# **Stockbot**
# Stock Price Prediction with Historical and Sentiment data


Eric Chan yee96@cs

Kai-Wei Chang kwchang2@cs

Andrew Wei nowei@cs


# Motivation

# The Stock Market

- Efficient Market Hypothesis
  - "… asset prices reflect all available information." – Wikipedia
  - What counts as available information?
- Stock prices fluctuate
  - "Buy low, sell high"
- How can we model stock prices?

# Sentiment

- Reflects perceptions and captures reactions in text
  - Public perceptions may reflect general trust/belief
- General positivity or negativity of text
- Can we capture sentiment associated with companies?

# Data Collection

# Financial Data

Obtained from Polygon

- Start date: January 1st, 2010
- End date: December 31st, 2019
- 15 companies
- Data
  - Open price
  - Closing price
  - High
  - Low
  - Volume

# Sentiment Data

Sentiment140 dataset on Kaggle

- 1,600,000 tweets
- Labels
  - Negative: 0 → 0
  - Neutral: 2 → 0.5
  - Positive: 4 → 1

# Scraping Tweets

- Scraped hashtag(#) and cashtag($) tweets associated with companies by stock ticker*
  - E.g. for Apple, #AAPL and $AAPL
  - ~2.5 million # tweets
  - ~1.7 million $ tweets
- Built with python
  - Using Selenium and BeautifulSoup4

* - avoiding usage collisions, e.g. KO is the stock ticker for CocaCola, but also the term for knocked out, so we looked up #CocaCola

# Companies Tracked

VISA

CHIPOTLE MEXICAN GRILL

Microsoft

UNITED

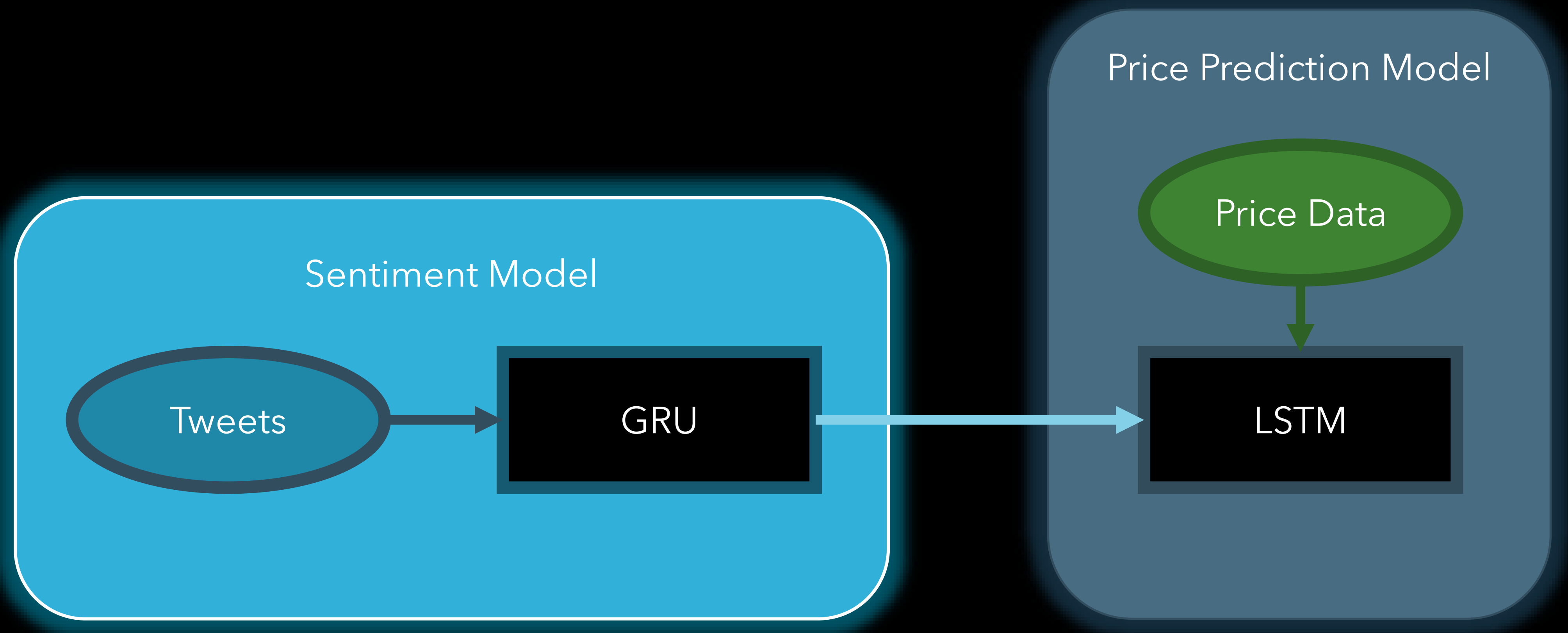
Coca-Cola

Bank of America

WELLS FARGO

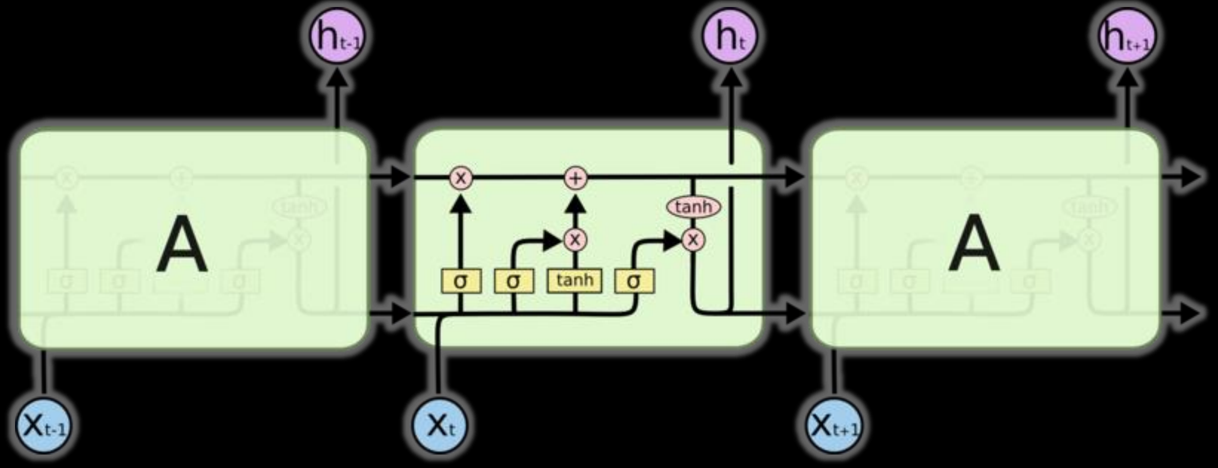
# Methods

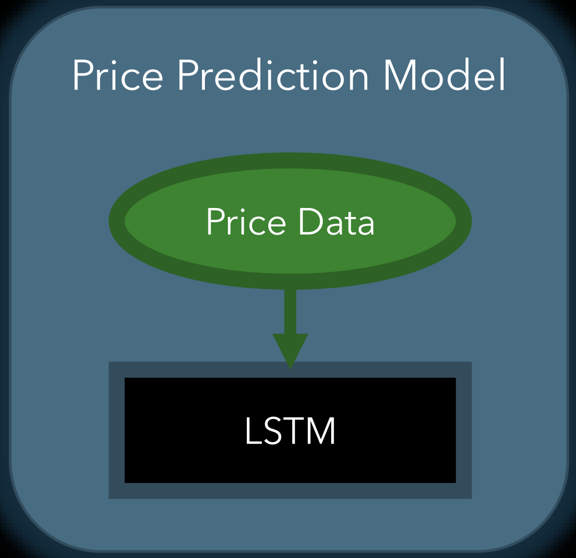
# Architecture



# Price Prediction Model

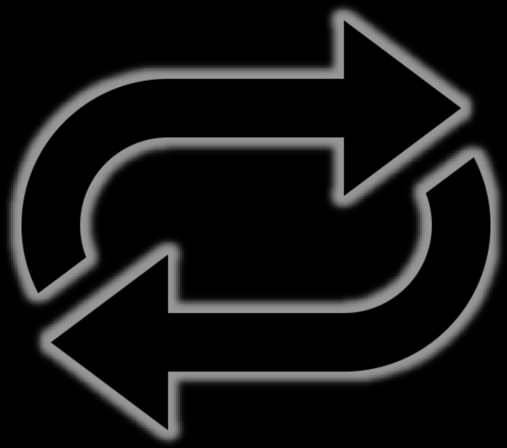
- Leverages financial data
- LSTMs
  - input dim = 5
    - open, high, low, closing price, volume
  - hidden dim = 32
  - number of layers = 2
  - output dim = 1
    - Price estimate for next date
- One model per company
- Uses previous 30 days to make a prediction

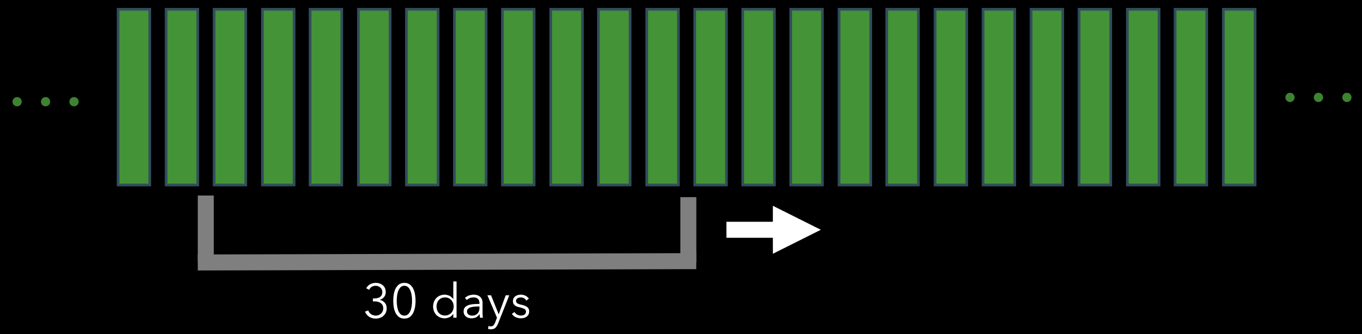


# Iterative Training

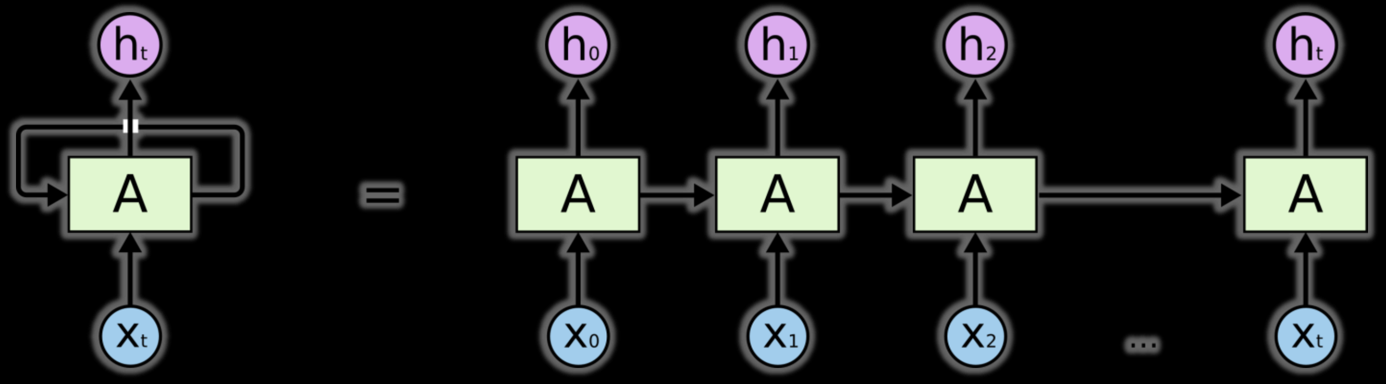
Once a prediction is made, include the actual test data point and retrain, then predict again

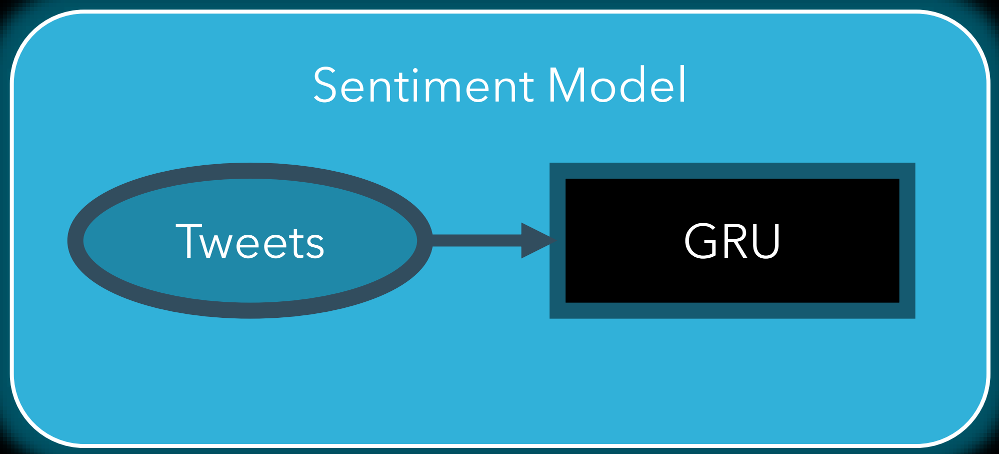




# Sentiment Model

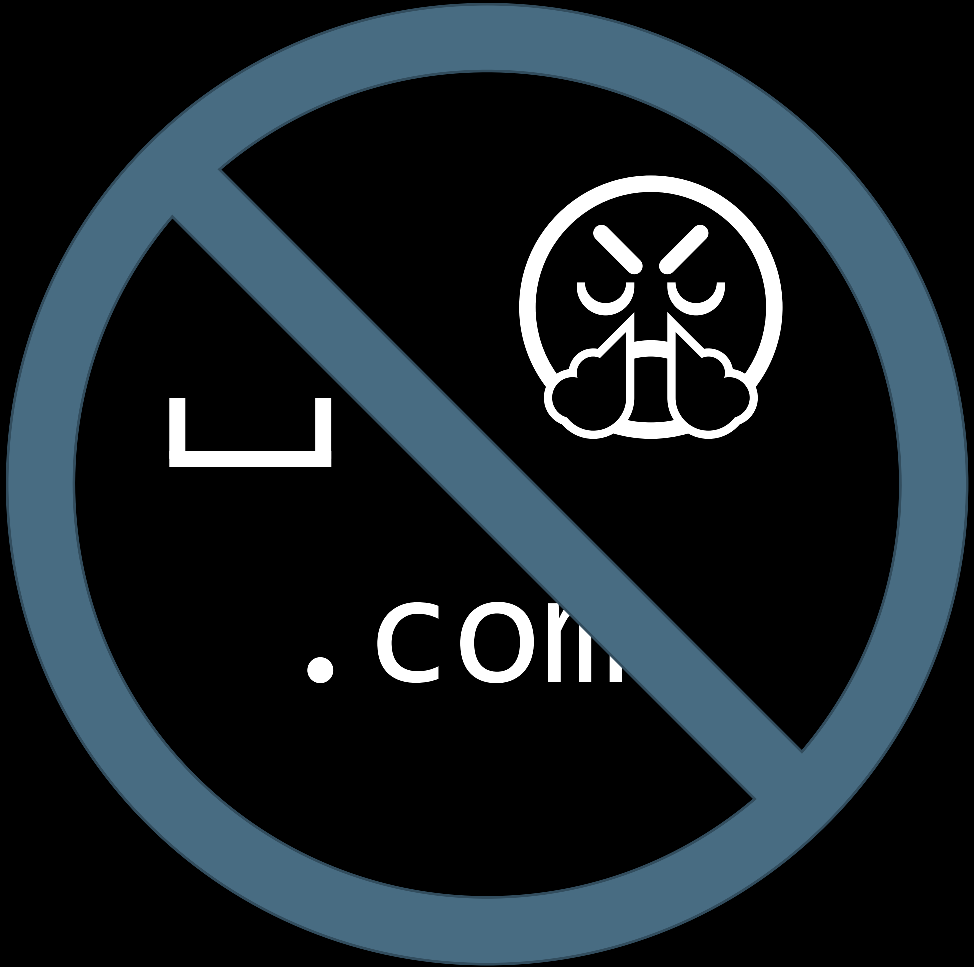
- Trained and tested on Sentiment140 dataset
- Used scraped Tweets
- GRU
  - embedding dim = 350
  - hidden dim = 350
  - number of layers = 2
  - output dim = 1
  - dropout = 0.025
  - batch size = 200



# Data Processing

- Removes:
  - Strips whitespace
  - Emojis
  - Links
- Performs UNK-ing
  - UNK probability = 0.6



# Price Change Labeling

- Labeling tweets using price changes
    - Labels need to be validated somehow
- Is there a correlation between tweet sentiment on a day and the price of the next day?

# Results

# Baseline

Simple Moving Average

- Smooths volatility
- Relatively effective in general
- Averages over 10 days, so $n = 10$

$$\frac{1}{n}\sum_{i=k}^{k+n} A_i$$

# Trend Prediction

When the price goes up or down, how often does our model predict an increase or decrease respectively?

- Roughly correct 50% of the time, but the error isn't too bad

# Subset of graphs generated

# Apple Inc. (AAPL)


# Microsoft Corporation (MSFT)


# Bank of America Corp (BAC)


# RMSE table

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\text{prediction}(i) - \text{actual}(i))^2}{n}}$$

| Ticker | AAPL | BAC | CMG | DAL | FB | GOOG | JPM | KO | LUV | MCD | MSFT | PEP | UAL | V | WFC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No Iter (Test) | 11.36 | 0.52 | 13.43 | 0.96 | 3.49 | 18.64 | 2.07 | 0.96 | 1.02 | 7.45 | 16.46 | 2.31 | 1.99 | 13.80 | 0.70 |
| **SMA** | **3.80** | **0.63** | **17.70** | **1.34** | **4.20** | **24.41** | **1.88** | **0.69** | **1.29** | **2.22** | **1.52** | **1.54** | **2.24** | **1.78** | **1.14** |
| Iter (Test) | 4.60 | 0.47 | 11.76 | 0.87 | 3.45 | 18.29 | 1.68 | 0.52 | 0.91 | 2.37 | 2.28 | 1.36 | 1.41 | 2.73 | 0.69 |

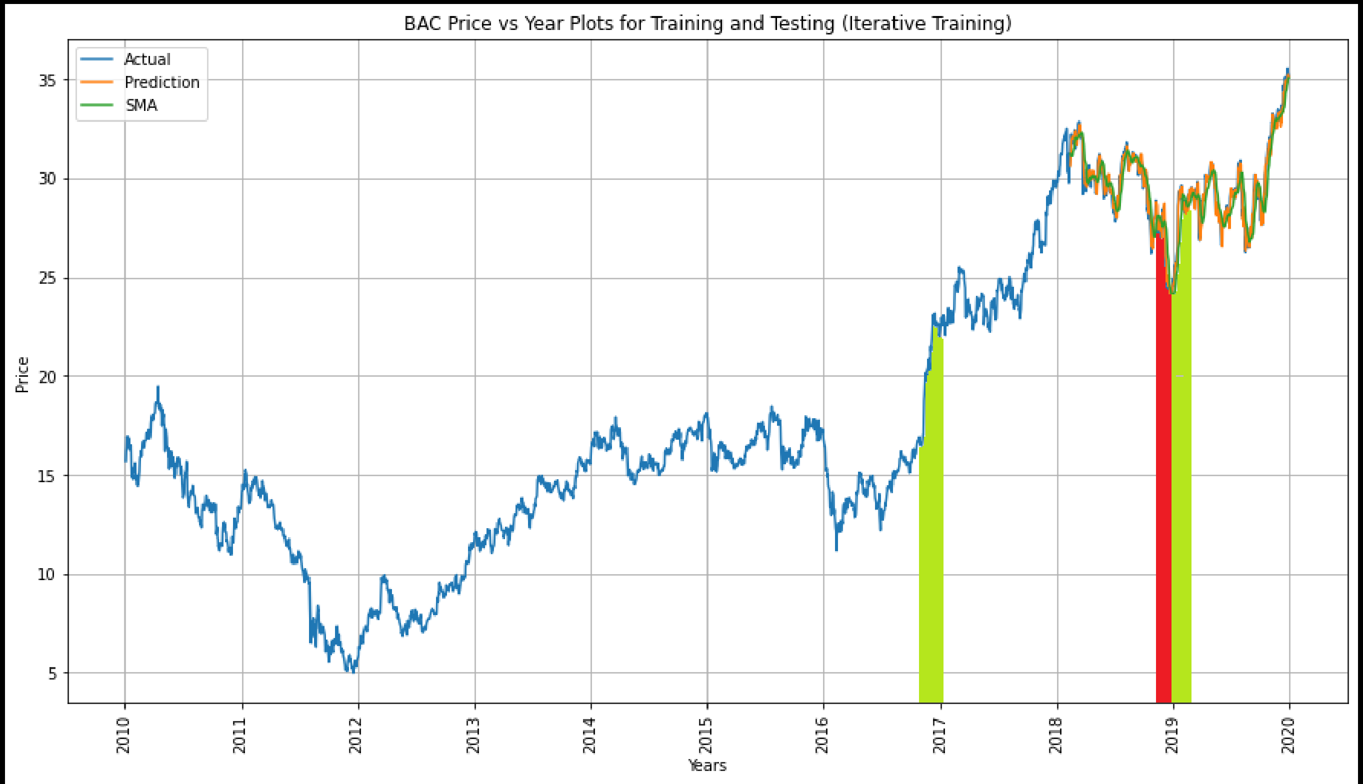
= Lower RMSE    = Higher RMSE

# Sentiment results

- Trained on Sentiment140
  - Train accuracy: 89%
  - Test accuracy: 88%
- We predicted the sentiment of scraped tweets
  - Give neutral rating if no tweets on the day
  - Otherwise give average sentiment score for that day

# Sanity check

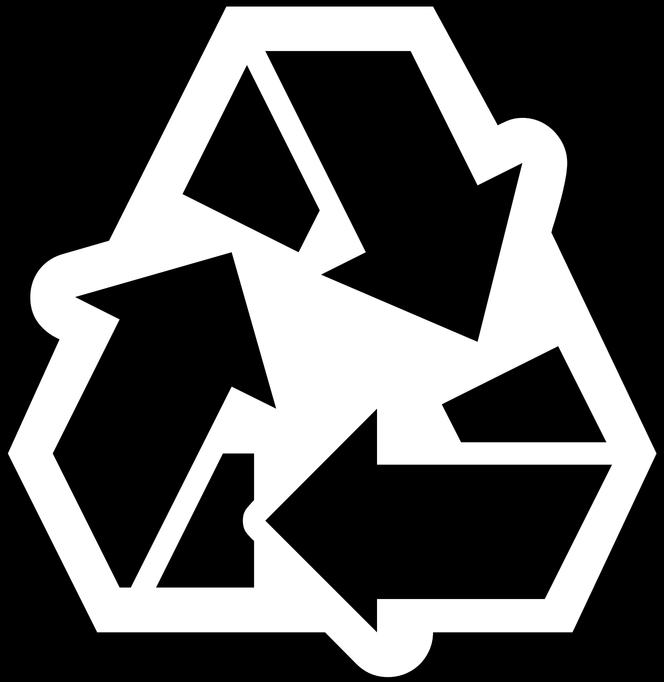
- We found that the predictions performed worse when we included them
- To sanity-check our model, we checked regions of increase and decrease for sentiment for BAC and found that they were all generally ~0.54, i.e. slightly positive



# Discussion

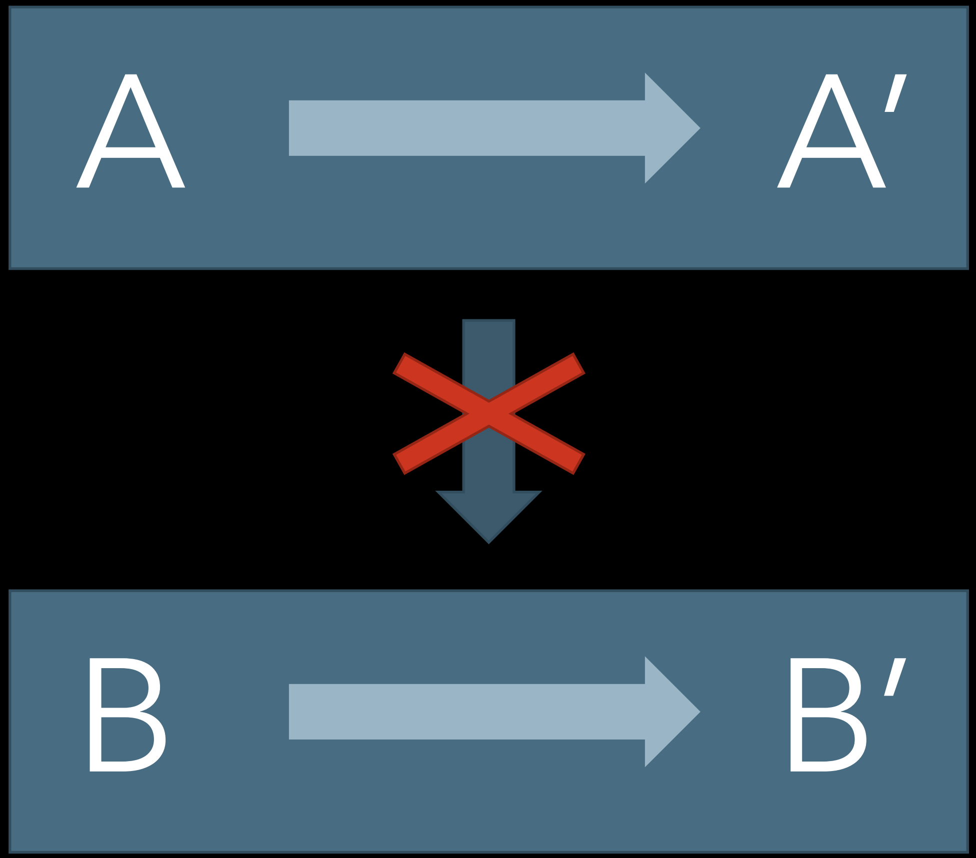
# Iterative Training

- Does it make sense?
  - Data is limited
  - Can’t generate new data for the past
- Not aiming for generalization



# Sentiment Generalization

- Didn't generalize very well
- Trained setting differs from applied setting
- Will likely perform better if we have more relevant training data



# Future Work

- Labeling the collected tweets
  - Training new sentiment model on labeled tweets
- Predicting up-to-date stock prices
- Test out the predictions with our own money

# Conclusion

Contributions

- Price Prediction model
  - Iterative training
- Sentiment model
- Scraped tweets for 15 companies stock tickers
  - January 1st, 2010 → December 31st, 2019