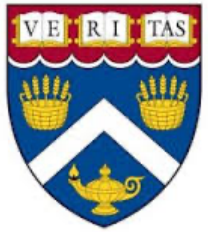
Final Project

Topics and Sentiment in Financial Markets from Twitter



Text Analytics and Natural Language Processing

Harvard University Extension School

Eduardo Peynetti

Financial Twitter

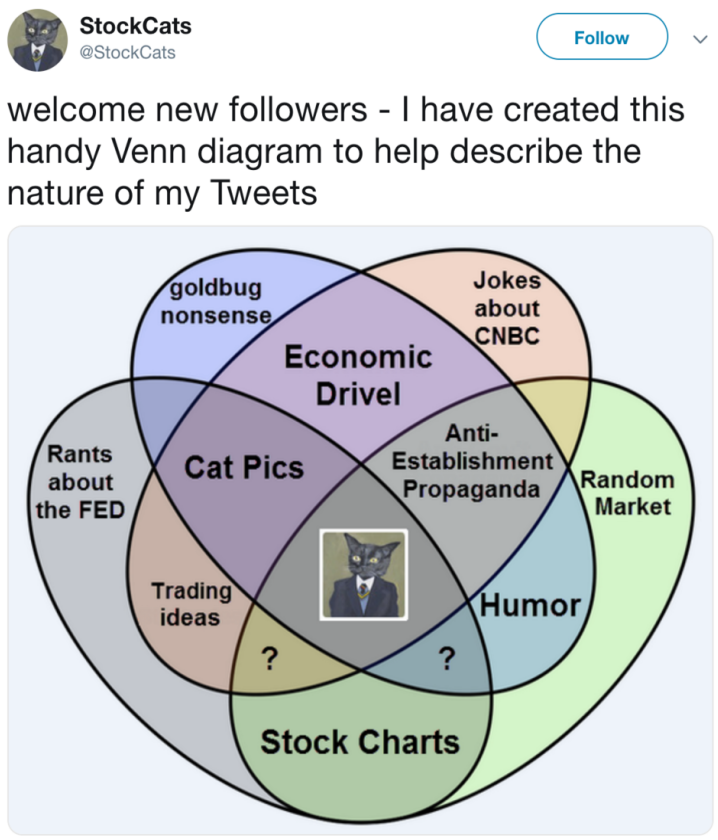
```
model = pipeline(
    "text-generation",
    model="pearkes/"
)
```

```
model("The future")
```

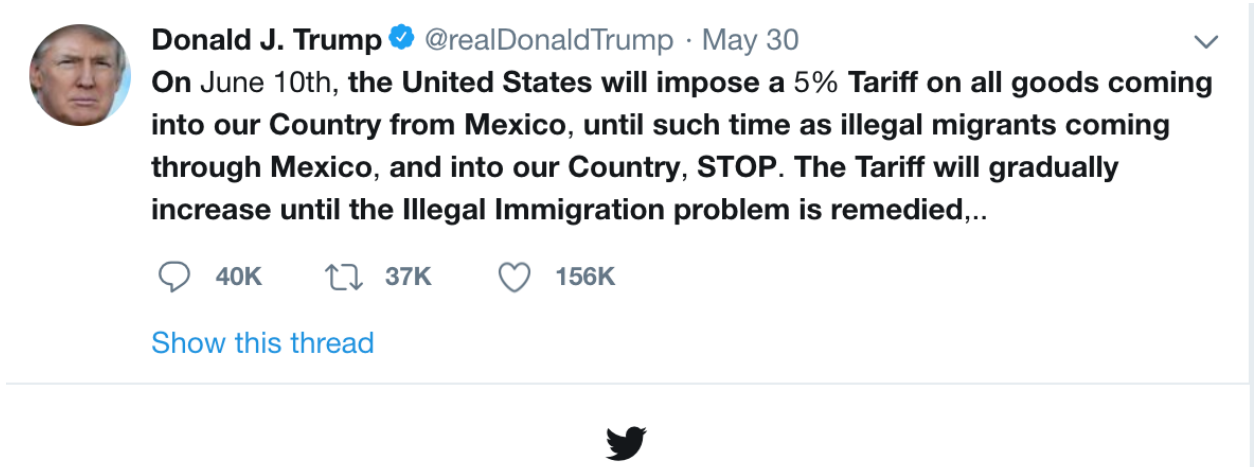
```
Setting `pad_token_id` to 50256 (first `eos_token_id`) to generate sequence
```

```
[{'generated_text': 'The future of finance. “Financial markets” is all the rage right now. \xa0…\nFintwit was a co
ol project. And with any luck, if it does okay I will get to read it, it is still going'}]
```

Financial Twitter



- Traders, analysts, reporters, anyone involved in markets
- Quick dissemination of news
- Expert (and not so expert) opinions on market relevant topics



Introduction

Objectives:

- Detect **relevant themes** and **sentiment** influencing financial markets
- **Quantify** changes in sentiment and theme relevance through time
- Attempt to **predict market movements** with this information

Data:

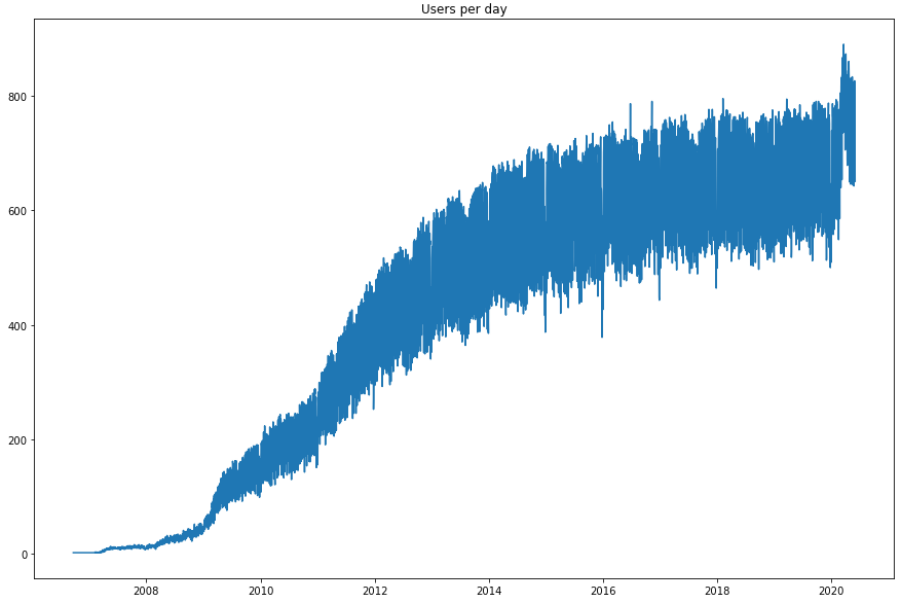
- Twitter feeds for over 1250 accounts
- Daily closing prices for S&P500

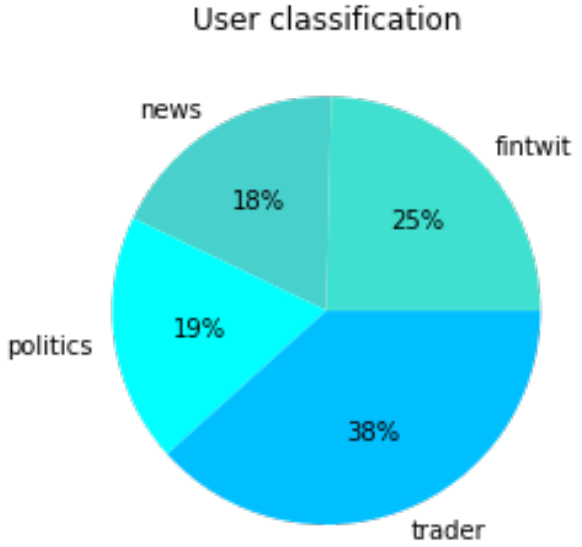
Approach:

- Topic modeling for market themes
- Pre-trained BERT for sentiment

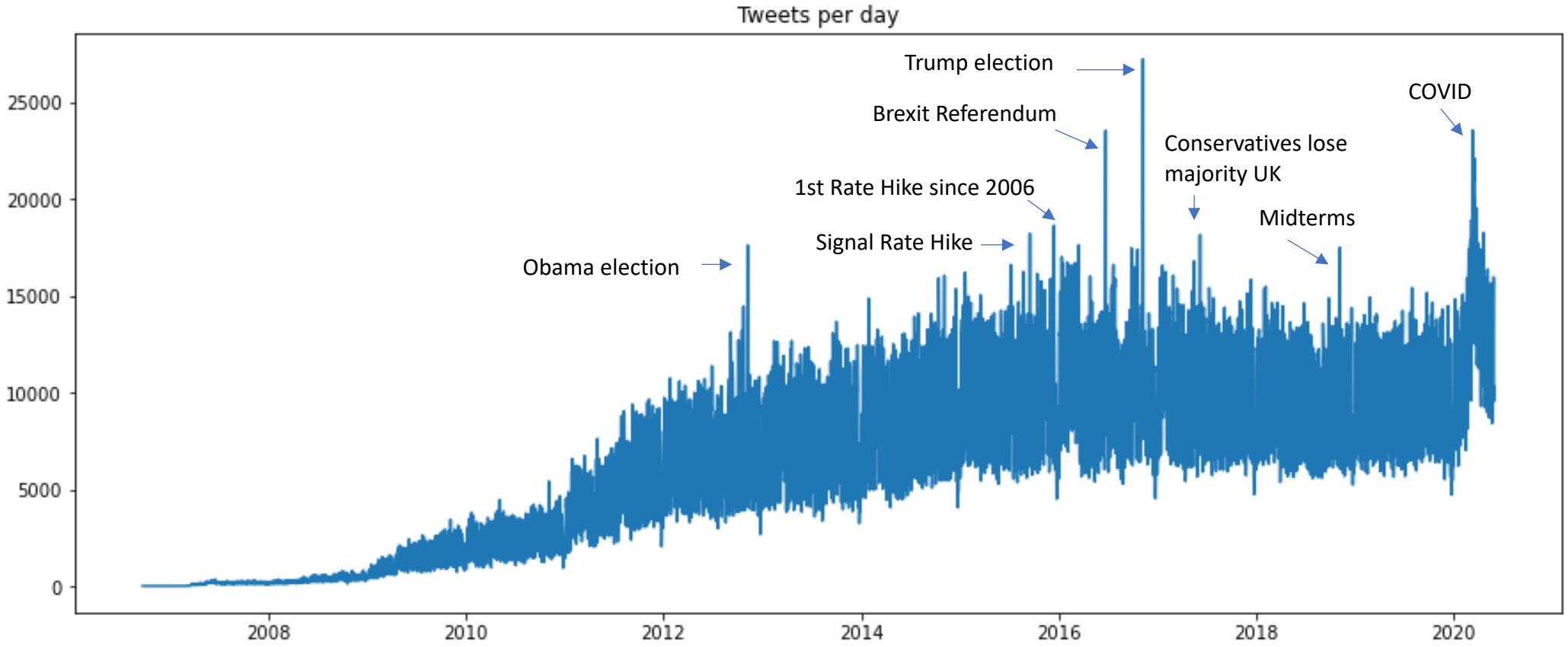
Data Set

- Over 35 million tweets, 1250 users, 6 GB raw data
- Scraped with **Twint**
- Timestamp up to second, unique tweet ID
- Twitter users labeled as **Fintwit, Trader, News Sources, Politics**

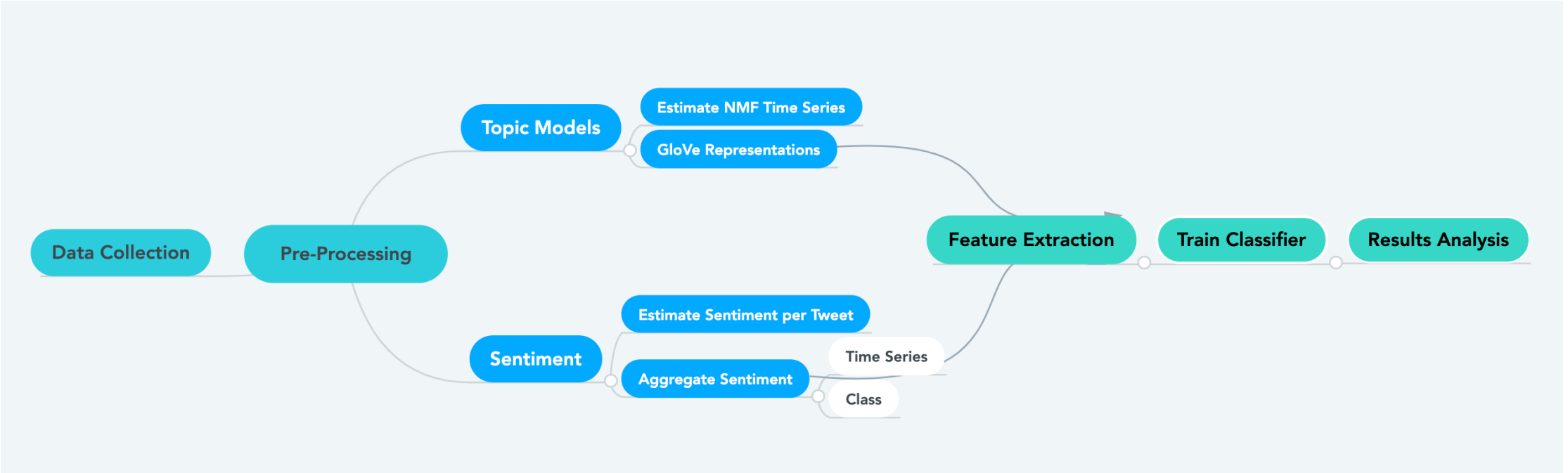




Data Set

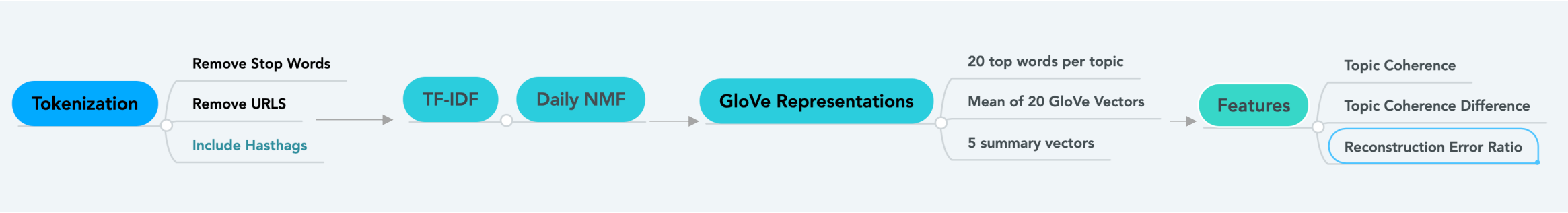

Tweets per day
Trump election
COVID
Brexit Referendum
Conservatives lose majority UK
1st Rate Hike since 2006
Midterms
Signal Rate Hike
Obama election
25000
20000
15000
10000
5000
0
2008
2010
2012
2014
2016
2018
2020

Pipeline



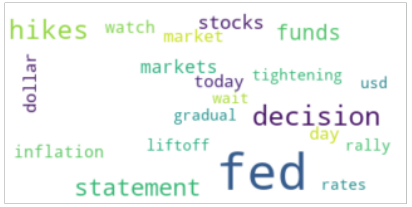
- Time Series with cutoff every market day at 3:45pm EST

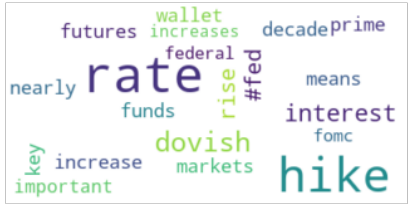
Topic Modeling

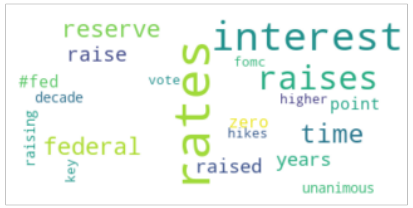


- 5 topics per model
- **Topic Coherence**: Pairwise average of cosine similarities between summary vectors of one day
- **Topic Coherence Difference**: Pairwise average of cosine similarities between summary vectors of two days
- **Reconstruction Error Ratio**: Transform data with previous day's model, calculate reconstruction error.
  Ratio between reconstruction error today / yesterday, to take account weekends

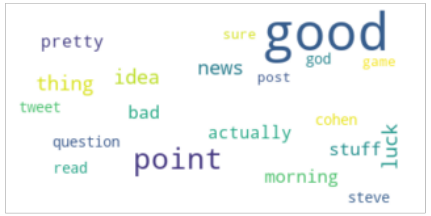
Topic Modeling

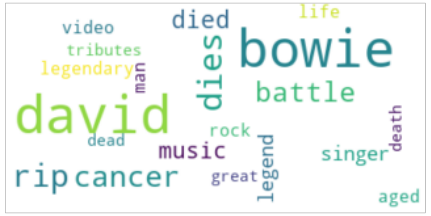






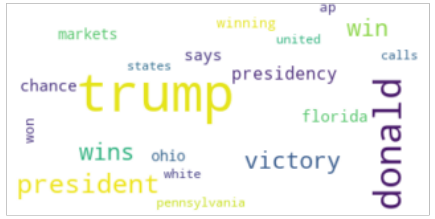
17 December 2015
Fed Rate Hike

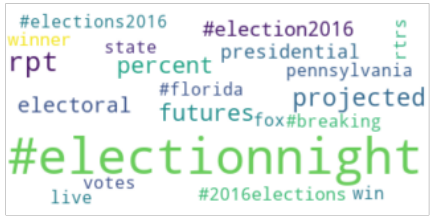


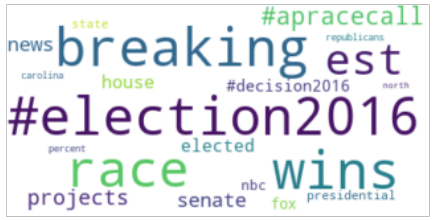


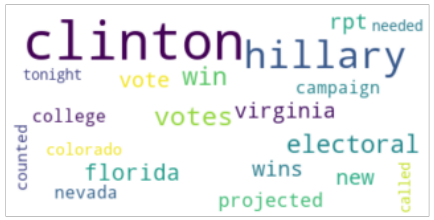
11 January 2016

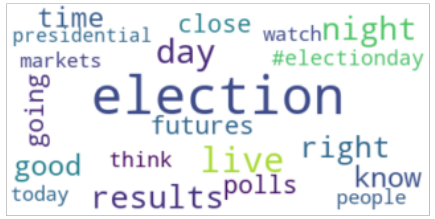
Topic Modeling



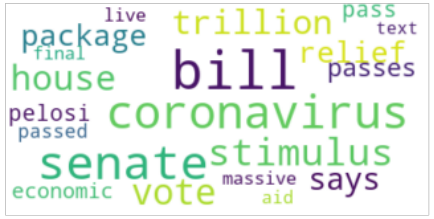


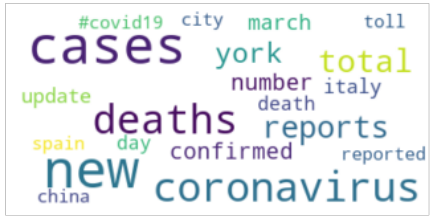


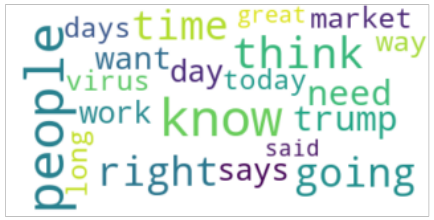


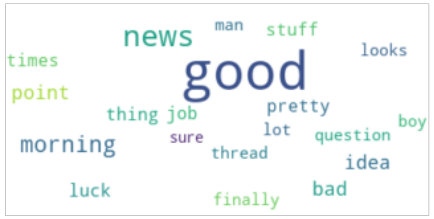


9 November 2016
Election Day







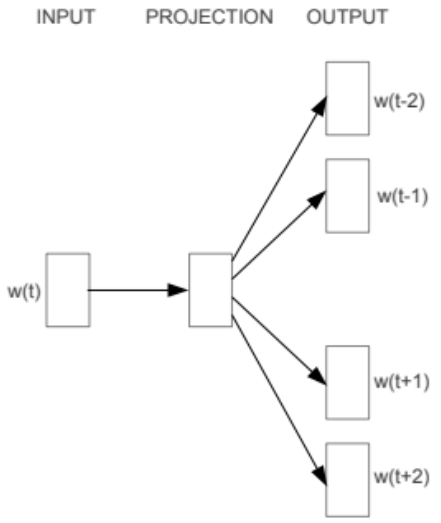


26 March 2020
Coronavirus

Topic Modeling

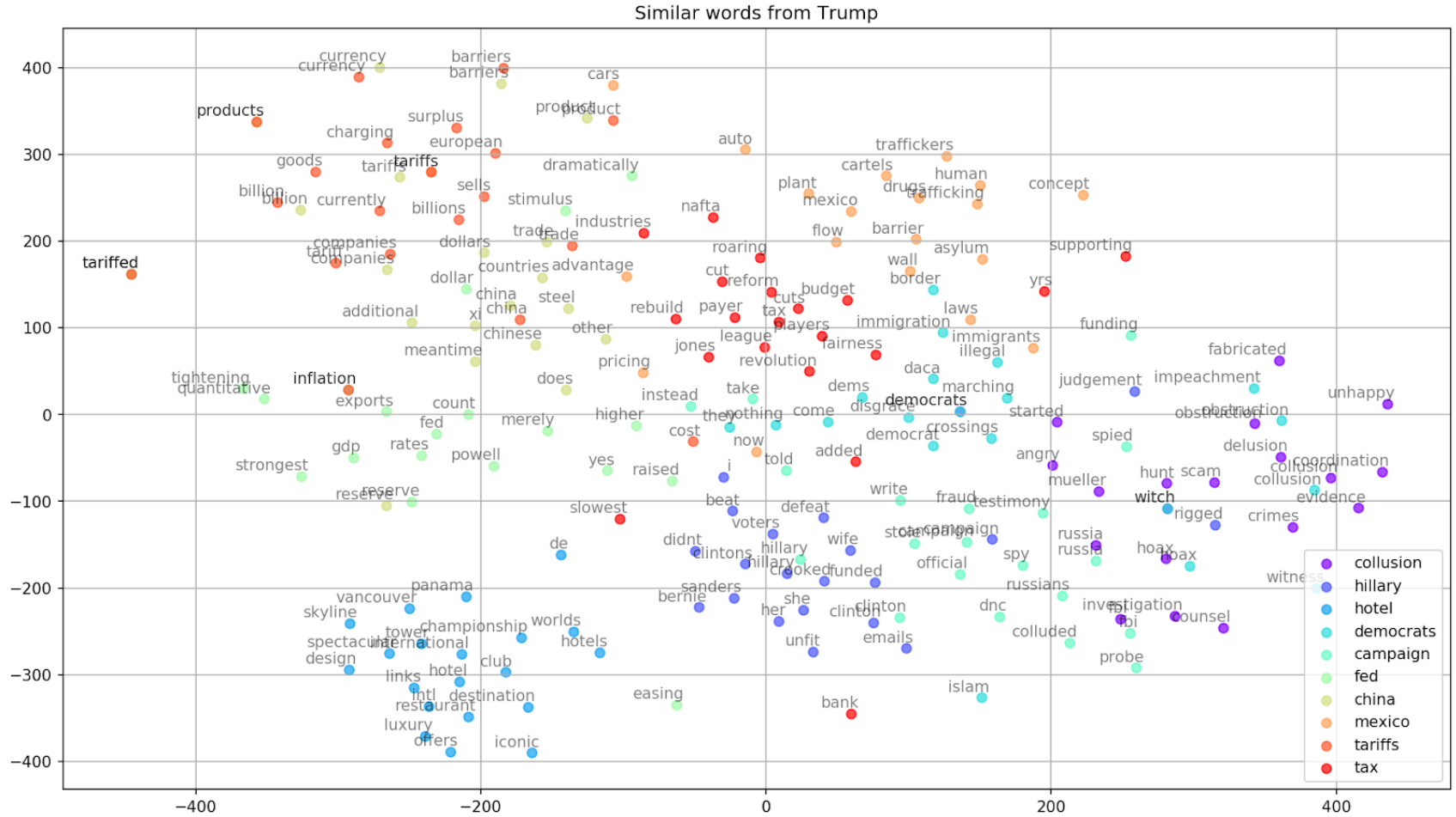
- Hand-crafted embeddings? Average of financial related words

- Train user or group-based embeddings
  - Example: Train on Trump's tweets
  - Word2Vec architecture
  - SkipGram: Predict surrounding words



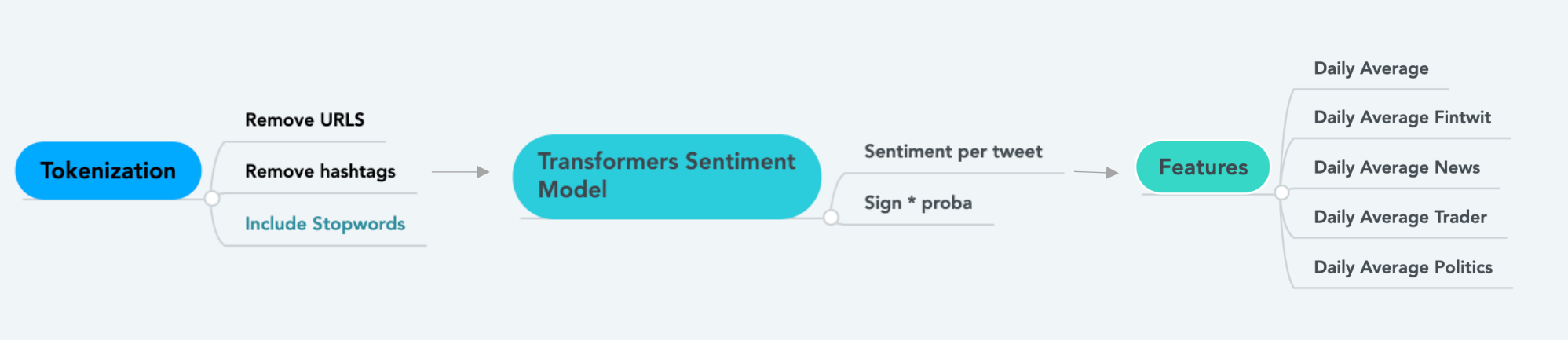
Skip-gram

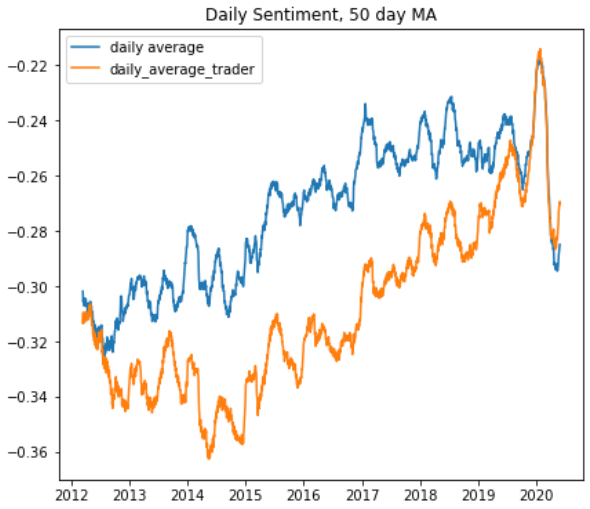
Topic Modeling

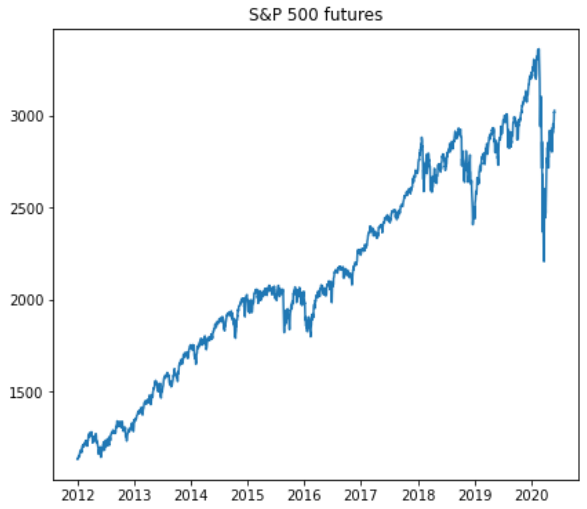


https://sites.google.com/view/trump-tweets/

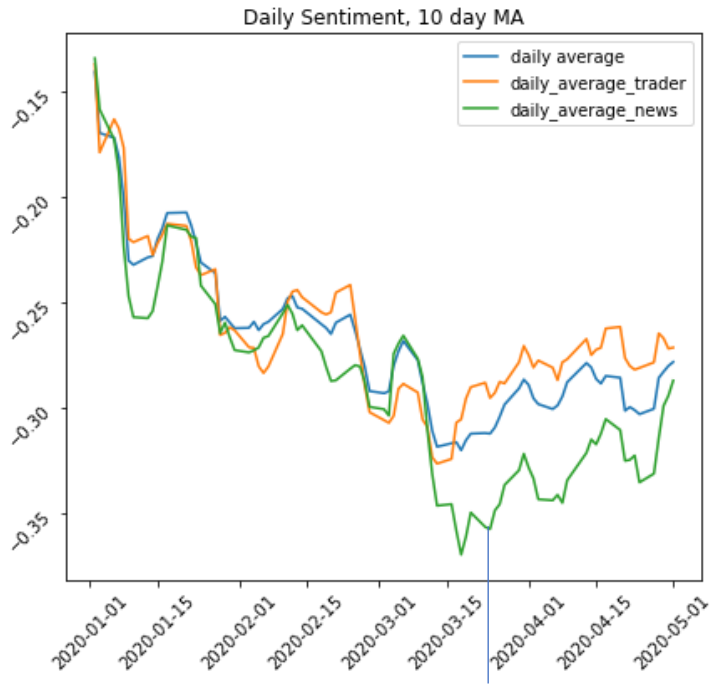
Sentiment Analysis

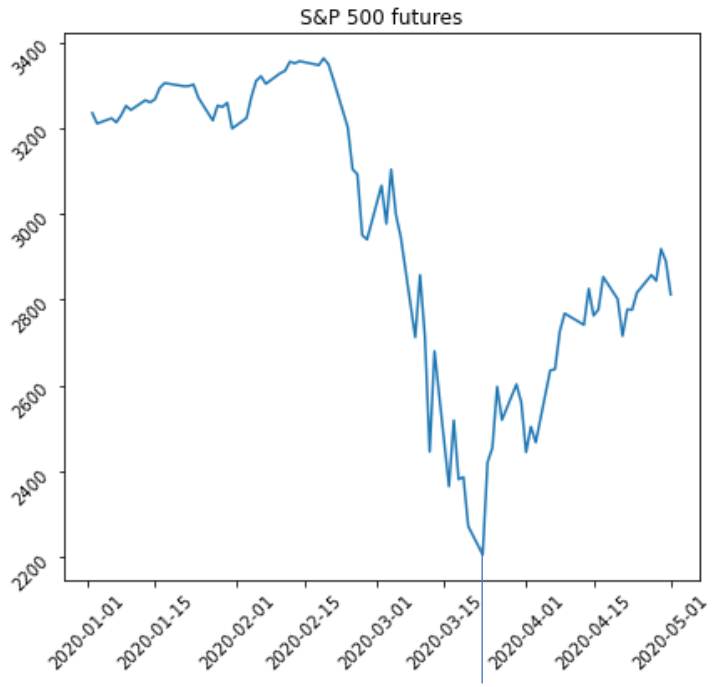






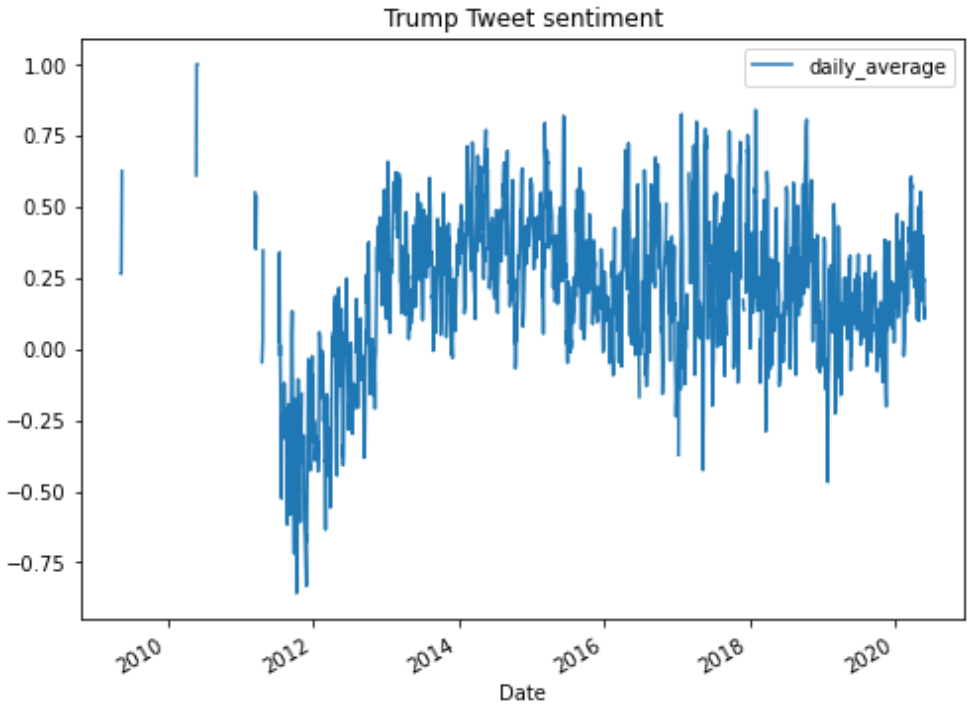
Sentiment Analysis





- Sentiment moved lower early in the year vs the market at highs
- The first sentiment to turn before the market bottom in March was traders (about a week before)
- News sentiment was the worst and bottomed 1 day before the market

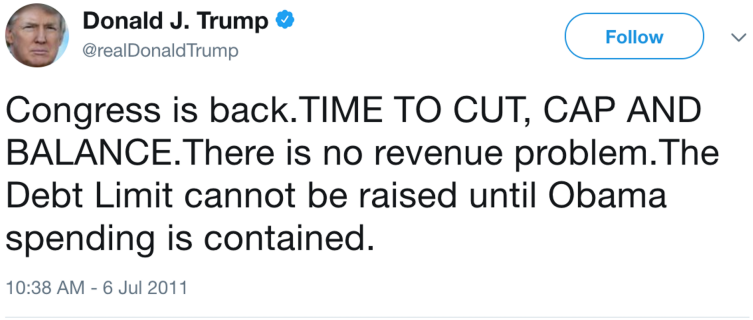
Sentiment Analysis



- **May 11, 2011**: Obama and Seth Meyers "discuss" Trump at White House Correspondents' Dinner

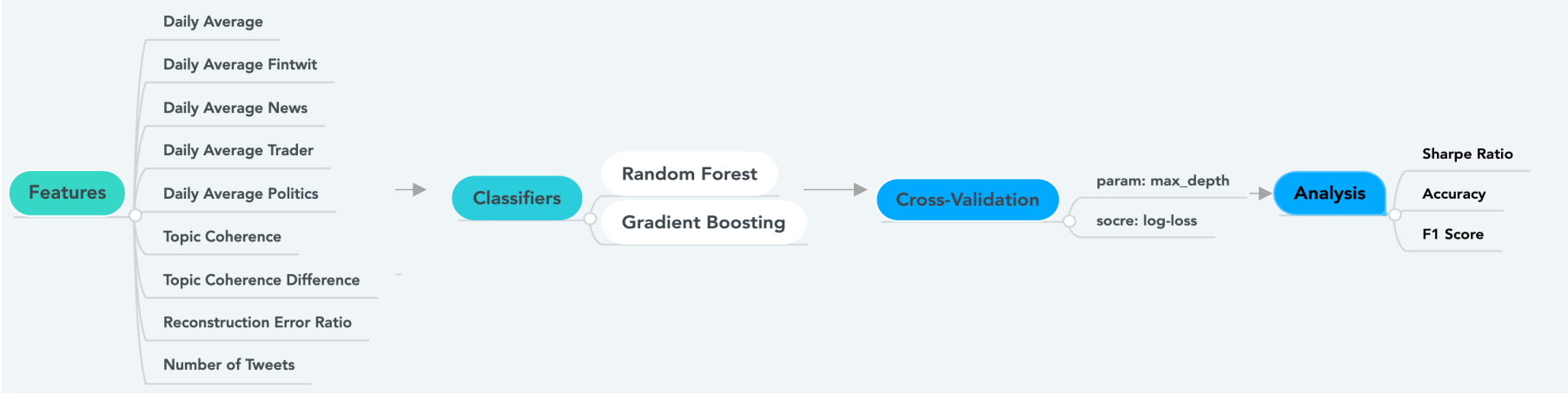
- **July 6, 2011, at 2 PM EST:** "Presidential Twitter town Hall with Jack Dorsey."

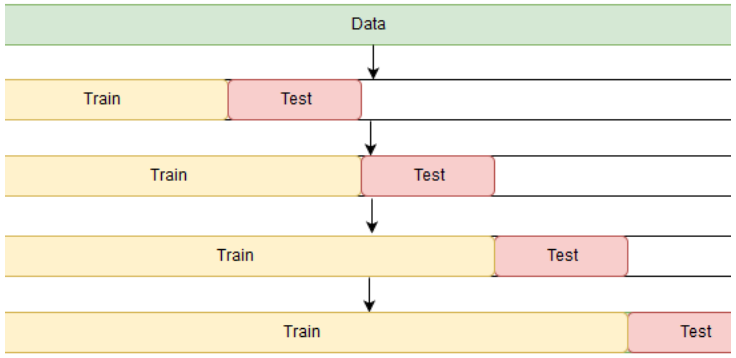
- Obama: "If you're going to communicate with the broad public, it is no longer sufficient to simply do it through traditional mainstream media."



https://theoutline.com/post/2445/trump-s-first-real-tweet-was-on-july-6-2011

Market prediction





- 9 features from the data, no market features
- Classify daily market direction: 1 if positive daily return, 0 if negative
- Train 2012-2018, test 2019 - June 2020
- 5-fold time series cross validation
- max_depth of 15, 9 for RF, GradBoost, respectively

Market Prediction

Sharpe Ratio

- Measure of a strategies' performance.
- Compare strategies with different levels of risk.
- **sqrt(252) * mean (daily return) / std (daily return)**

F1-Score

- Harmonic average of precision and recall
- A strategy that is always long the S&P will have high accuracy, 0 recall

Buying **S&P500** at the close has about **55% accuracy** historically
Logistic Regression will always choose to buy regardless of the data

Market Prediction

	Accuracy	F1	Sharpe Long Total	Sharpe Short Total	Sharpe Long Test	Sharpe Short Test	Days Long
SPX	**0.549**		0.691	0.691	0.630	0.630	1.000
RF	0.545	**0.647**	**0.809**	**0.707**	**1.347**	**1.626**	0.740
GradBoost	0.520	0.601	0.595	0.254	0.798	0.562	0.653

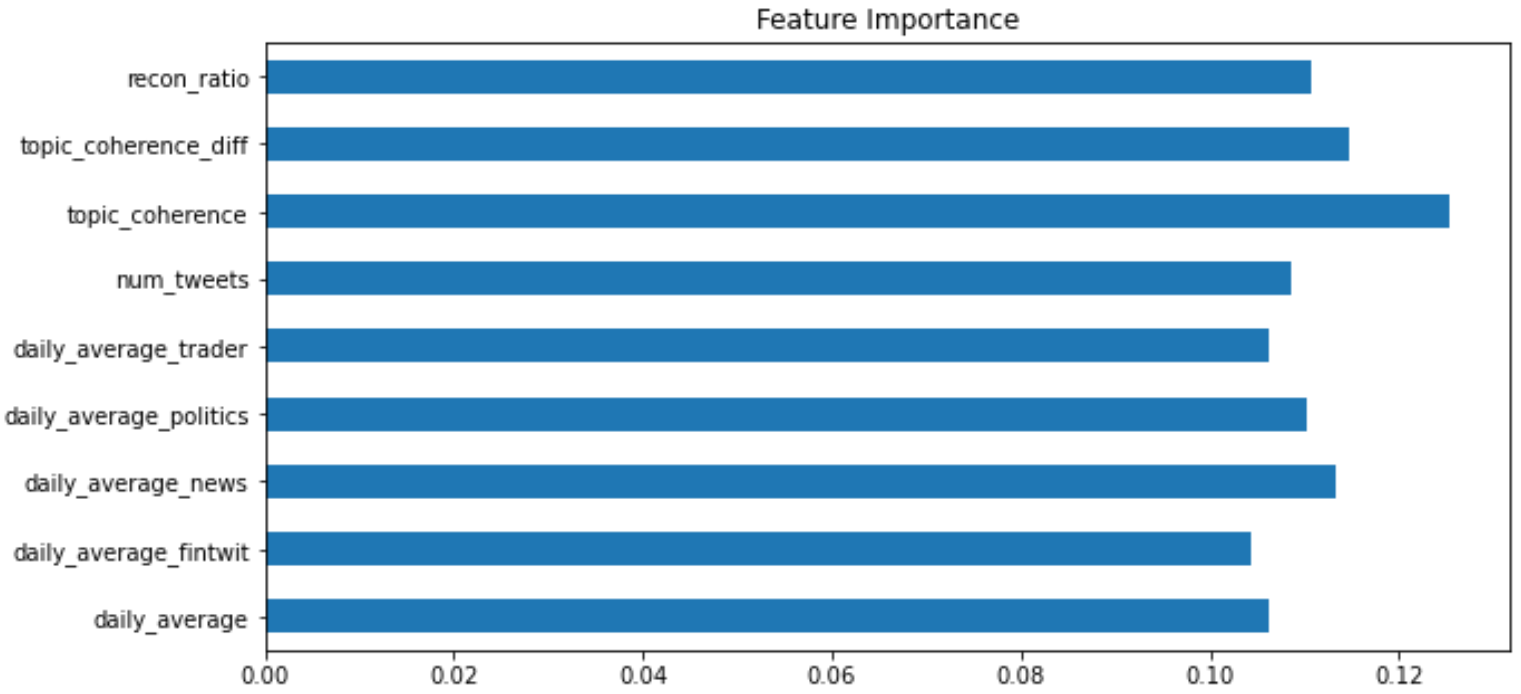
- No market features, only features from Topics and Sentiment

- **Trading Strategy Long**:
  - Buy S&P500 at the close if prediction is positive, else flat

- **Trading Strategy Short**:
  - Buy S&P500 at the close if prediction is positive, else sell S&P500 at the close

- Random Forest outperforms in almost all measures
- At similar accuracy, seems to be able to pick days with better returns
- Just noise?

Market Prediction



- All features seem to contribute in a **similar scale**
- **Topic coherence** seems the most important
- **Sentiment is split** between different categories

Work to Do

- Market Sentiment Measure
  - Transformer sentiment model is **trained on movie reviews**
  - Fine tune sentiment model to market returns
- Topics
  - Only looking at **abstract** changes in time, not looking at content
  - Compare to handcrafted embeddings, or user/group-specific embeddings
- Market
  - Introduce market features such as technical indicators, volatility
  - Explore different strategies. **Weight buys by prediction probability**.

Conclusions

- Natural Language Processing models and tools we saw in class can be very useful to understand alternative data such as Twitter

- Modeling and summarizing market relevant data from Twitter with NLP can help us understand markets better

- Promising results for detecting signal to predict market movements with these tools

Thank you!

Appendix: Features

- **Reconstruction Error Ratio, *recon_ratio***:
  - For tweets in a day, see topic model from previous day. Use it to transform data, calculate reconstruction error.
  - The feature is the ratio between the reconstruction error in the original model with the new reconstruction error minus 1. This is done to standardize for weekends
- **Topic Coherence, *topic_coherence***:
  - Average of all cosine similarities between each of 5 vectors for each topic
- **Topic Coherence Difference *topic_coherence_diff***:
  - Same as topic coherence but comparing vectors from yesterday with today

Appendix: Features

- **Number of tweets, *num_tweets***:
  - Counting number of tweets per day, *num_tweets*. The quantity is number of tweets in a day divided by the 200-day moving average of tweets in a day
  - Use MA since number of tweets has increased steadily over time
  - Since a weekend or holiday counts as a "day", need to standardize
  - Looking at the average number of tweets for 2, 3, 4, 5 days respectively
- **Daily Average Sentiment, *daily_average*:**
  - Average sentiment score over all tweets in a day
- **Daily Average *Class, daily_average_class:***
  - Average sentiment score over all tweets in a day for users classified as *Class*