Prediction of share price movement based on Earnings Call transcripts

Rita Palfine Tolgyesi

December 2020

Agenda

- Introduction and problem statement
- Baseline vs outcomes
- How I got there
- Conclusions and Lookout

Introduction of topic and problem statement

- Problem Statement

  Can we predict movement of share prices based on how the managers talk on the quarterly earnings call? (Without actually looking at the financial results)

- Hypothesis:

  Above average use of negative words predicts below average share price reaction from day of call to day of next call.

- Hypothesis revised:

  Manager's word choices reflect the company's performance

- Underlying assumption

  The way management talks about the company reflects how the company actually is doing, and gives a better understanding to investors as opposed to financial statements.

Baseline vs outcomes

Baseline	Fraction
DOWN	0.527
UP	0.473

Acc 0.634	Pred D	Pred U
True D	1238	949
True U	567	1393

Acc 0.60	Pred D	Pred U
True D	1440	895
True U	762	1079

	Test results	Roc auc	Accuracy	Precision	Recall
1	Logistic Regression on engineered features	0.675	0.62	0.64	0.62
2	Random Forest – cvec on all words in text, limited by frequency	0.645	0.59	0.58	0.81
3	Random Forest – tvec on all words in text, limited by frequency	0.640	0.59	0.58	0.81
4	Logistic Regression – cvec on all words in text, limited by frequency	0.627	0.59	0.59	0.69
5	Logistic Regression – tvec on all words in text, limited by frequency	0.622	0.58	0.58	0.74
6	Random search of Random Forest – all words, not limited	0.616	0.58	0.56	0.88

Average or desirable hit ratio in Finance

'The average **hit rate** for the portfolios in the Inalytics database was **49.6%**' [1]

'However numerous professionals and empirical experience on different type of portfolios is hinting to fact that **50% is rather a good Hit Ratio**. Very limited number of skilled investment professional are able to consistently maintain this level.' [2]

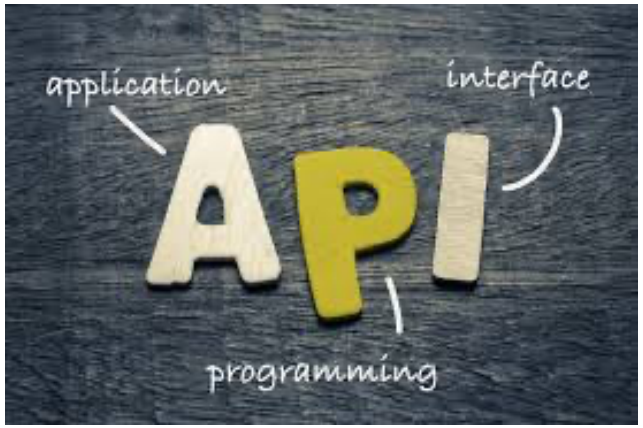
	Base Case "very good" theoretical Fund	What's needed for "Buffett"
Fund Under Management	100%	100%
Number of stocks	50	50
Individual Stock Weighting	2%	2%
Hit Rate	55%	60%
Loss Rate	45%	40%
Number of winning stocks	28	30
Number of Losing Stocks	23	20
Alpha per winner (excess return)	20%	25%
Alpha per loser	-15%	-15%
Winners Alpha %	11%	15%
Losers Alpha %	-7%	-6%
Total Alpha pre fees (%)	4.3%	9.0%

[1] https://thehedgefundjournal.com/identifying-manager-skill/

[2] https://www.linkedin.com/pulse/high-investment-hit-rate-too-good-true-ali-chabaane

https://morphicasset.com/what-batting-averages-can-tell-you-about-funds-management/

How I got there



Scraping Earnings Call Transcripts of US listed companies using Beautiful Soup from

https://www.fool.com/earnings-call-transcripts/

21818 transcripts scraped – covering 3 years

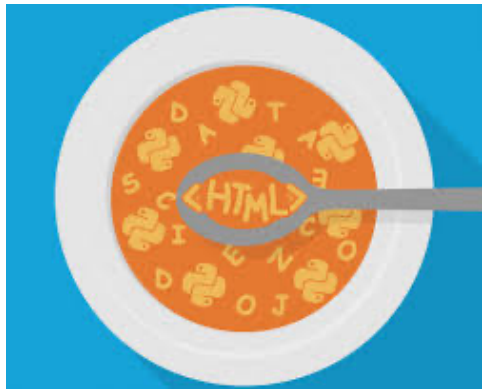
Share prices collected through Alpha Vantage AP

Sentiment word collection from

https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists

2355 negative words

354 positive words



Combining scrapes

Removing duplicates

Extracting details from scrapes

Merging transcripts with share price data – twice

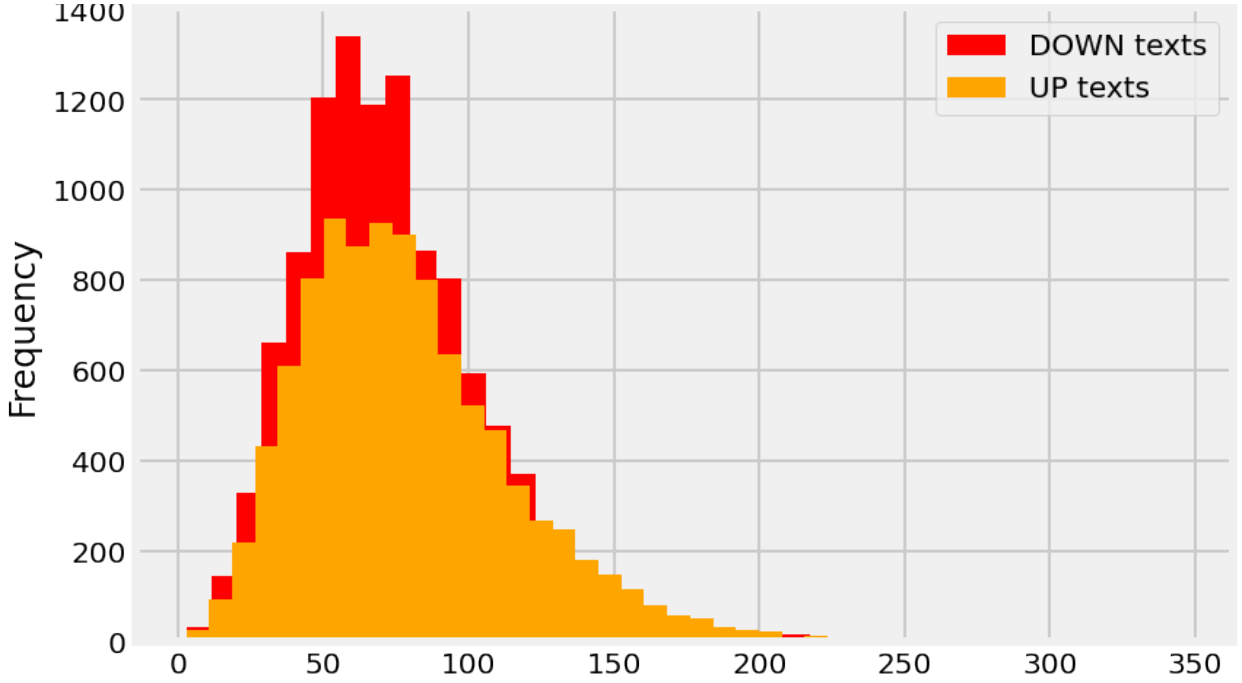
Labeling texts: "DOWN" and "UP"
(call date – next period end)

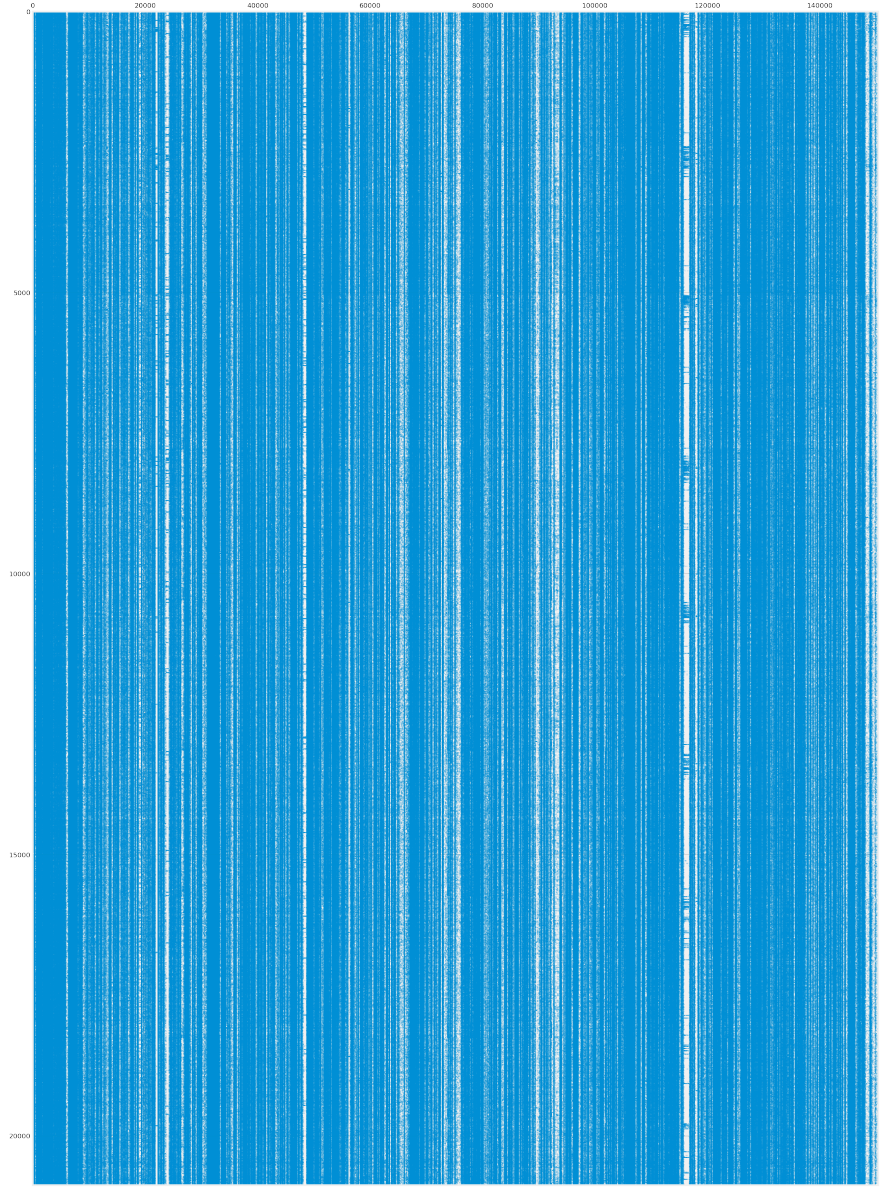
- Not all pages are organized in the same way just because the first few appears so
- Not all pages contain all information you thought they do

Finalizing data for modelling
20877 observations in dataset


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
DIRTY DATA

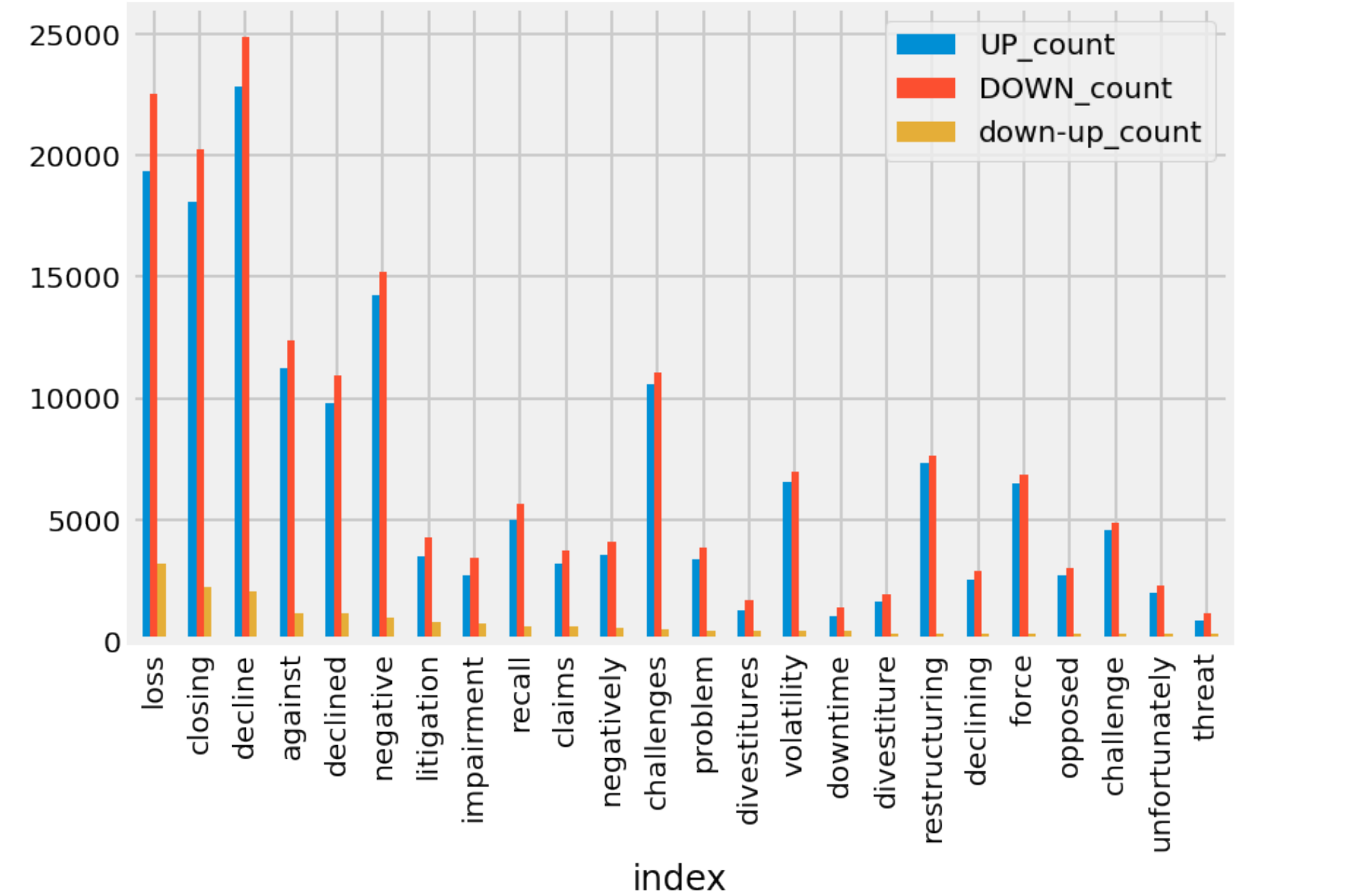
Histogram of total negative word frequency in “UP and “DOWN texts

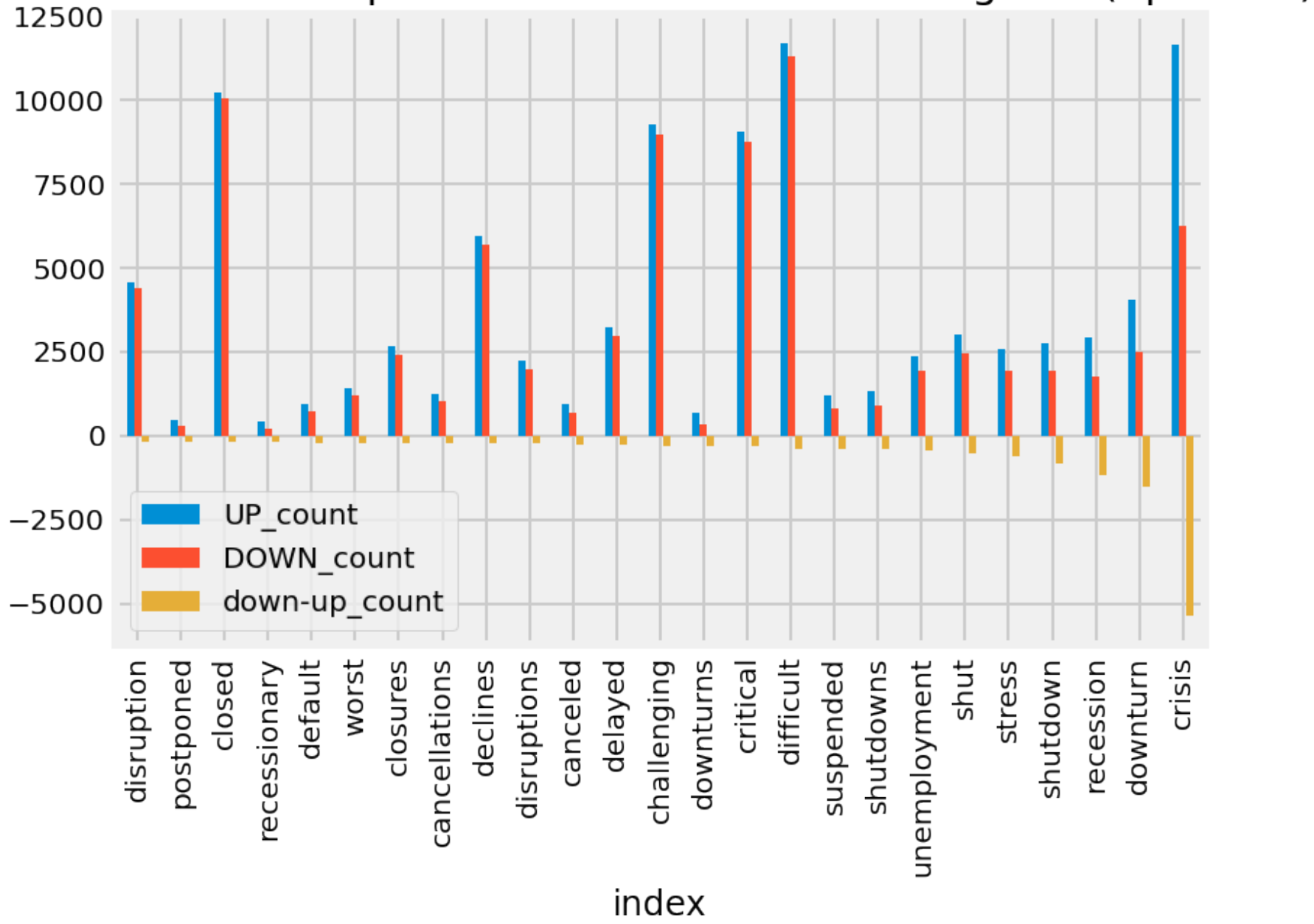




Most frequent words in the different text categories

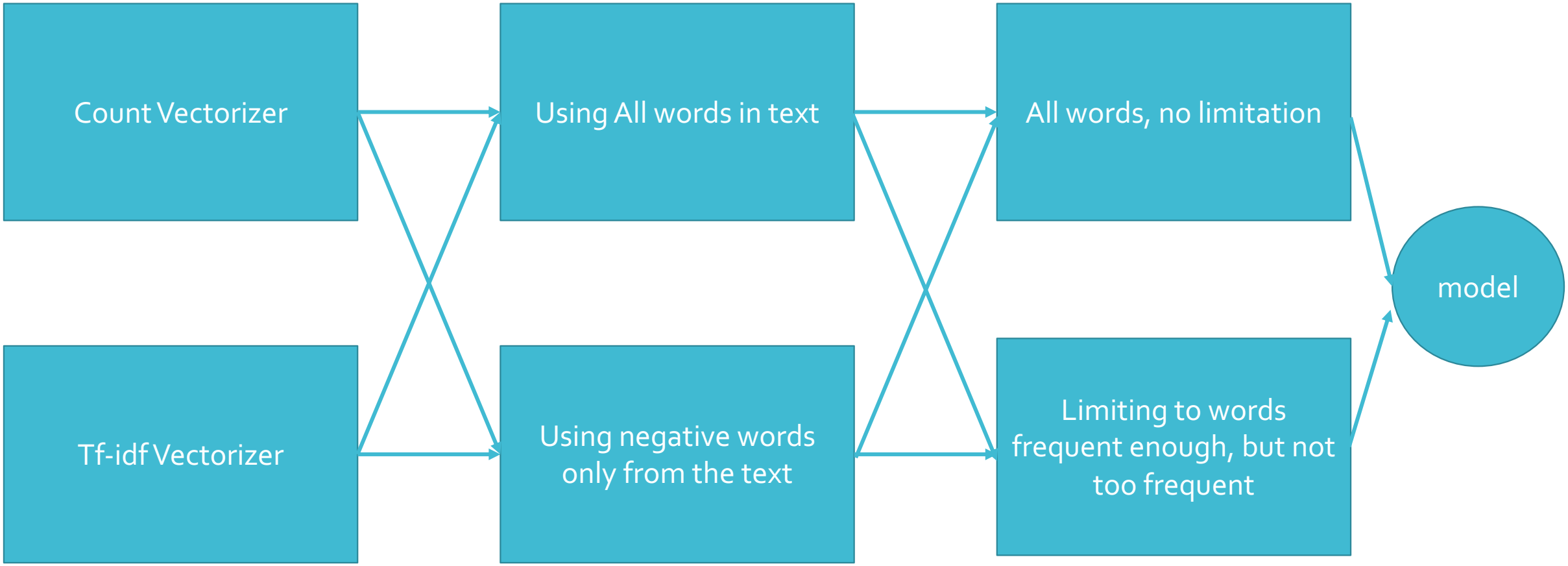
	"UP" texts			"DOWN" texts
quarter	442990		year	511566
year	441121		quarter	485908
think	313211		million	357438
chief	305182		think	333019
officer	304965		chief	330633
million	302487		officer	330317
just	274812		just	299032
analyst	242210		analyst	262752
executive	225536		executive	244000
business	203281		growth	221319


word count comaprision where difference is highest (down - up)
25000
20000
15000
10000
5000
0
UP_count
DOWN_count
down-up_count
loss
closing
decline
against
declined
negative
litigation
impairment
recall
claims
negatively
challenges
problem
divestitures
volatility
downtime
divestiture
restructuring
declining
force
opposed
challenge
unfortunately
threat
index

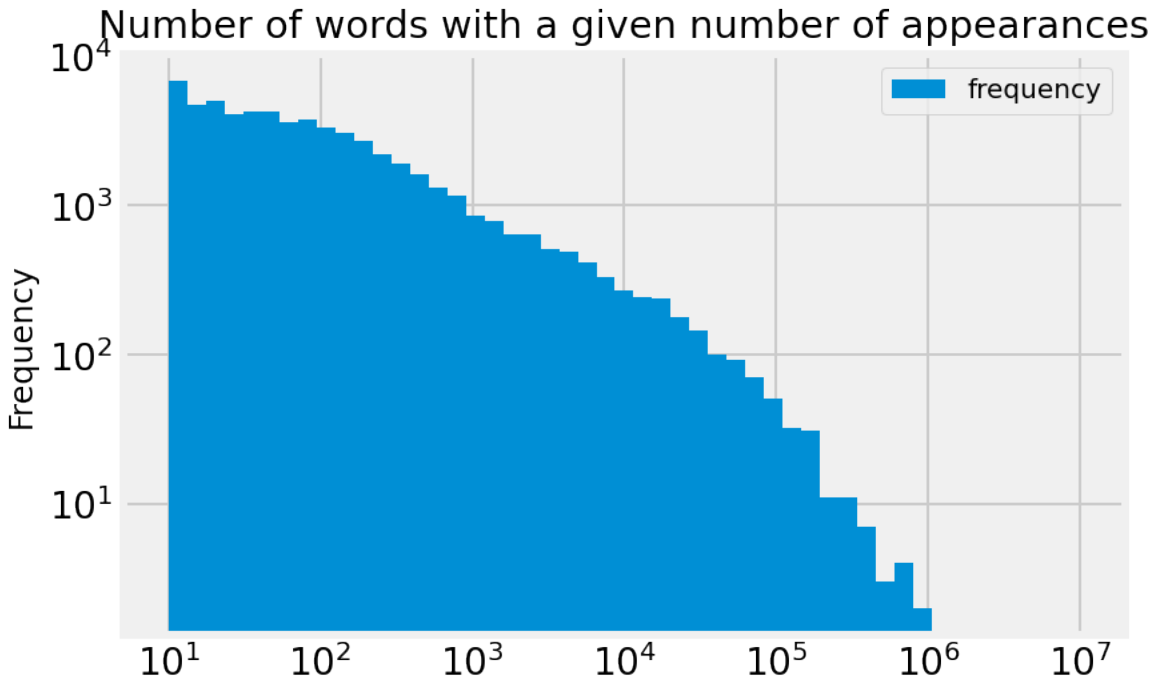

word count comaprision where difference is highest (up-down)
12500
10000
7500
5000
2500
0
−2500
−5000
UP_count
DOWN_count
down-up_count
disruption
postponed
closed
recessionary
default
worst
closures
cancellations
declines
disruptions
canceled
delayed
challenging
downturns
critical
difficult
suspended
shutdowns
unemployment
shut
stress
shutdown
recession
downturn
crisis
index

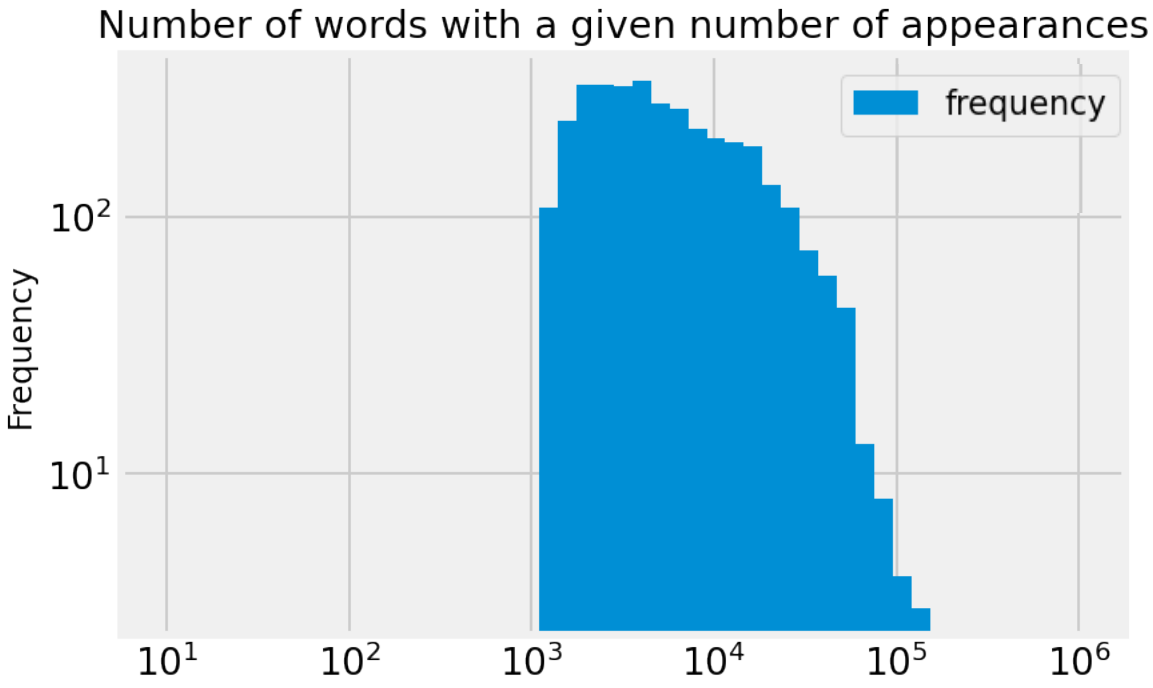
Modelling


YOU ARE
HERE



Eight different set of train features





- **Used words:**
- analysed words: 150739

- **After limitation**
- analysed words: 3430
- Negative words: 105
- Positive words: 127

Model selection:

- Stratified train-test split
- Cross validation with 5 folds
- ROC_AUC Scoring:
  - + give importance to both classes
  - + adjust to baseline

Models tried:

- Logistic Regression: performs well, relatively quick
- Random Forest: most accurate results with words features only, slower
- KNN: slow and weaker results (0.48)
- Naïve Bayes: relatively quick, results close to best models (0.58)
- Support Vector Classifier: very slow (0.56)

Tuning methods tried:

- Randomized Search CV
- Grid Search CV

Other techniques:

- Change threshold for prediction
- Combine results from different model

Further features:

- Sentiment score
  (positive – 2 x negative
- Time aspects of call
  (year, month, day of week)

Baseline	Fraction
DOWN	0.527
UP	0.473

Acc 0.634	Pred D	Pred U
True D	1238	949
True U	567	1393

Acc 0.60	Pred D	Pred U
True D	1440	895
True U	762	1079

	Test results	Roc auc	Accuracy	Precision	Recall
1	Logistic Regression on engineered features	0.675	0.62	0.64	0.62
2	Random Forest – cvec on all words in text, limited by frequency	0.645	0.59	0.58	0.81
3	Random Forest – tvec on all words in text, limited by frequency	0.640	0.59	0.58	0.81
4	Logistic Regression – cvec on all words in text, limited by frequency	0.627	0.59	0.59	0.69
5	Logistic Regression – tvec on all words in text, limited by frequency	0.622	0.58	0.58	0.74
6	Random search of Random Forest – all words, not limited	0.616	0.58	0.56	0.88

1) Logistic Regression on engineered features

Acc 0.62	Pred D	Pred U
True D	1363	824
True U	754	1206

Acc 0.634	Pred D	Pred U
True D	1238	949
True U	567	1393

Modified threshold:

0.55

2) Random Forest Cvec all words, limited frequency

Acc 0.59	Pred D	Pred U
True D	1791	411
True U	1287	687

Acc 0.60	Pred D	Pred U
True D	1440	895
True U	762	1079

Modified threshold:

0.589025

4) Logistic Regression Cvec all words, limited frequency

Acc 0.59	Pred D	Pred U
True D	1518	684
True U	1037	937

Modified threshold:

no improvement

Comparing features from a random forest and a logistic regression model

2) Random Forest feature importance

Cvec, all words, limited frequency

feature	importance
callfeb	0.0173
coronavirus	0.0095
callmay	0.0088
march	0.0073
april	0.0049
2021	0.0041
2018	0.0041
pandemic	0.0033
2017	0.0033
crisis	0.0032

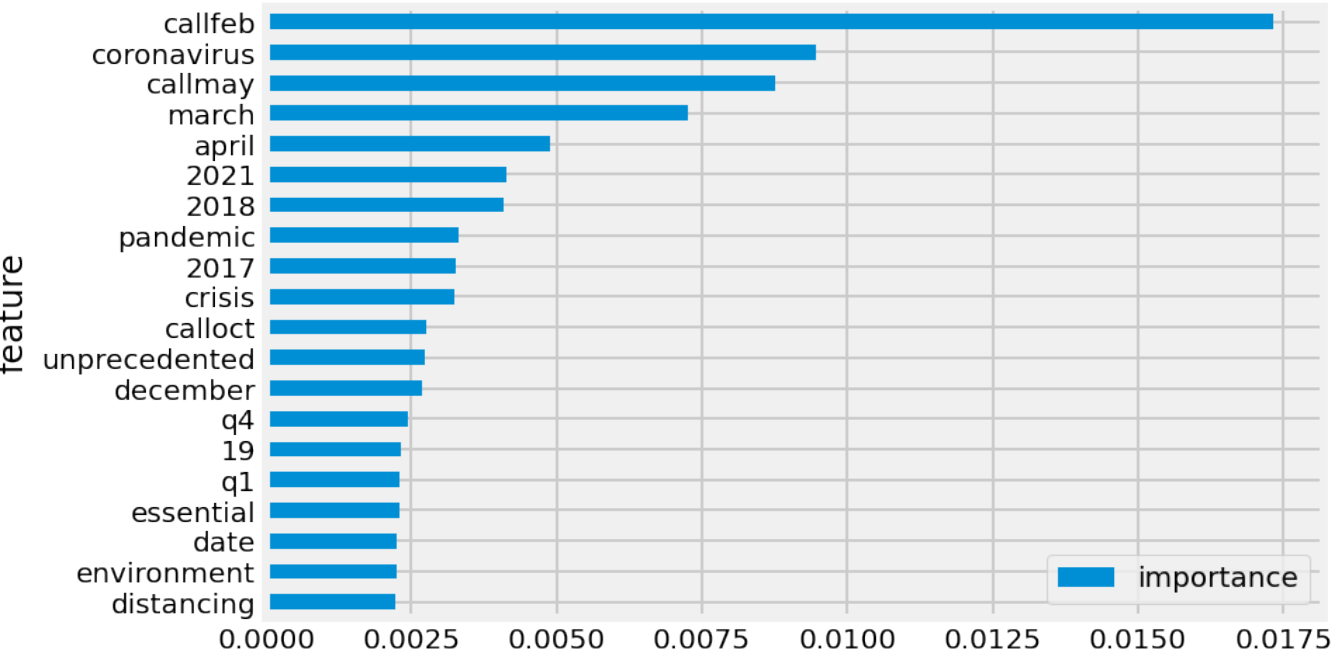
Avg importance 1/3450 = 0.00029

4) Logistic Regression coefficients

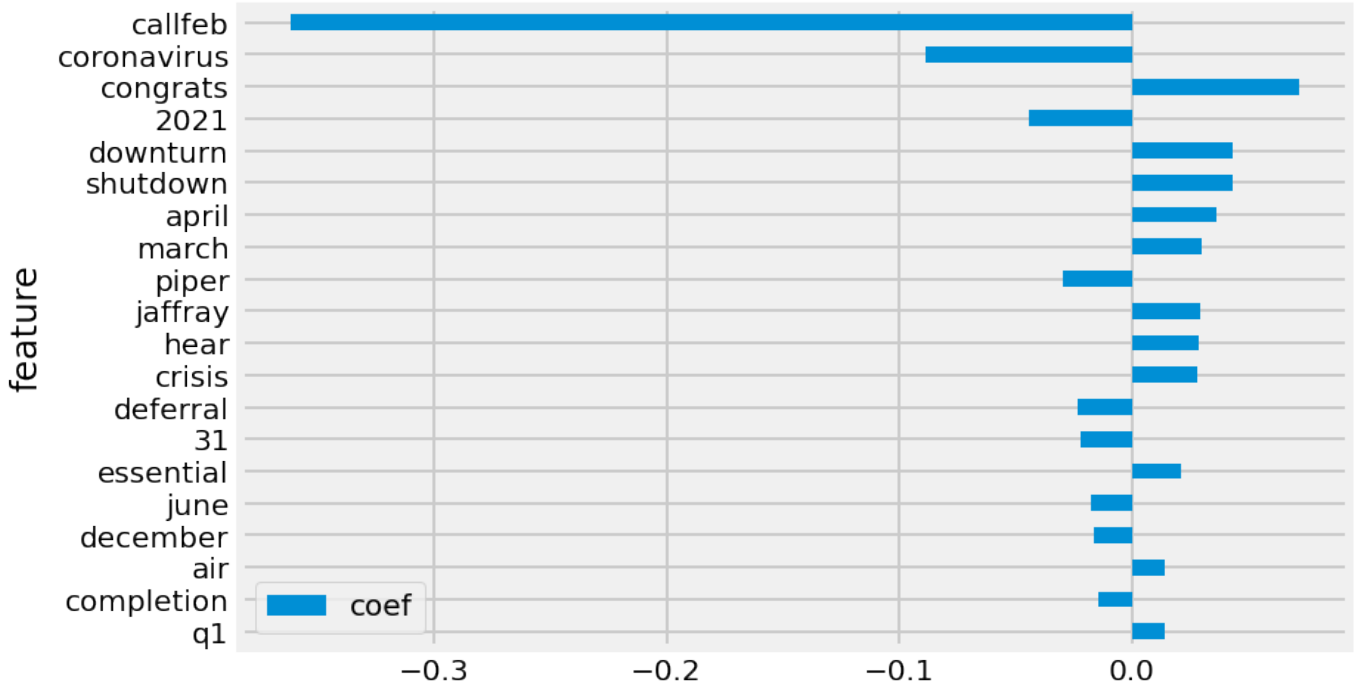
Cvec, all words, limited frequency

feature	coef	abs_coef
callfeb	-0.362	0.362
coronavirus	-0.089	0.089
congrats	0.072	0.072
2021	-0.044	0.044
downturn	0.043	0.043
shutdown	0.043	0.043
april	0.036	0.036
march	0.030	0.030
~~piper~~	~~-0.029~~	~~0.029~~
~~jaffray~~	~~0.029~~	~~0.029~~
hear	0.029	0.029
crisis	0.028	0.028

Random Forest – feature importance

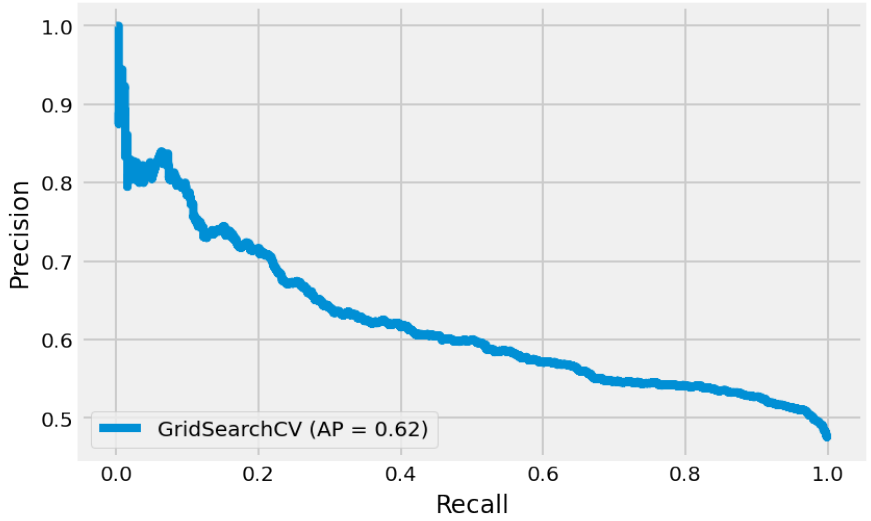


Logistic regression - coefficients

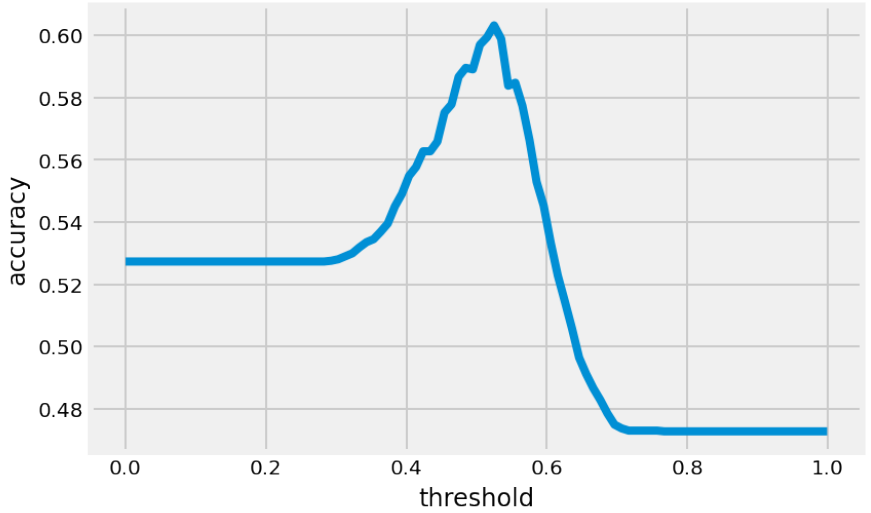


feature	coef	abs_coef
year_2018	-1.094703	1.094703
month_5	1.077782	1.077782
month_4	0.995323	0.995323
month_10	0.760236	0.760236
year_2020	-0.725518	0.725518
month_6	0.716034	0.716034
month_11	0.554249	0.554249
month_8	0.382188	0.382188
month_2	-0.337885	0.337885
month_12	0.332375	0.332375
month_7	0.307449	0.307449
month_9	0.292956	0.292956
month_3	0.237999	0.237999
weighted score	**-0.223201**	**0.223201**
year_2019	0.079950	0.079950
dayofweek_3	-0.069410	0.069410
dayofweek_2	-0.040657	0.040657
dayofweek_4	0.011870	0.011870
duration	0.010399	0.010399
dayofweek_1	0.004304	0.004304

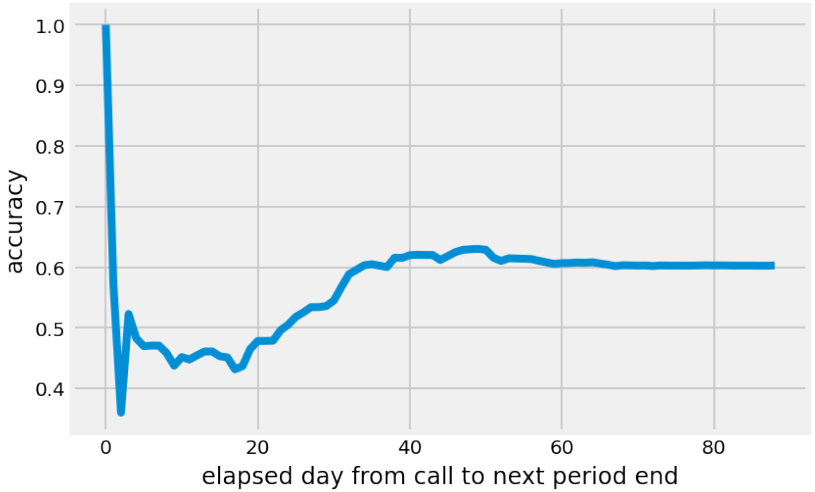
Random Forest – cvec – all words



Accuracy at different prediction thresholds



Threshold: 0.52477
Accuracy: 0.6032



Max at
Delta days: 52
Accuracy: 0.63

Conclusions for project

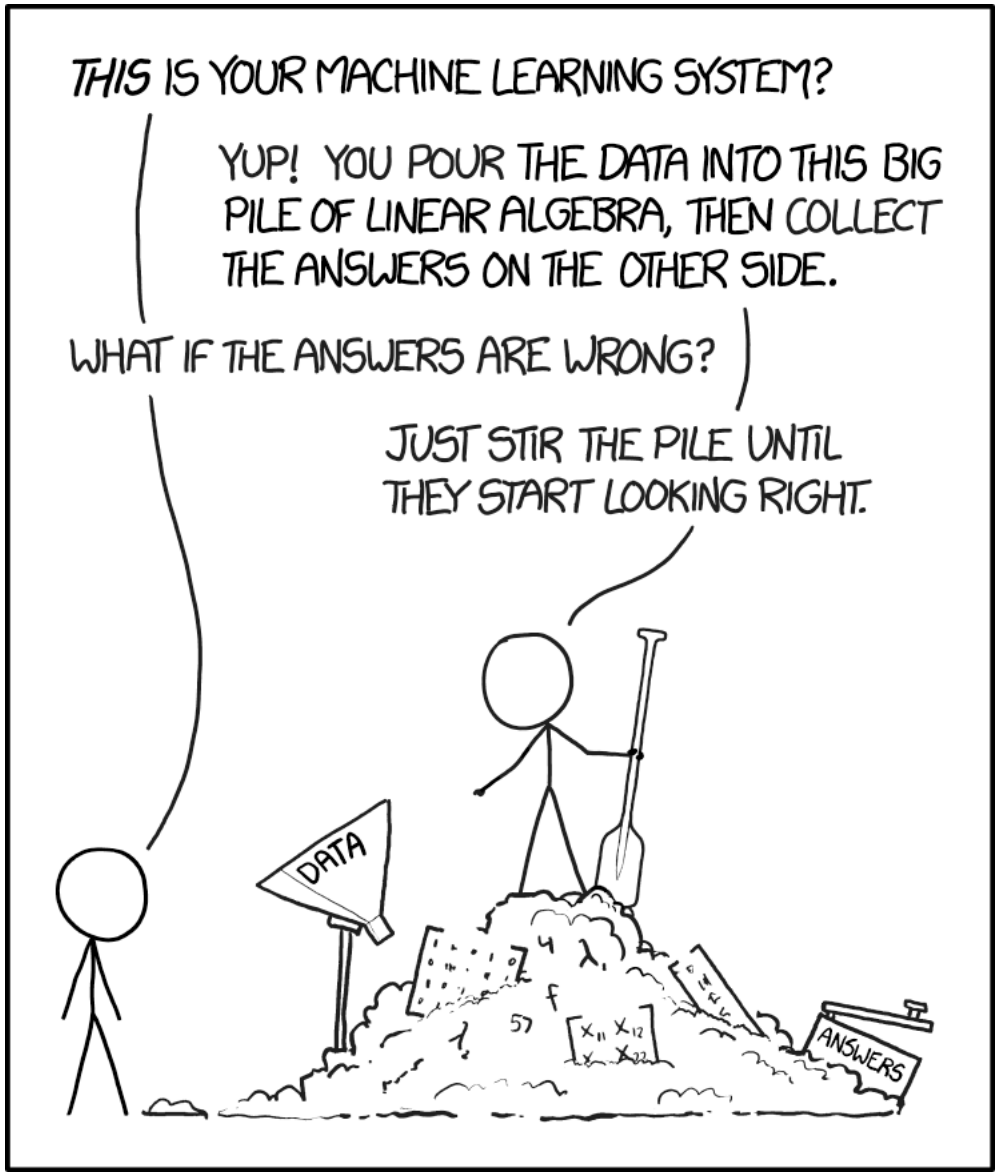
Conclusions:

- There is explanatory power of words
  - Just using text features accuracy was 58-59%
- Features from all words performed better than negative words alone (roc_auc around 0.56, accuracy around 0.55)
- Most important features contain reference to time aspect

Lookout:

- Evaluate the model performance based on a simulated portfolio
- Re-label my data and define UP-DOWN compared to a global index's movement
- Find a more consistent timeframe that works (52 days)
- Explore ngrams and use stemming
- Split up the categories:
  - significantly outperforms
  - outperforms
  - underperform
  - significantly underperforms

Thank you!


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS