Predicting Stock Prices from Tweets

By:
Dhammatorn Riewcharoon
Yumiko Suwannaroj
Preston Akwule
Farah Jardaneh

Motivation

- In today's Financial Markets everyone is looking for any edge they can get over others
- Many financial institutions are already integrating Financial News sources into their analysis (such as Bloomberg)
- We want to see if we can extract extra value using **unconventional text** resources such as **twitter** that can give us an edge over what everyone else is doing.

Quantity of Interest

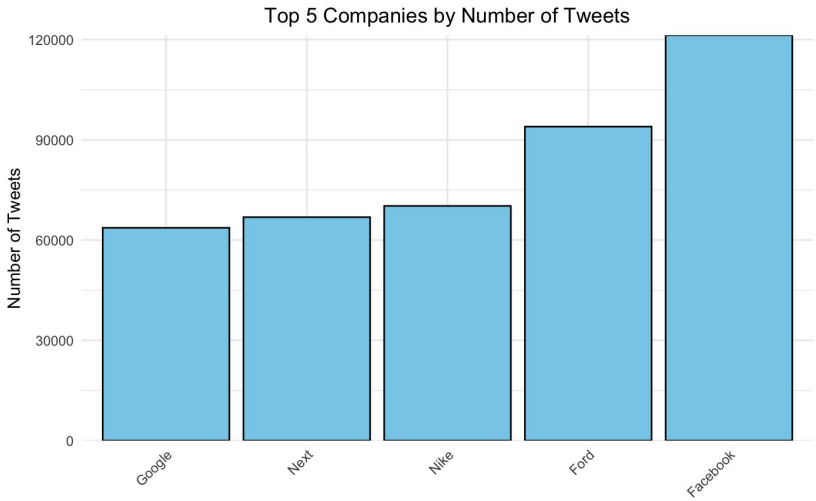
- Initially: Daily Returns (as a binary)
- Later on: Industry (to tackle a more straightforward problem)

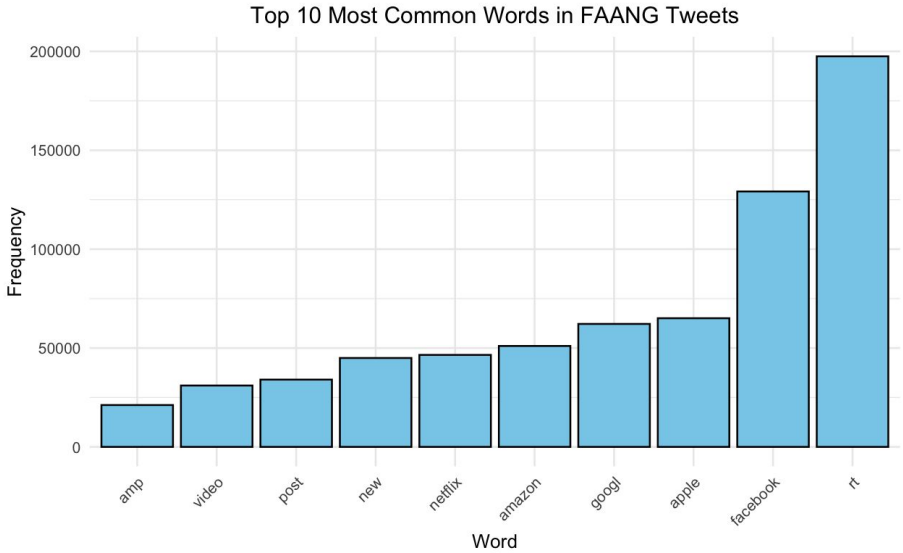
Population of Interest

- Random tweets between 2017-2018 tagging a set of 100 companies. Our tweets do not focus on a single topic, it could be news about the company, it could be a customer complaint, it could be a review etc. (Open domain)

Data Overview

- 100 companies (FAANG vs Non FAANG)
- Distribution of data (800k+ labeled tweets prior to pre-processing)





Analysis Overview (1)

- Preprocessing steps:
  - Replacing HTML
  - Replacing URL
  - Replacing Emoji
  - Removing missing values
  - Created binary indicator (made_money) based on stock return
  - Generated DFMs to identify common bigrams and unigrams, one excluding company names using a custom stop list to avoid bias

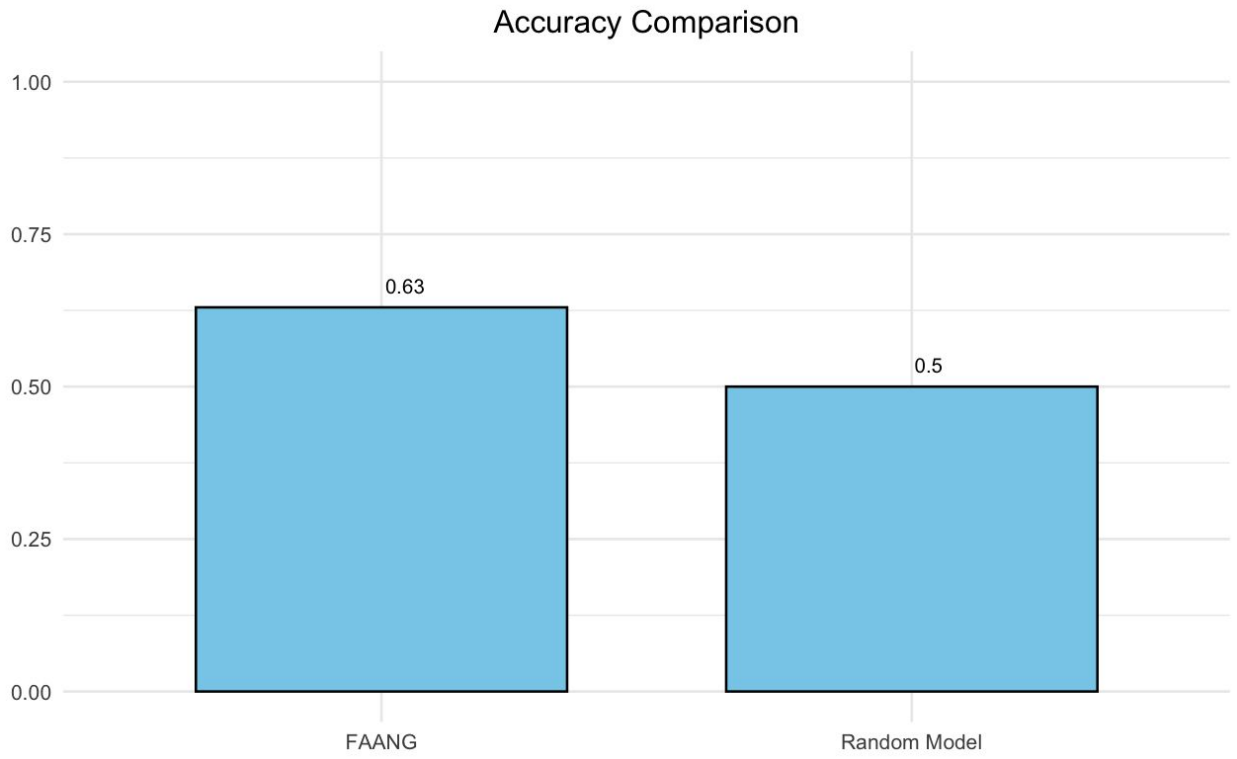
Analysis Overview (2)

- Employed techniques:
  - N-gram LASSO and vector embeddings on FAANG companies
  - Transfer learning to generalize models to non-FAANG companies
  - Integrating sentiment dictionaries
  - Using sentence structure with Spacy
  - These techniques can capture semantic patterns and allow models to adapt to diverse dataset for better market predictions

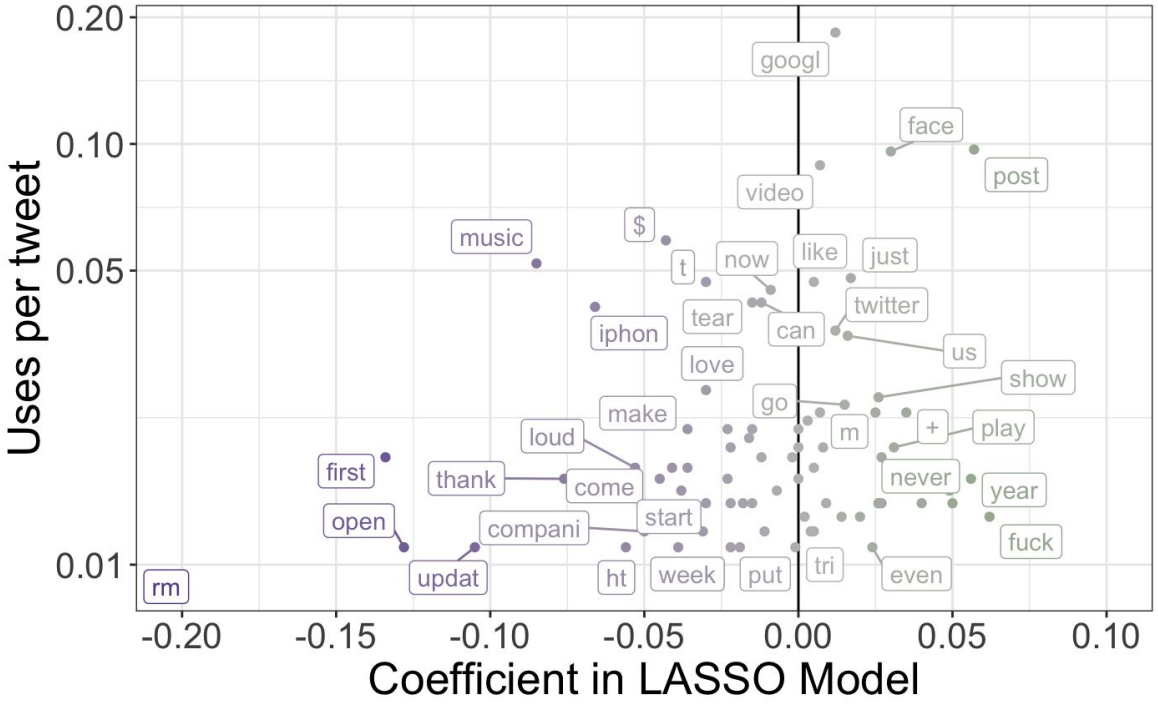
Looking at Daily Return as Binary Outcome

- 0 means price decreased, 1 means stock price increased
- Started by training on FAANG only companies, as they have a substantial weight on the performance of the S&P 500

N-gram LASSO on FAANG companies



N-gram LASSO coefficient plot



Tweets with words having large coefficients

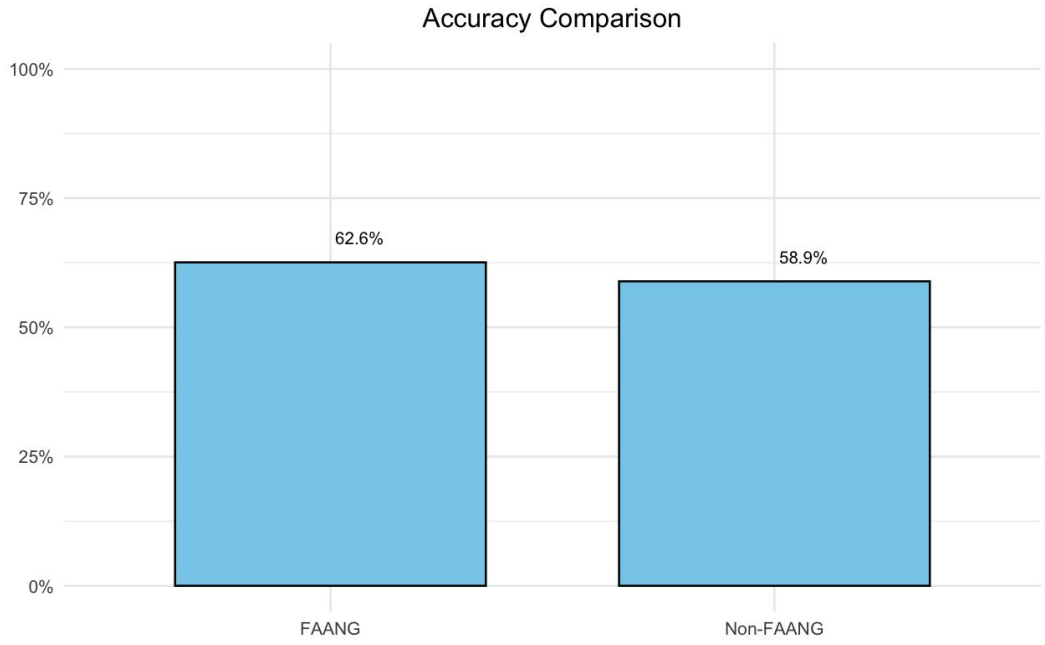
I just finished the haunting of hill house and now i feel empty, **thank** you @netflix

"Why the **f*** ** can't I stop watching Will Arnett talk about a toaster and microwave?! @netflix what the **f** ** is this #NetflixLive"

@Apple what a coincidence this shit happens to specifically **iphone** 7 s a couple weeks before the 8 & X drop

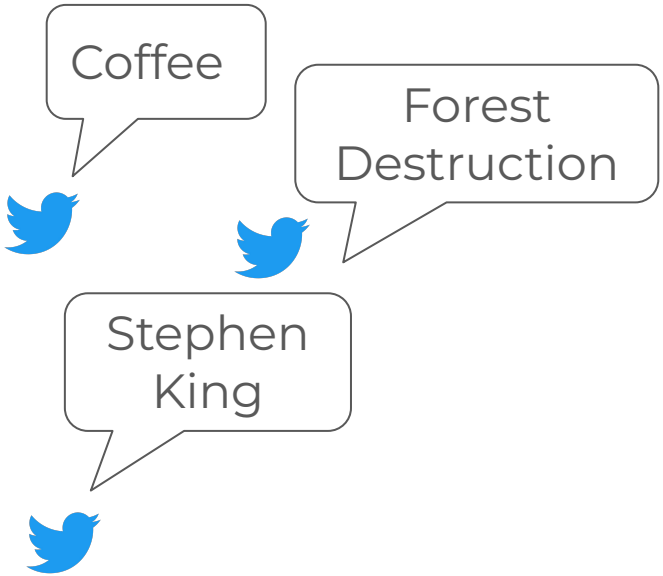
Transfer learning : FAANG model on Non-FAANG

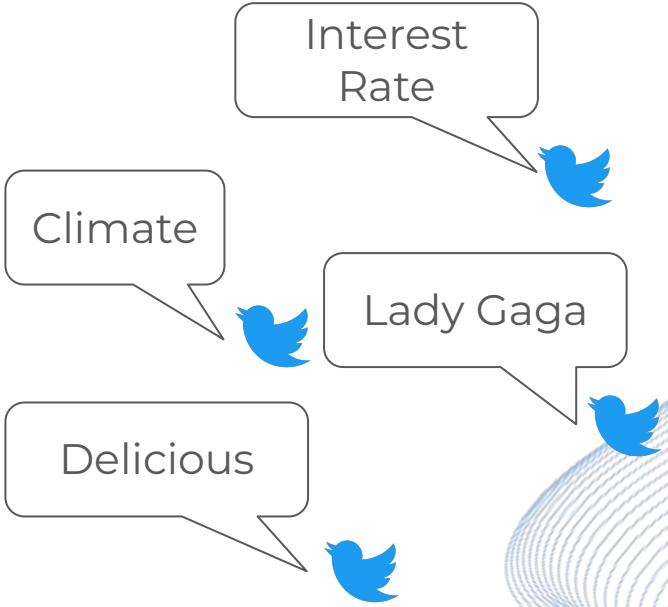
Predicted	Actual 0	Actual 1
0	4,987	7,758
1	20,2143	295,749



Why do we think the performance decreases?

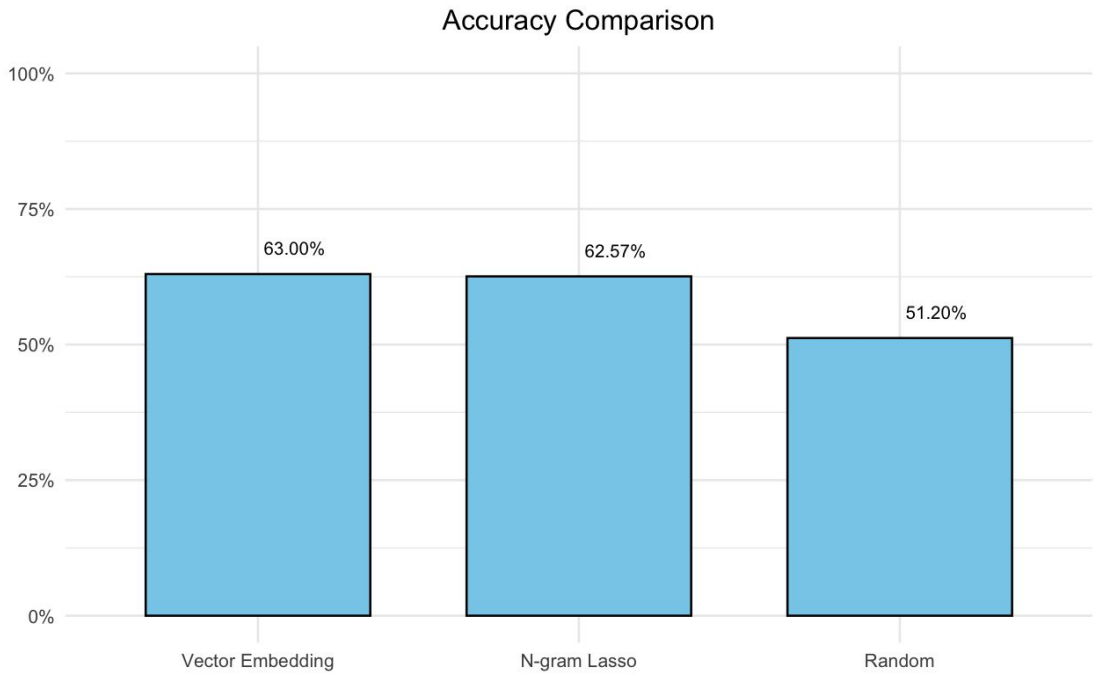
- FAANG companies are all tech companies
- Model hasn't seen words outside of tech



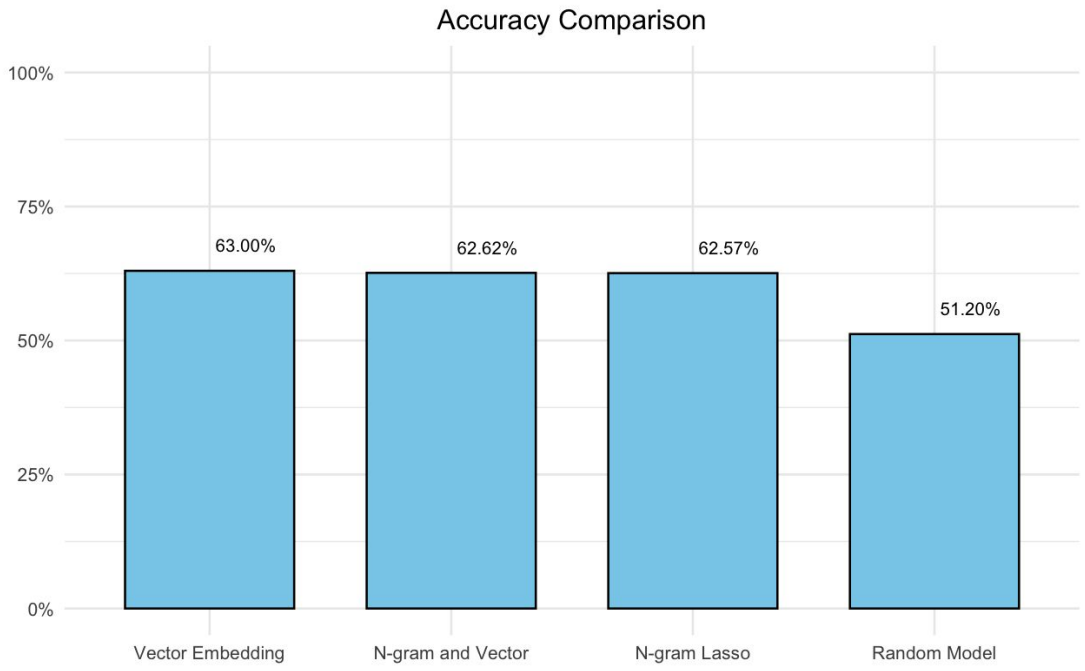


Vector Embedding on FAANG companies

Predicted \ Actual	0	1
0	2,912	1,667
1	24,641	41,069

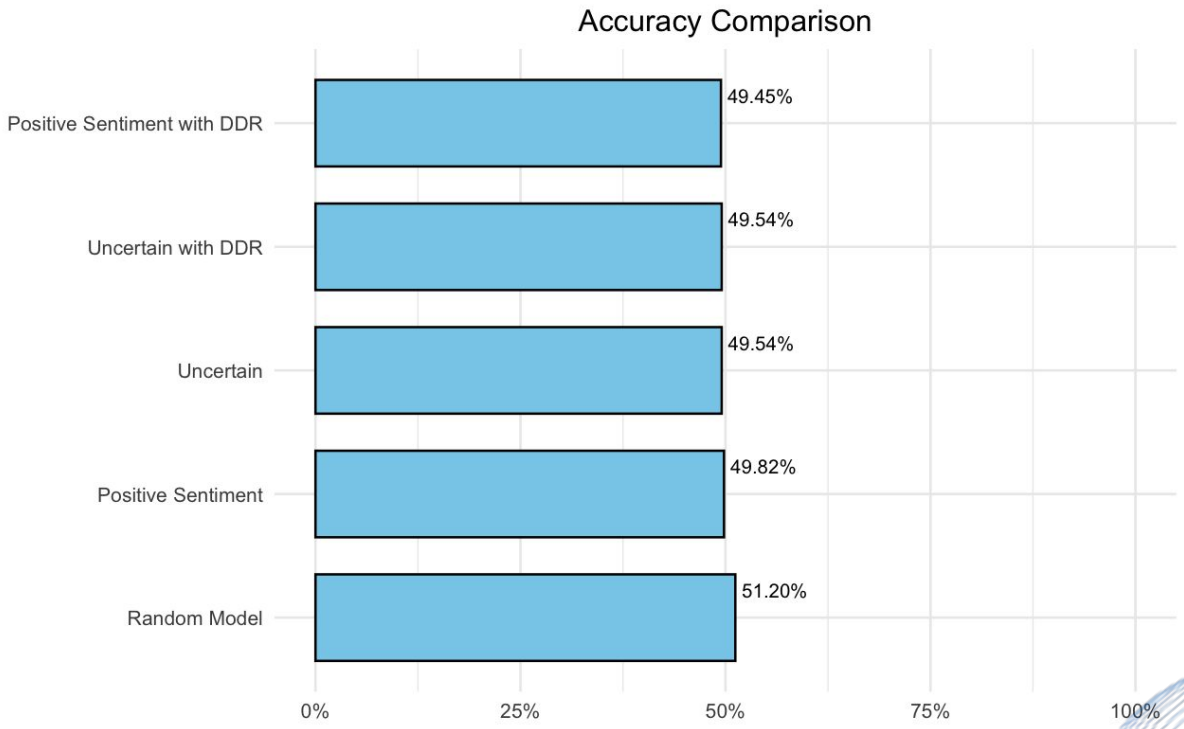


Vector Embedding and LASSO model



Predicted \ Actual	0	1
0	3246	1727
1	24282	41034

Positive Sentiment and Uncertain Dictionary Model on FAANG



Limitations and potential future improvements

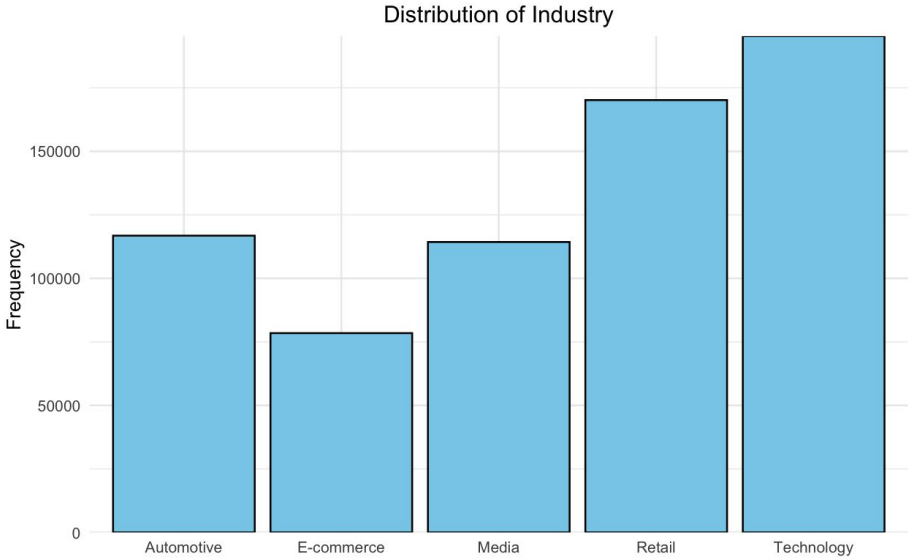
- Predict on risk premium instead of returns
- We're still predicting a lot of False Positives
- Training the mode on specific industries to ensure it captures the semantics related to the industry
- More data from different time period to ensure our data is not biased

Looking at Industry

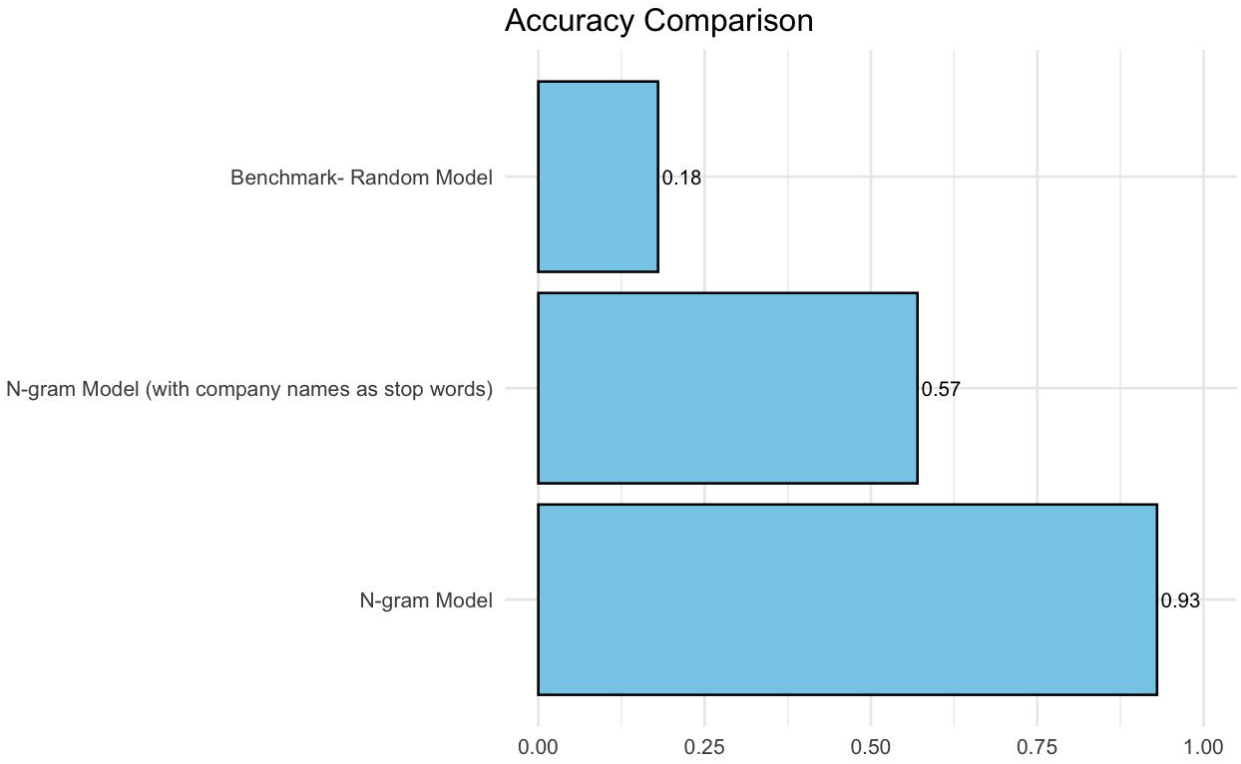
Tackling a more straightforward problem

Industry

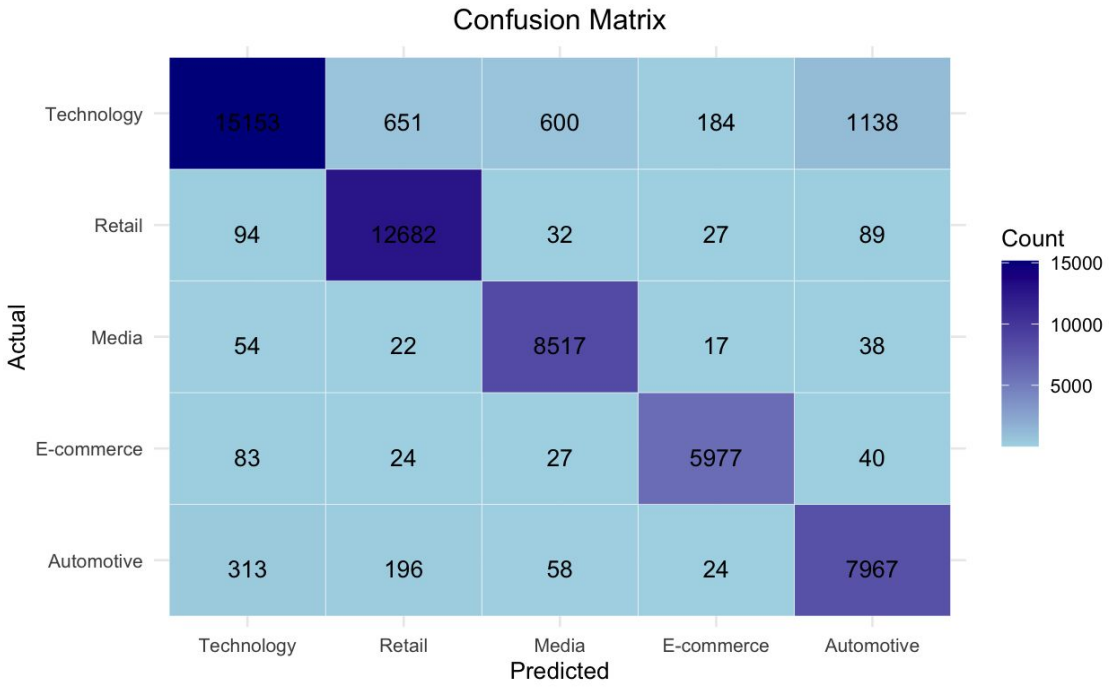
- A label we added ourselves to the dataset from the “stock” column
- We only used a subset of our data due to computational limitations, as well as ensuring we had enough tweets for each industry



Initial Models



Best Performing Model (N-grams)



Mislabeled Industries

RT @peta: BREAKING VICTORY! 👏 After over a decade of campaigning by PETA around the world, @Burberry is banning fur and angora. Burberry'...

Predicted: Technology

Label: Retail (Burberry)

Check out what I found on eBay : adidas Alphabounce EM Shoes Men's https://t.co/CGzczy4F3i https://t.co/Q6yQDdwix

Predicted: E-commerce

Label: Retail (Adidas))

Toyota says in talks with Geely on cooperation in hybrid vehicle tech | Reuters https://t.co/tRwVPh4jYZ

Predicted: Media

Label: Auto (Toyota)

Dropbox uploads a $600M credit line, Audi acquires Silvercar and more on #CrunchReport with titoyooo https://t.co/YGiijeyQuf

Predicted: Technology

Label: Automotive (Audi)

Mislabeled Industries

RT @peta: BREAKING VICTORY! 👏 After over a decade of campaigning by PETA around the world, @Burberry is banning fur and angora. Burberry'...

Predicted: Technology

Label: Retail (Burberry)

Check out what I found on eBay : adidas Alphabounce EM Shoes Men's https://t.co/CGzczy4F3i https://t.co/Q6yQDdwix

Predicted: E-commerce

Label: Retail (Adidas))

Toyota says in talks with Geely on cooperation in hybrid vehicle tech | Reuters https://t.co/tRwVPh4jYZ

Predicted: Media

Label: Auto (Toyota)

Dropbox uploads a $600M credit line, Audi acquires Silvercar and more on #CrunchReport with titoyooo https://t.co/YGiijeyQuf

Predicted: Technology

Label: Automotive (Audi)

Mislabeled Industries

- There seems to be an issue in the model where there are two companies within a tweet.
- Sometimes the confusion is justified: this tweet is truly about two companies!

Dropbox uploads a $600M credit line, Audi acquires Silvercar and more on #CrunchReport with titoyooo https://t.co/YGiijeyQuf

Predicted: Technology

Label: Automotive (Audi)

Mislabeled Industries

- There seems to be an issue in the model where there are two companies within a tweet.
- Other times, one company is clearly the subject of the tweet!

Toyota says in talks with Geely on cooperation in hybrid vehicle tech | Reuters https://t.co/tRwVPh4jYZ

Predicted: Media

Label: Auto (Toyota)

Suggested Solution

- Extract tweet subject “Nsubj” using Spacy
- If the model had access to the main subject(s) of the tweet, maybe it can perform better!

```r
extract_subject <- function(text) {
  parsed <- spacy_parse(text,
                        lemma = T,
                        dependency = T)

  main_nouns <- parsed$lemma[parsed$dep == "nsubj"]

  if (length(main_nouns) > 1) {
    return(paste(paste0(main_nouns, "_subj"), collapse = ", "))
  }

  if(length(main_nouns) == 0) {
    return("")
  }
  return(paste0(main_nouns, "_subj"))
}
```

Result on our Tweet

Toyota says in talks with Geely on cooperation in hybrid vehicle tech | Reuters https://t.co/tRwVPh4jYZ

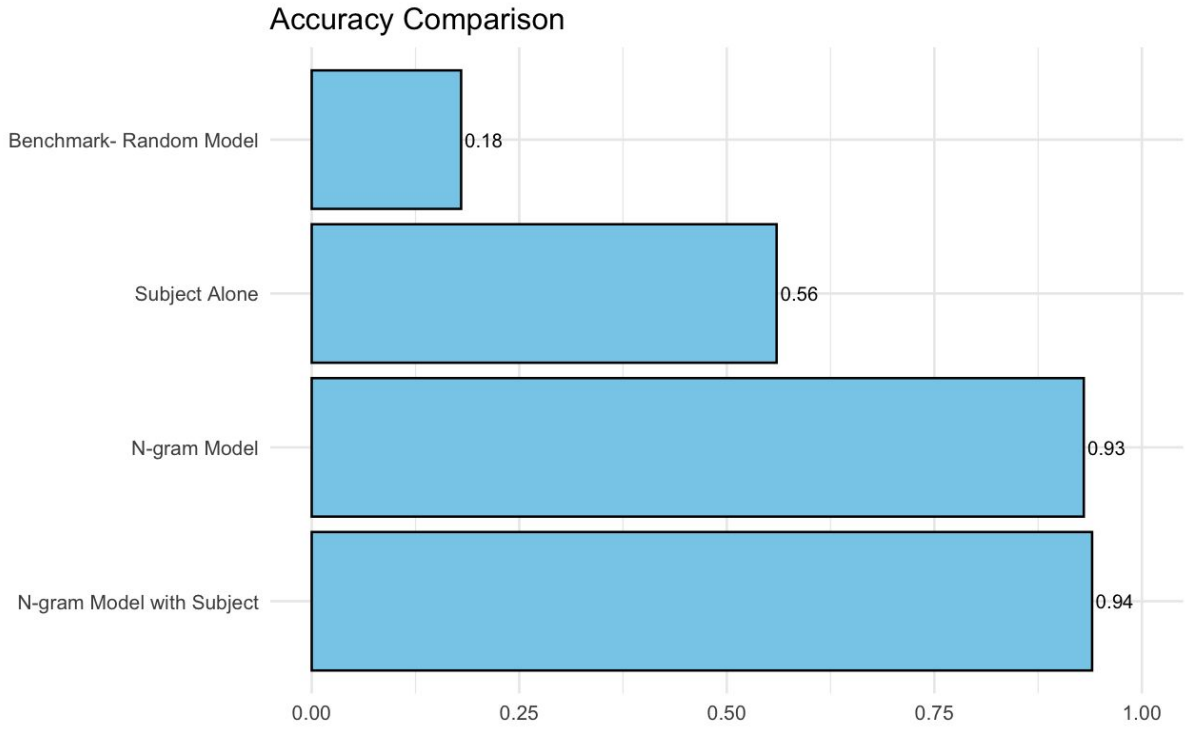
Subject Identified by Spacy:

Toyota

New Classification:

Automotive Industry

Updated Models



Only **1%** increase in accuracy!

Juice is **not worth** the squeeze, given how computationally expensive Spacy is

Thank you!