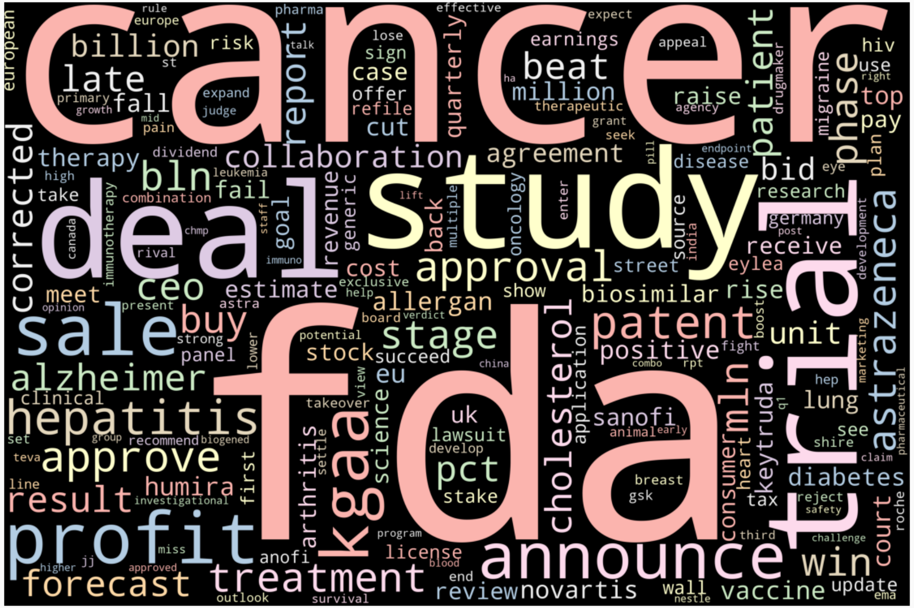
REMOVAL OF OUTLIERS BY YEAR (COVID ADJUSTMENT)

BEFORE: 2012 to 2021
About a decade's worth of news heavily biased due to high frequency reporting on recent COVID-19 developments

BEFORE: 2012 to 2019
Remove news headlines dated to 2020 and 2021 to avoid having too much noise regarding COVID/vaccination situation

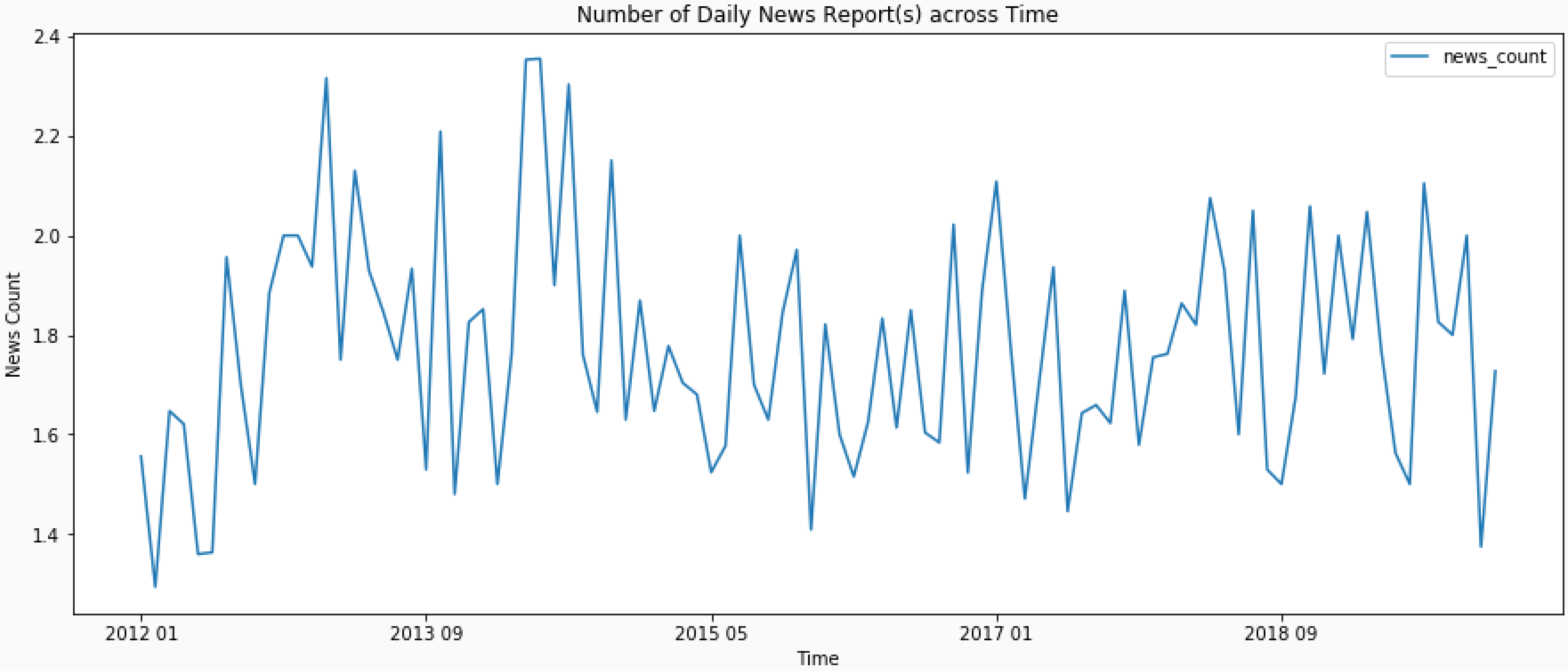


Word Cloud for Biotech News Headlines



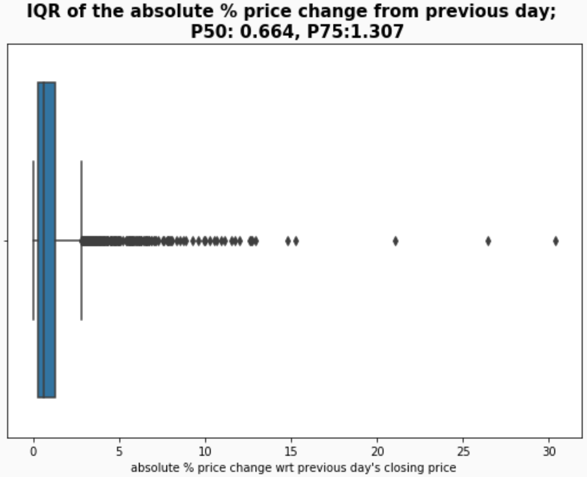
Word Cloud for Biotech News Headlines

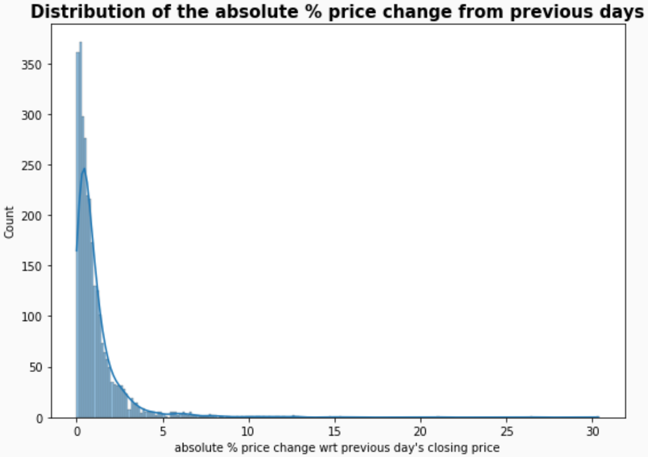
REMOVAL OF OUTLIERS - NUMBER OF NEWS > P99*



* Days with number of news report > P99 (i.e. 7) for a particular company are also removed

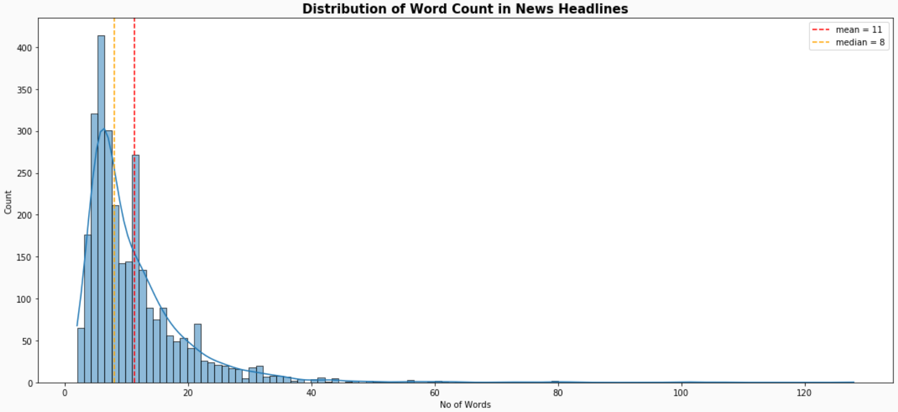
DISTRIBUTION OF WORD COUNT AND PRICE CHANGE VARIABLES





% Price Change

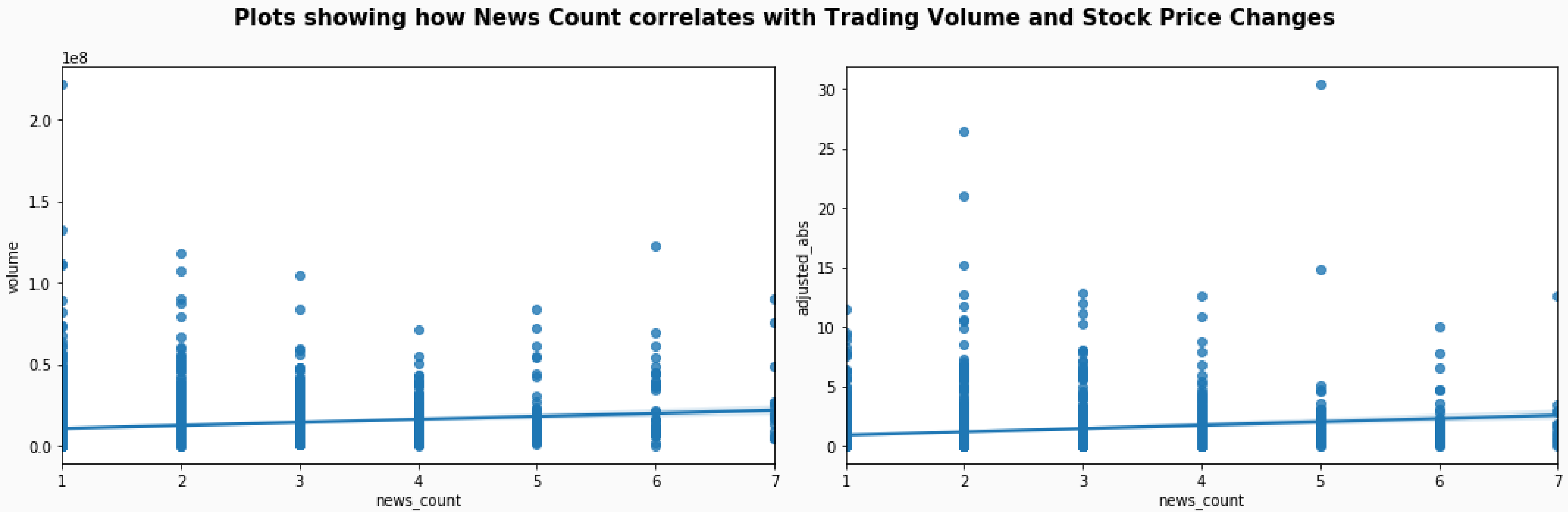
- Right skewed distribution with a long tail
- **P75** to be referenced as **threshold** of 'price jump' for target variable



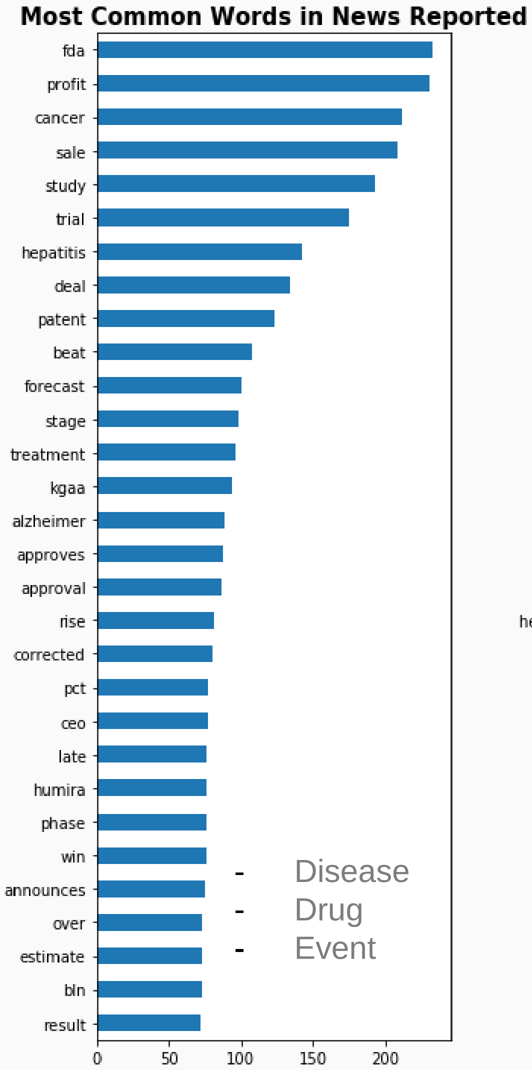
Word Count

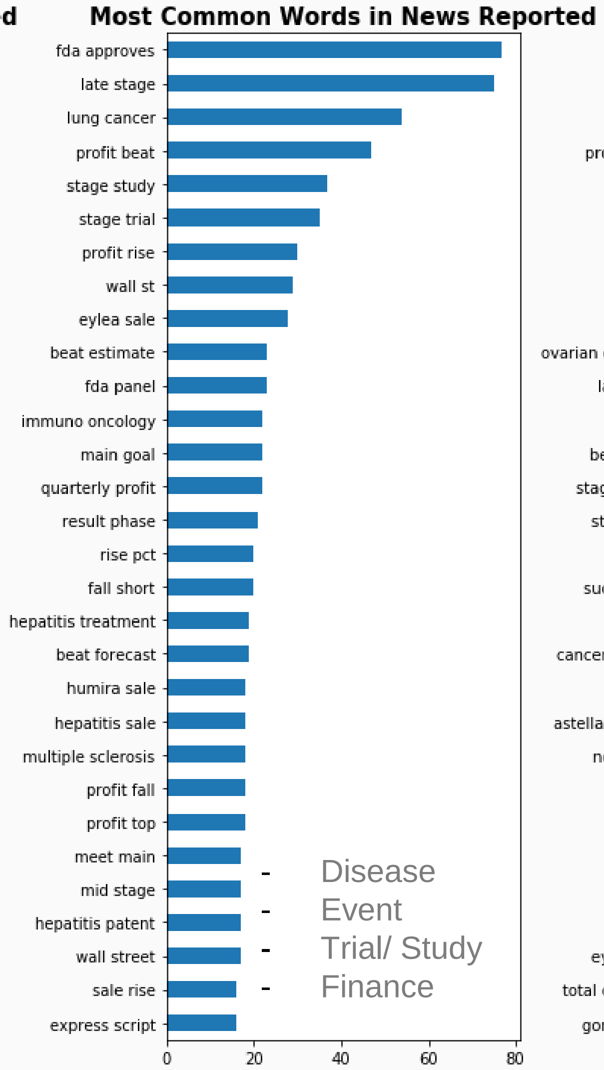
- Right skewed distribution with a long tail
- Mean = 11, median = 8 words; question of sufficiency of text info for NLP analysis

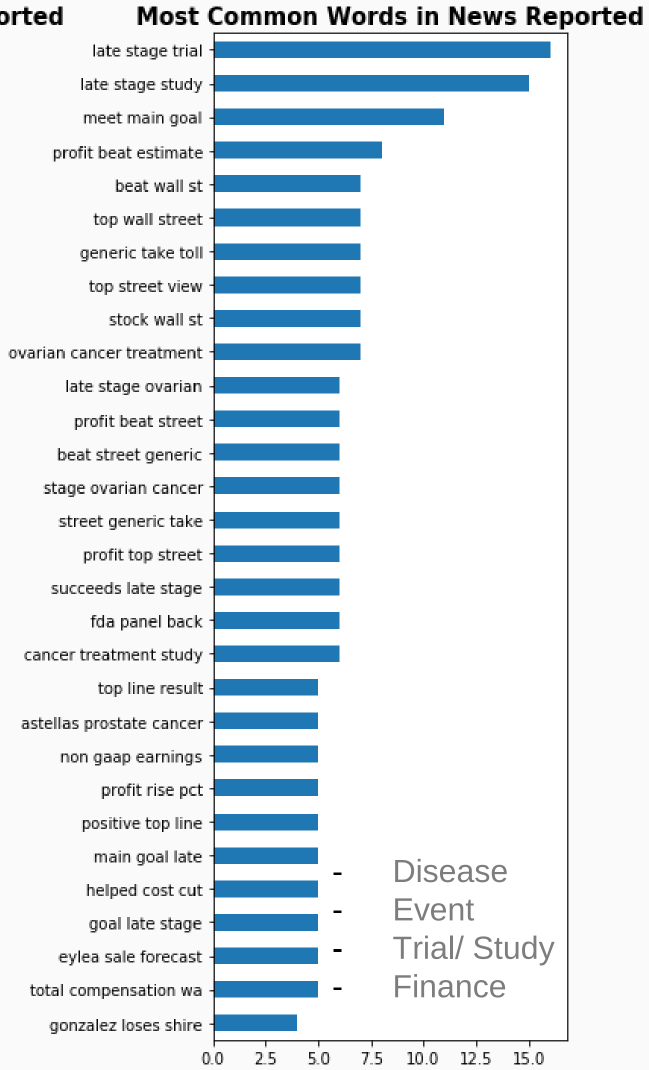
MORE NEWS REPORT SEEM MILDLY ASSOCIATED WITH HIGHER TRADING VOLUME AND STOCK PRICE CHANGES

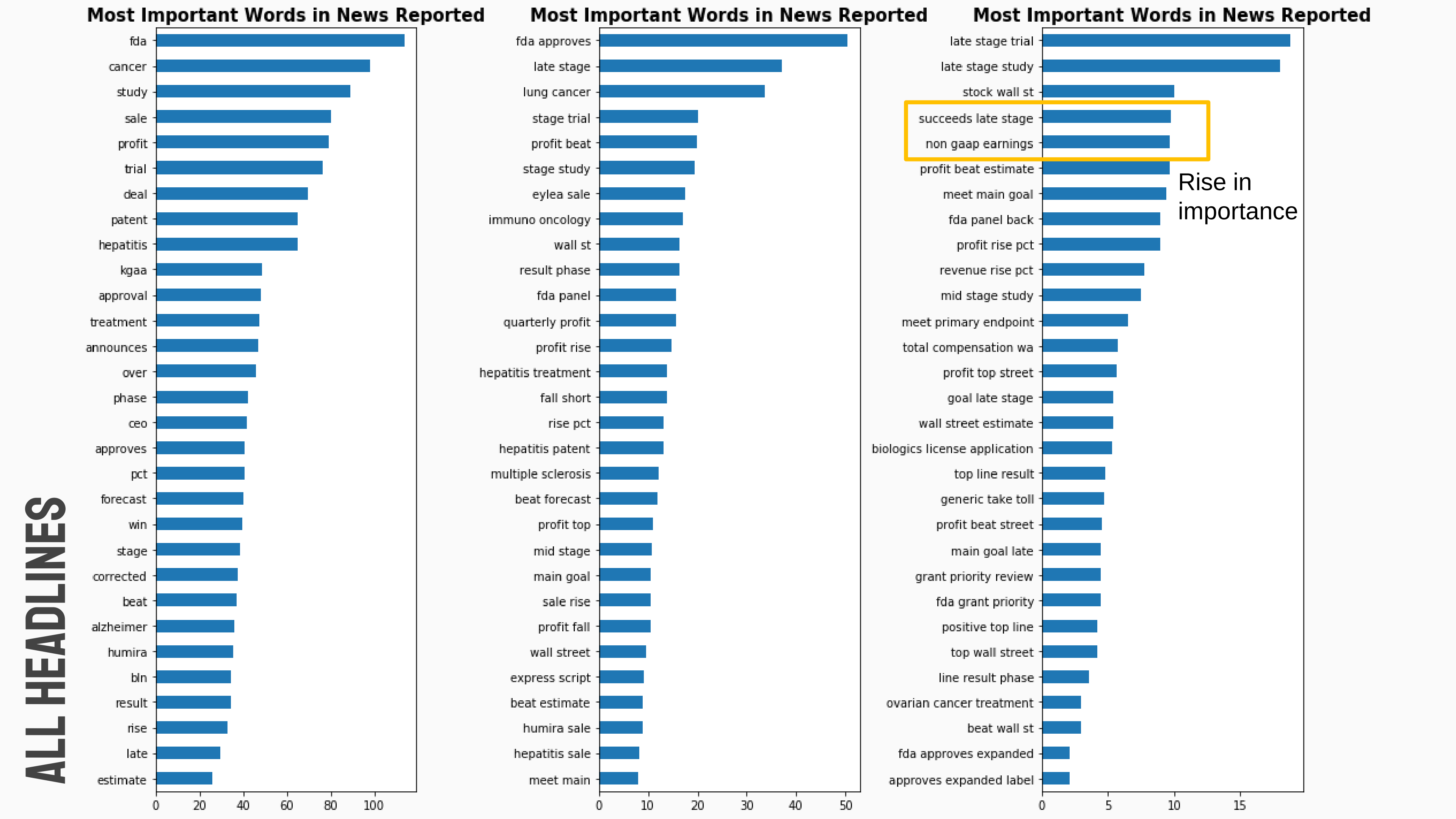


ALL HEADLINES



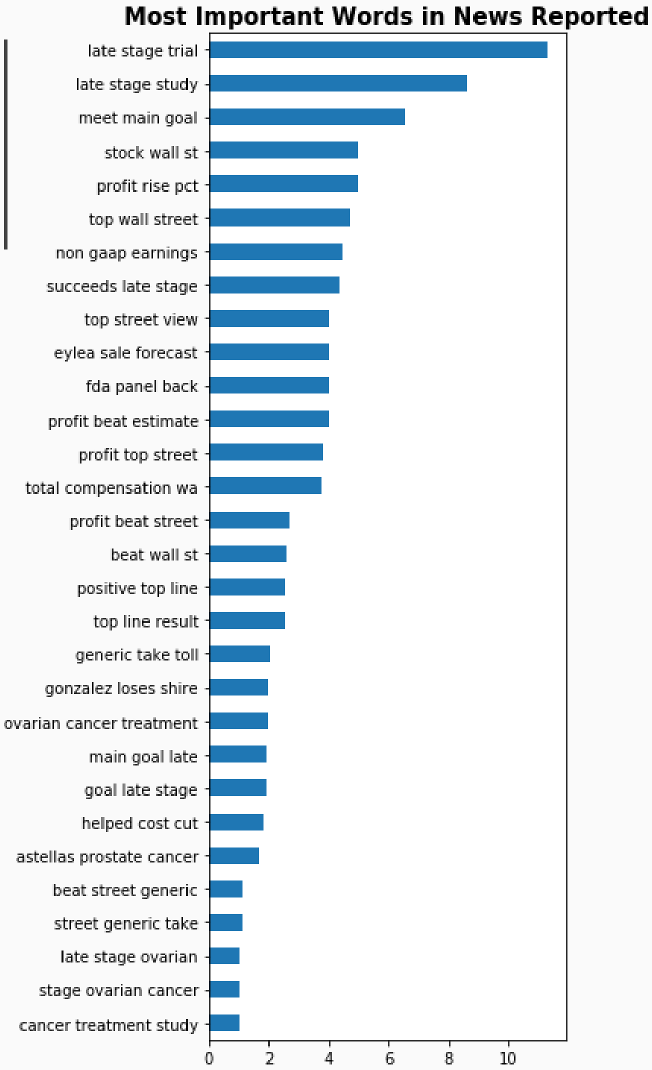




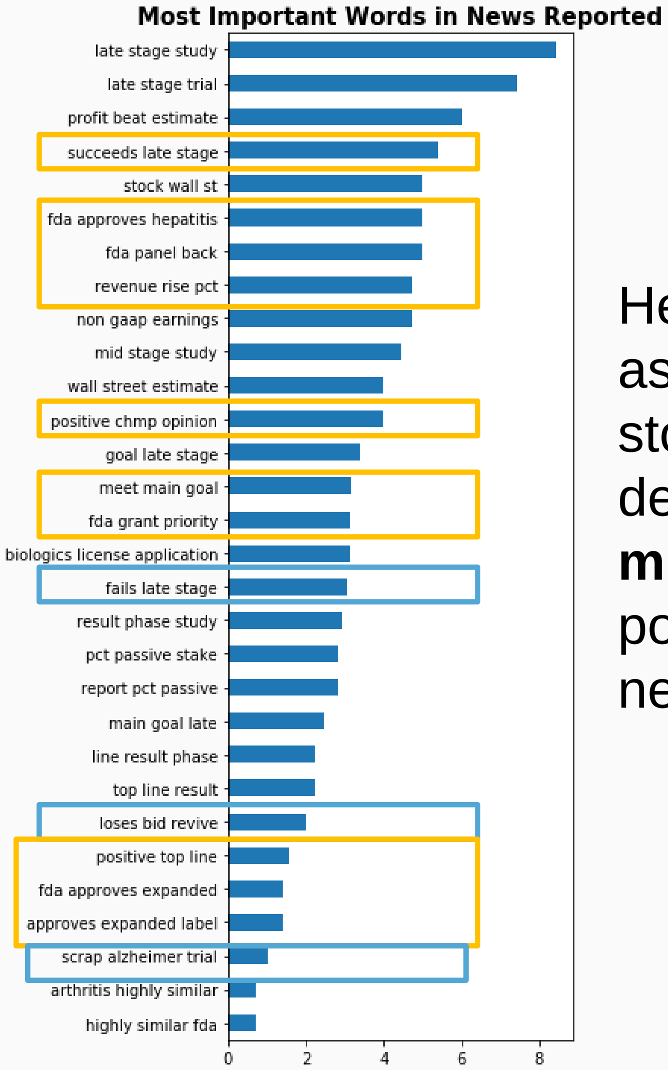


ALL HEADLINES
Most Important Words in News Reported
fda
cancer
study
sale
profit
trial
deal
patent
hepatitis
kgaa
approval
treatment
announces
over
phase
ceo
approves
pct
forecast
win
stage
corrected
beat
alzheimer
humira
bln
result
rise
late
estimate
0
20
40
60
80
100
Most Important Words in News Reported
fda approves
late stage
lung cancer
stage trial
profit beat
stage study
eylea sale
immuno oncology
wall st
result phase
fda panel
quarterly profit
profit rise
hepatitis treatment
fall short
rise pct
hepatitis patent
multiple sclerosis
beat forecast
profit top
mid stage
main goal
sale rise
profit fall
wall street
express script
beat estimate
humira sale
hepatitis sale
meet main
0
10
20
30
40
50
Most Important Words in News Reported
late stage trial
late stage study
stock wall st
succeeds late stage
non gaap earnings
profit beat estimate
meet main goal
fda panel back
profit rise pct
revenue rise pct
mid stage study
meet primary endpoint
total compensation wa
profit top street
goal late stage
wall street estimate
biologics license application
top line result
generic take toll
profit beat street
main goal late
grant priority review
fda grant priority
positive top line
top wall street
line result phase
ovarian cancer treatment
beat wall st
fda approves expanded
approves expanded label
0
5
10
15
Rise in importance

HEADLINES ASSOCIATED WITH PRICE INCREASE

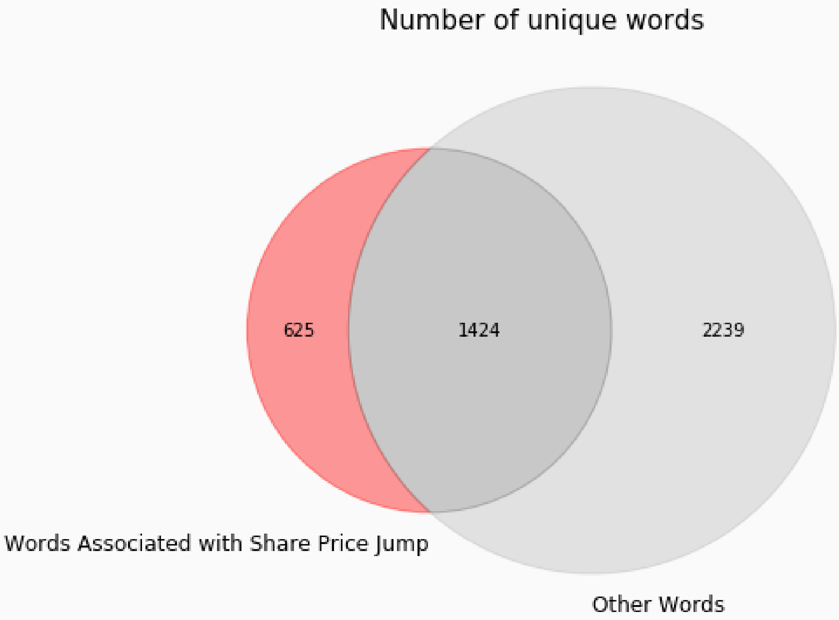


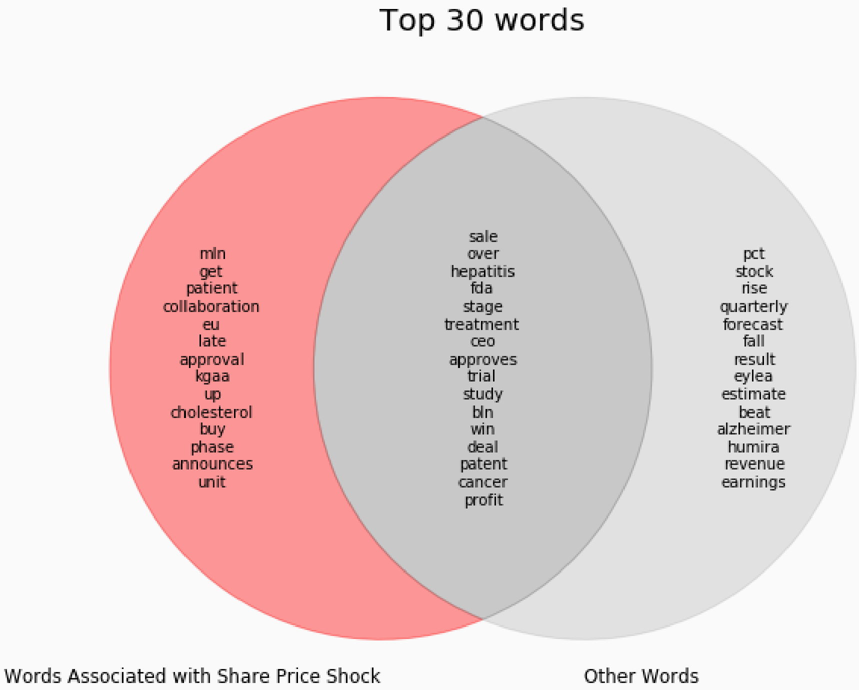
HEADLINES ASSOCIATED WITH PRICE DECREASE



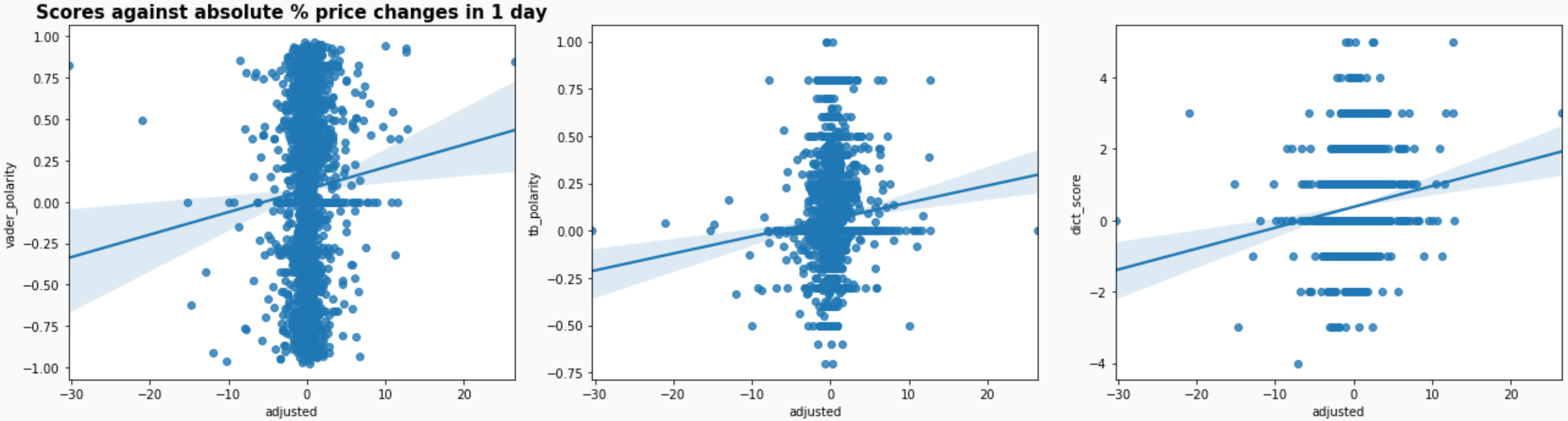
Headlines associated with stock price decrease are a **mixed bag** of positive and negative events

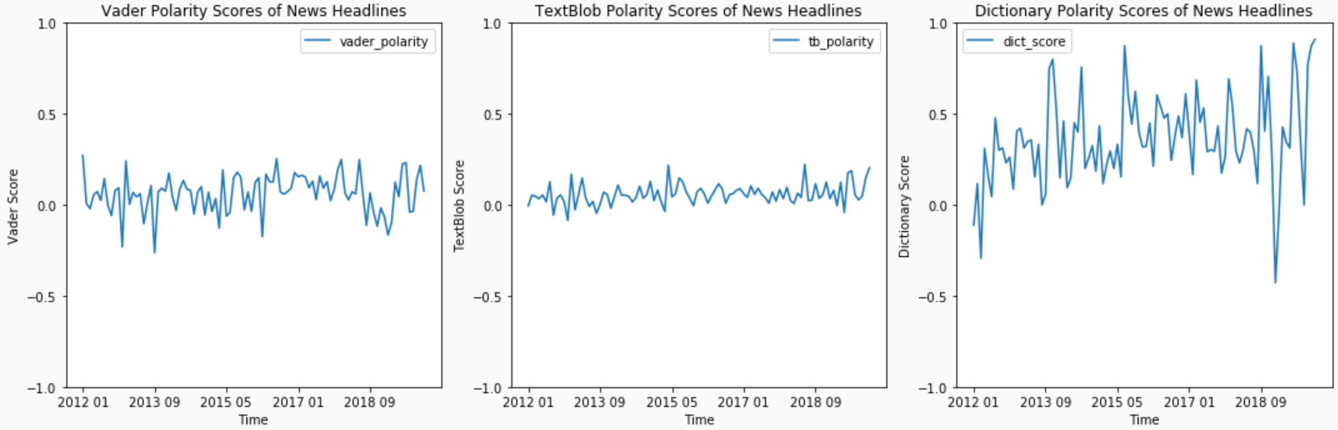
FEWER UNIQUE WORDS ARE ASSOCIATED WITH SHARE PRICE JUMP, BUT THERE ARE NO DISCERNABLE PATTERNS OR WORDS THAT SEEM INDICATIVE OF SHARE PRICE JUMP



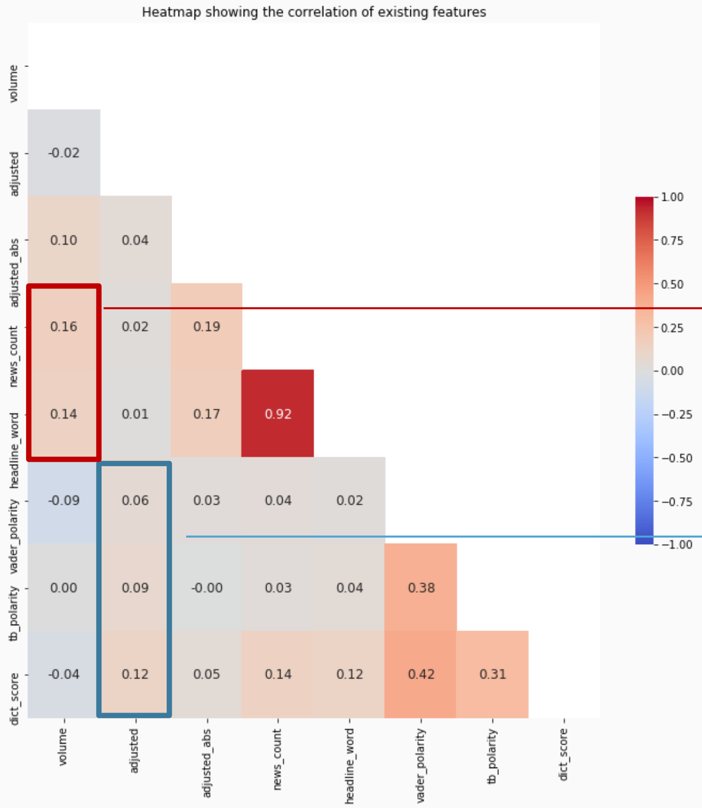


SCATTER PLOT OF SENTIMENT SCORES SUGGEST EXISTING LIBRARIES/FUNCTIONS (E.G. VADER & TEXTBLOB) ARE NOT EFFECTIVE AT DIFFERENTIATING NEWS HEADLINES





WEAK POSITIVE CORRELATION BETWEEN NEWS AND TRADING VOLUME, AND DICTIONARY SCORE AND ADJUSTED STOCK PRICE CHANGES IN %



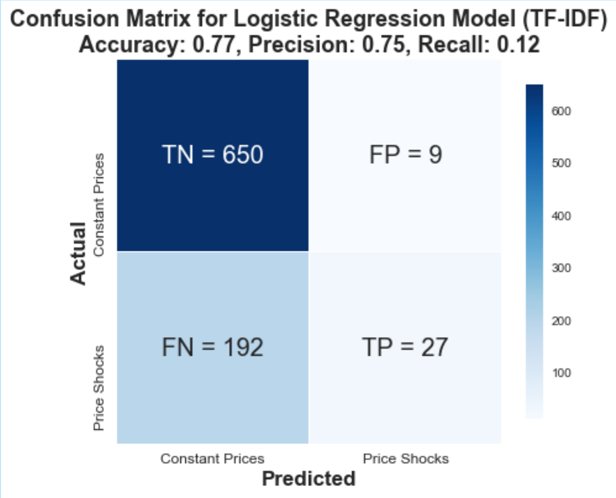
Higher news count weakly & positively correlated with higher trading volumes

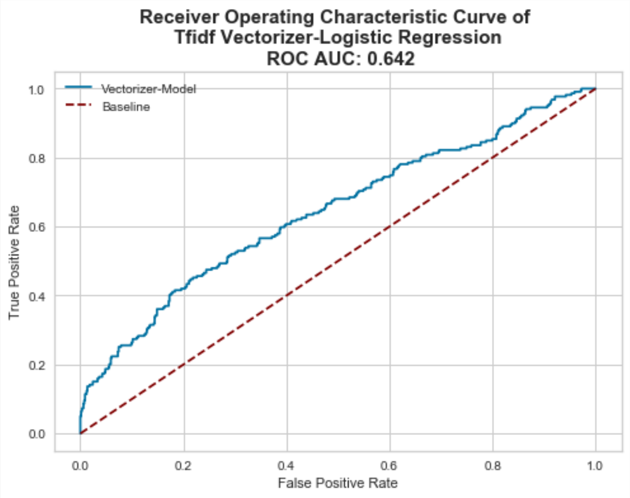
The manually determined dictionary score shows a higher correlation with the percentage change in stock prices than the vader and TextBlob polarity scores

MODEL SELECTION BASED ON RESULTS ARISING FROM PYCARET'S MODEL COMPARISON FUNCTION

S/N	Model	Vectorizer	N-Gram	Accuracy	AUC	Recall	Precision	F1
1	Light Gradient Boosting Machine	Count	1-gram	0.764	0.625	0.248	0.549	0.339
2	Logistic Regression	Count	1-gram	0.764	0.651	0.307	0.529	0.386
3	Gradient Boosting Classifier	Count	2-gram	0.747	0.560	0.068	0.663	0.120
4	Logistic Regression	Count	2-gram	0.745	0.594	0.165	0.516	0.247
5	Gradient Boosting Classifier	Count	3-gram	0.749	0.522	0.046	0.496	0.084
6	Logistic Regression	Count	3-gram	0.747	0.538	0.060	0.493	0.106
7	**Logistic Regression**	**Tf-idf**	**1-gram**	**0.758**	**0.654**	**0.117**	**0.672**	**0.199**
8	K Neighbors Classifier	Tf-idf	1-gram	0.748	0.568	0.114	0.577	0.187
9	Logistic Regression	Tf-idf	2-gram	0.766	0.600	0.043	0.636	0.079
10	Gradient Boosting Classifier	Tf-idf	2-gram	0.764	0.600	0.080	0.524	0.135
11	Logistic Regression	Tf-idf	3-gram	0.748	0.570	0.006	0.150	0.011
12	Ada Boosting Classifier	Tf-idf	3-gram	0.748	0.533	0.066	0.511	0.116

RESULTS OF A ONE-GRAM TD-IDF LOGISTIC REGRESSION

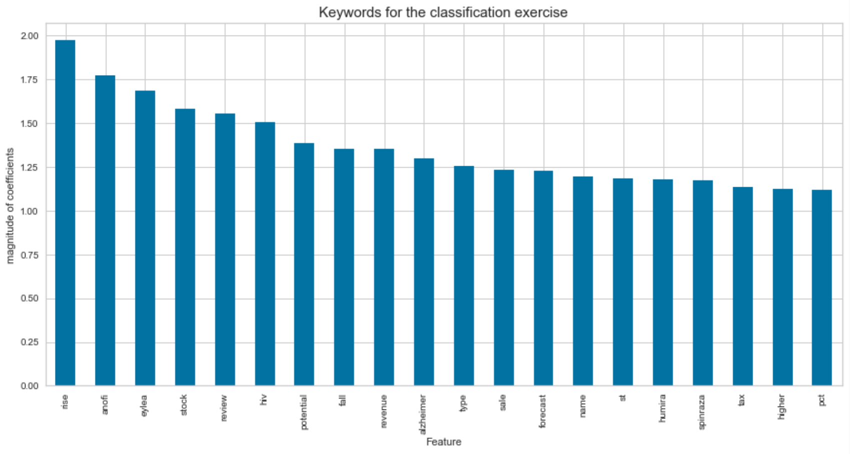




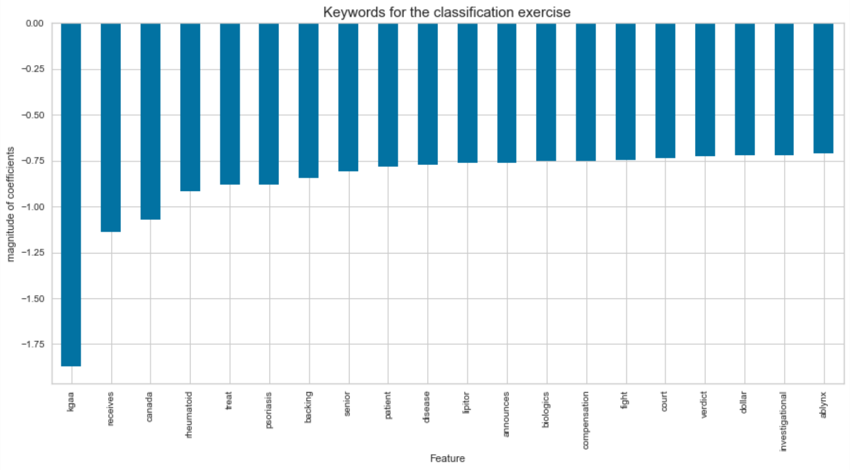
The model has an accuracy score of 0.77, a ROC AUC score of 0.642 and a 0.75 precision score. While the precision score is high, it is obtained at the expense of a lower recall score of 0.12. The **model will be relatively good at predicting when a news headline would likely be a price jumper**. **However, the model may misclassify other news which are 'price jumper' as 'non-price jumper'.** This may lead to traders missing out on a significant news events that would affect stock prices significantly.

FEATURE IMPORTANCES – SINGLE WORD

Words, e.g. ‘**rise**’, ‘**stock**’, ‘**hiv**’, ‘**potential**’, ‘**fall**’, are predictive of stock price jumps



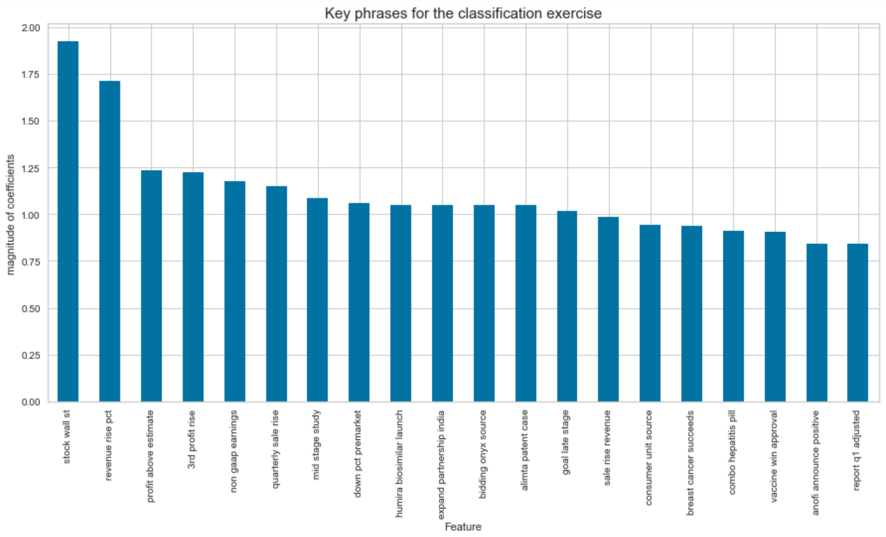
Words, e.g. ‘**kgaa**’, ‘**receives**’, ‘**canada**’, ‘**rheumatoid**’, are associated with constant stock prices

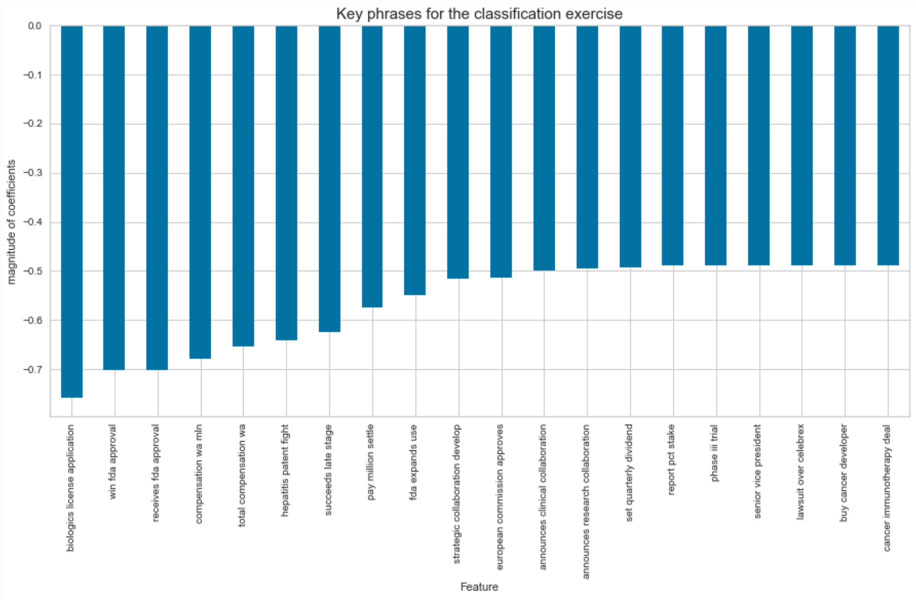


FEATURE IMPORTANCES – THREE-WORD PHRASES

Phrases, e.g. **'revenue rise pct'**, **'profit above estimate'**, **'quarterly sale rise'**, **'mid stage study'**, are predictive of stock price jumps

Phrases, e.g. **'win fda approval'**, **'receives fda approval'**, **'succeed late stage'**, are associated with **constant** stock prices

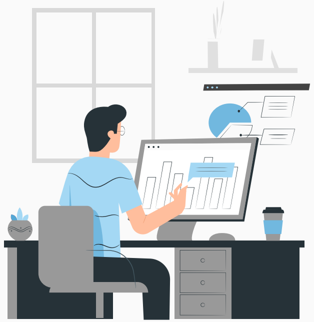




CONCLUSION & RECOMMENDATIONS

1. **News headlines can be used to predict changes to stock prices.** The tf-idf logistic regression model is able to achieve > 70% accuracy and precision scores, and > 60% AUC ROC scores.

2. **The release of positive/ negative news may not necessarily lead to a jump in stock prices.** From this modelling experience, we learn that the announcement of significant milestones (such as FDA approval) may not lead to significant stock price changes.



Considerations for Future Project(s):

a. Aggregate news headlines from more sources (e.g. Financial Times, Seeking Alpha) for analysis
b. Study how the long-form news articles affect stock prices
c. Aggregate news headlines for more healthcare, pharmaceutical or biotech companies for analysis

THANK YOU

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Storyset**

Please keep this slide for attribution.