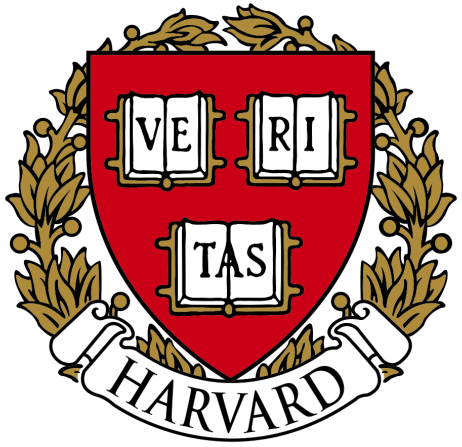
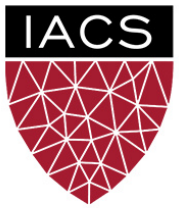


News Sentiment Analysis for Stock Return Prediction

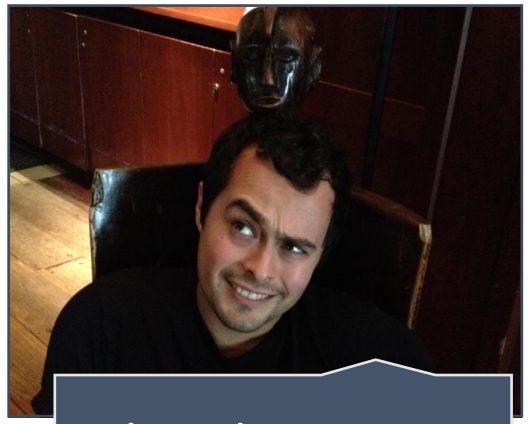
***αβnormal* Distriβution**

AC295/CS115

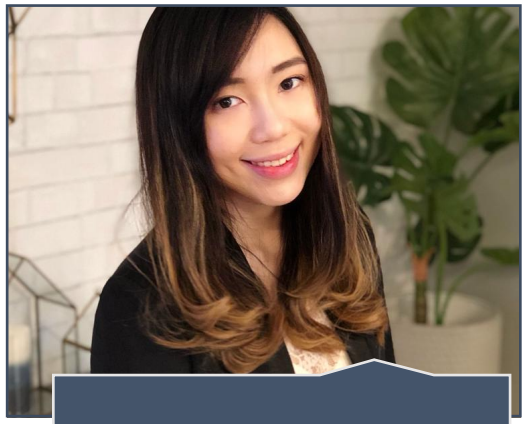
Eduardo, Jessica, Rohit & Stuart



$\alpha\beta normal$ Distriβution



Eduardo Peynetti

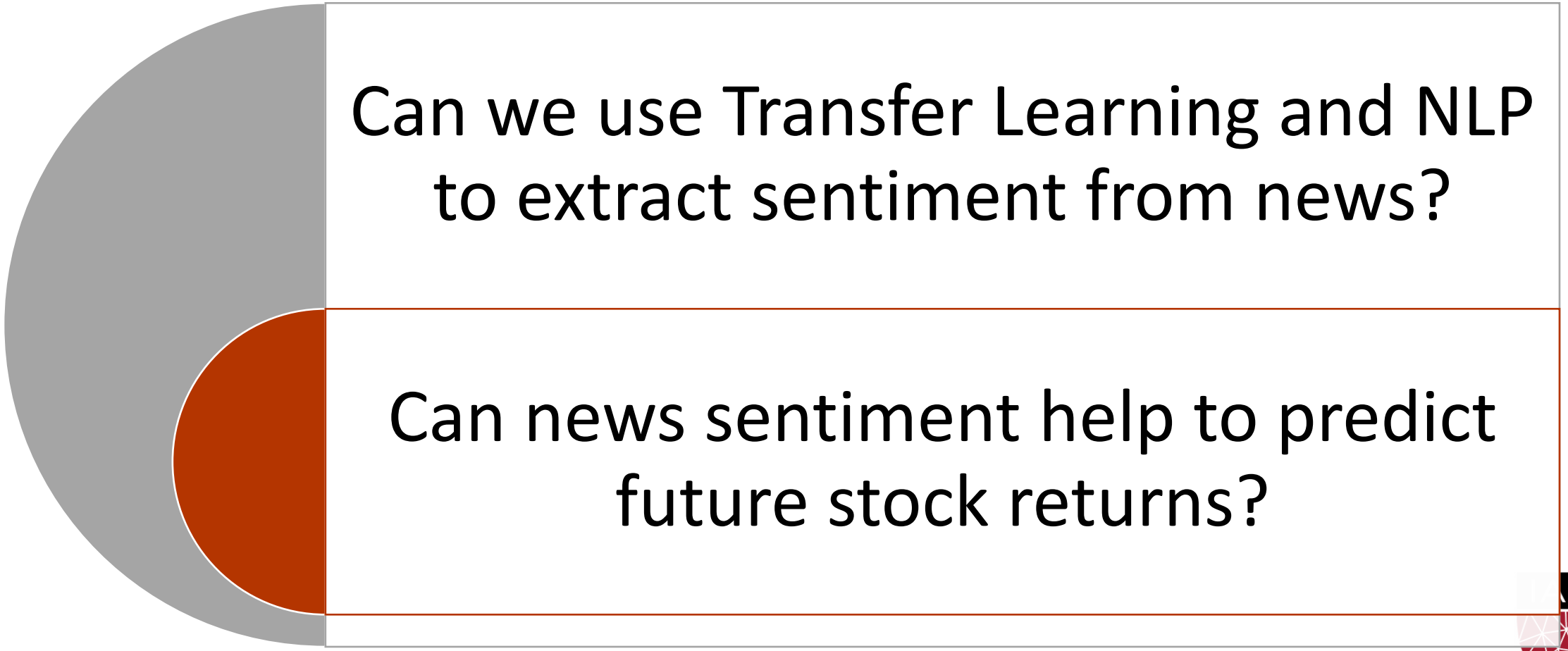


Jessica Wijaya

Rohit Beri

Stuart Neilson

Problem Statement



Datasets

27 million+ News Summaries
- 2.7 Million selected

3.3 million+ New Summaries
- 1.0 Million selected

Key Developments for S&P500 Companies
- 900K+ press releases and event announcements

Sharadar Financial Data
- Stocks Returns

10Ks for S&P500 Companies
- 6000+ documents with focus on Risk Factor, MD&A, and Market Risk

Baseline Models

Loughran McDonald

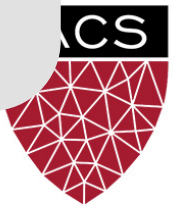
- Dictionary based model that analyzes word frequencies
- TF-IDF scores using positive and negative words

BERT for Sentiment Analysis

- Pre-trained BERT model
- Trained on SST/2, baseline sentiment dataset

Fin-BERT for Sentiment Analysis

- Language Model trained on financial news
- Sentiment Model trained on pre-labeled phrases (+/=/-)



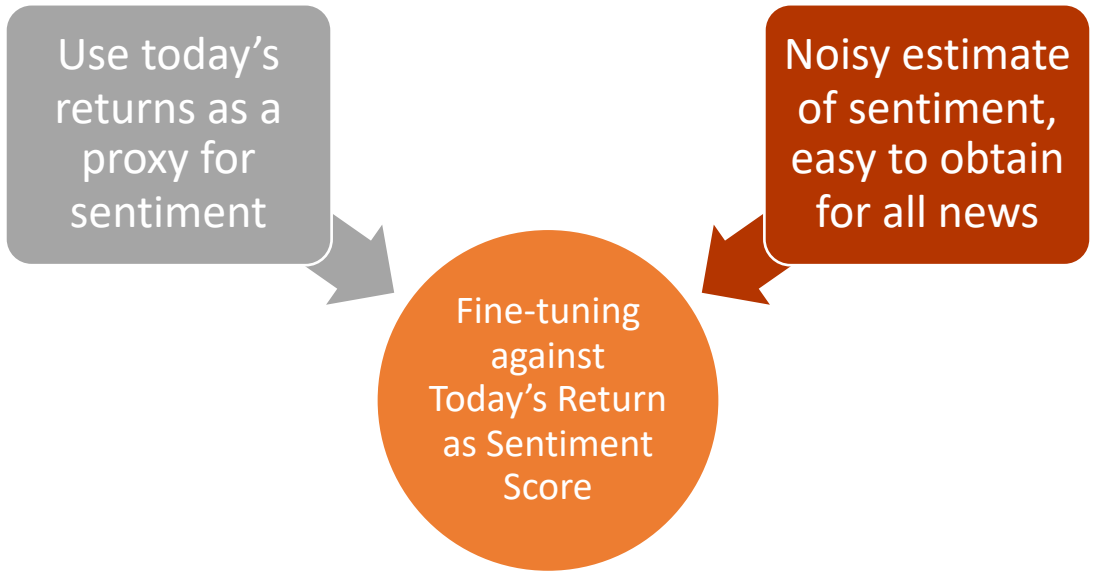
Our model

Fine-tuned Fin-BERT

Fine-tuned by using todays returns as a target

-1/+1 for on positive/negative returns

0 when news doesn't have a word in LM dictionary



IACS

Performance Evaluation

Long-Short Portfolio Construction

Every day:

Estimate average sentiment for each stock which had news

Buy stocks with positive sentiment, sell stocks with negative sentiment

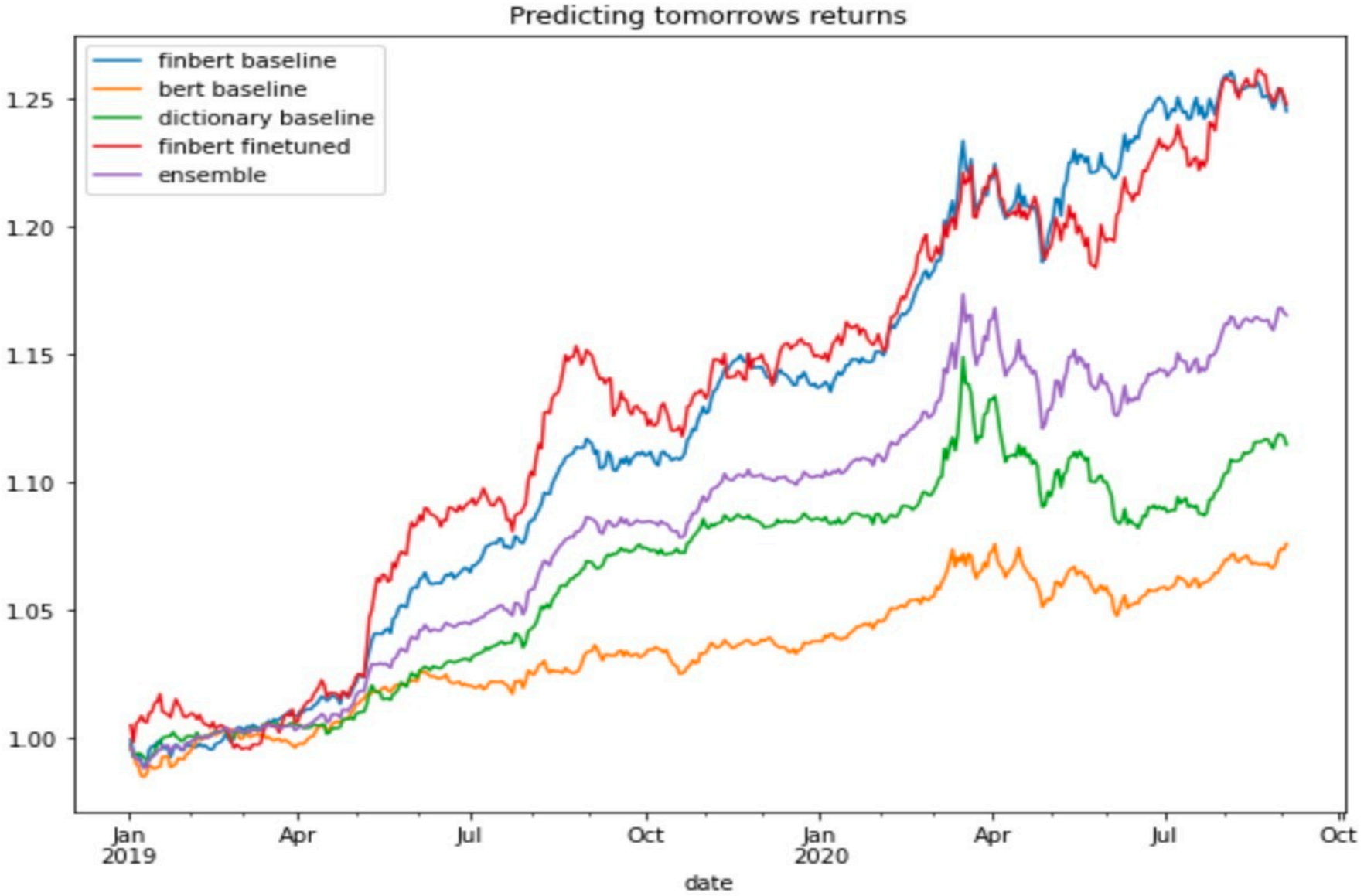
Performance Measure

All models have accuracies around 51-52%. Wrong measure!

Use portfolio returns as performance measure

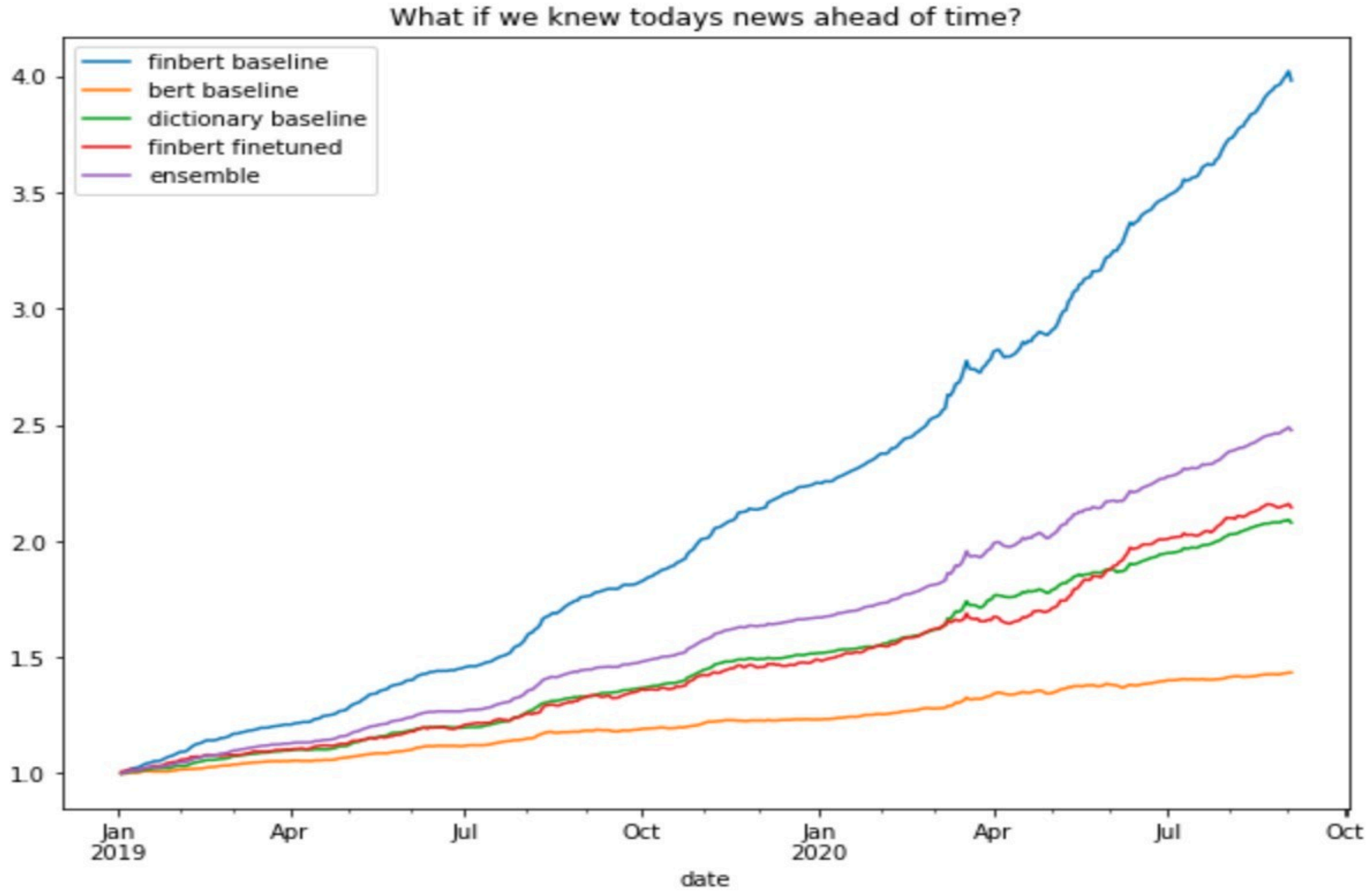
Performance Evaluation

Predicting next day's returns



Sentiment Extraction

If we knew today's news beforehand?



Sharpe Ratio

$$\frac{\bar{\mu}_{Returns}}{\sigma_{Returns}}$$

A way to standardize returns across stocks with different volatilities

Units of reward expected per unit of risk

LM baseline	BERT base	FinBert base	FinBert Fine-Tuned
1.75	1.6	**3.05**	2.45

IACS

Decay in return predictability

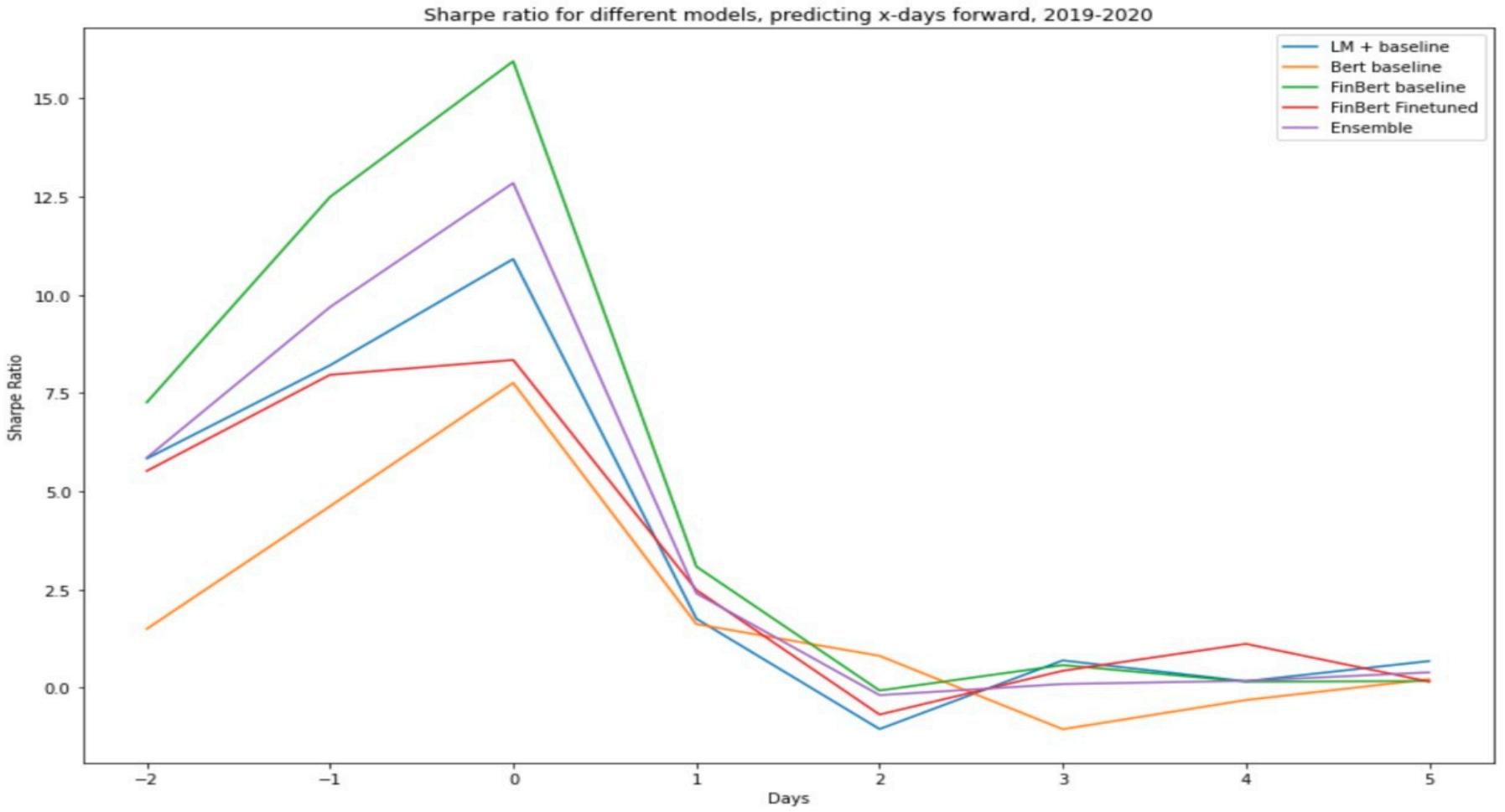
Todays news are predictive of yesterday's returns!

Old news / market knew the news before-hand

Small but persisent effect of news into the future

Sharpe Ratio for different models

Predicting x days forward



BERT vs. Fin-BERT – Negative ("oil prices fell")

BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
negative	positive (0.69)	positive	1.73	[CLS] oil prices fell on tuesday amid concerns that a possible rise in co ##vid - 19 cases following the u . s . labor day long weekend , which also marks the end of the peak u . s . driving season , could squeeze demand for fuel . [SEP]

Fin-BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
negative	negative (1.00)	negative	-1.48	[CLS] Oil prices fell on Tuesday amid concerns that a possible rise in Co ##vid - 19 cases following the U . S . Labor Day long weekend , which also marks the end of the peak U . S . driving season , could squeeze demand for fuel . [SEP]

IACS

BERT vs. Fin-BERT – Positive (shares "doubled")

BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (0.66)	positive	1.22	[CLS] shares of the cpu and gp ##u developer have more than doubled in 2019 , as wall street gives a thumbs - up to its product launches and share gains . [SEP]

Fin-BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (1.00)	positive	1.09	[CLS] S ##hare ##s of the CPU and GP ##U developer have more than doubled in 2019 , as Wall Street gives a thumbs - up to its product launches and share gains . [SEP]

IACS

BERT vs. Fin-BERT – Positive ("signed a large deal")

BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (0.64)	positive	0.45	[CLS] bas ##ware corporation stock exchange release august 31 , 2010 at 16 : 25 bas ##ware signed a large deal with an international industrial group bas ##ware will deliver in ##vo ##ice automation solution and connectivity services to an international industrial group [SEP]

Fin-BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (1.00)	positive	1.81	[CLS] Ba ##s ##ware Corporation stock exchange release August 31 , 2010 at 16 : 25 Ba ##s ##ware signed a large deal with an international industrial group Ba ##s ##ware will deliver In ##vo ##ice Auto ##mation solution and Con ##nect ##ivity Services to an international industrial group [SEP]

IACS

Challenges

Financial Data is noisy!

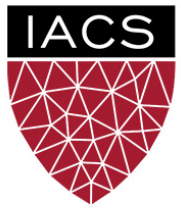
- Wrong timestamps mean fake performance
- Repeated news across datasets

Long training and evaluation time!

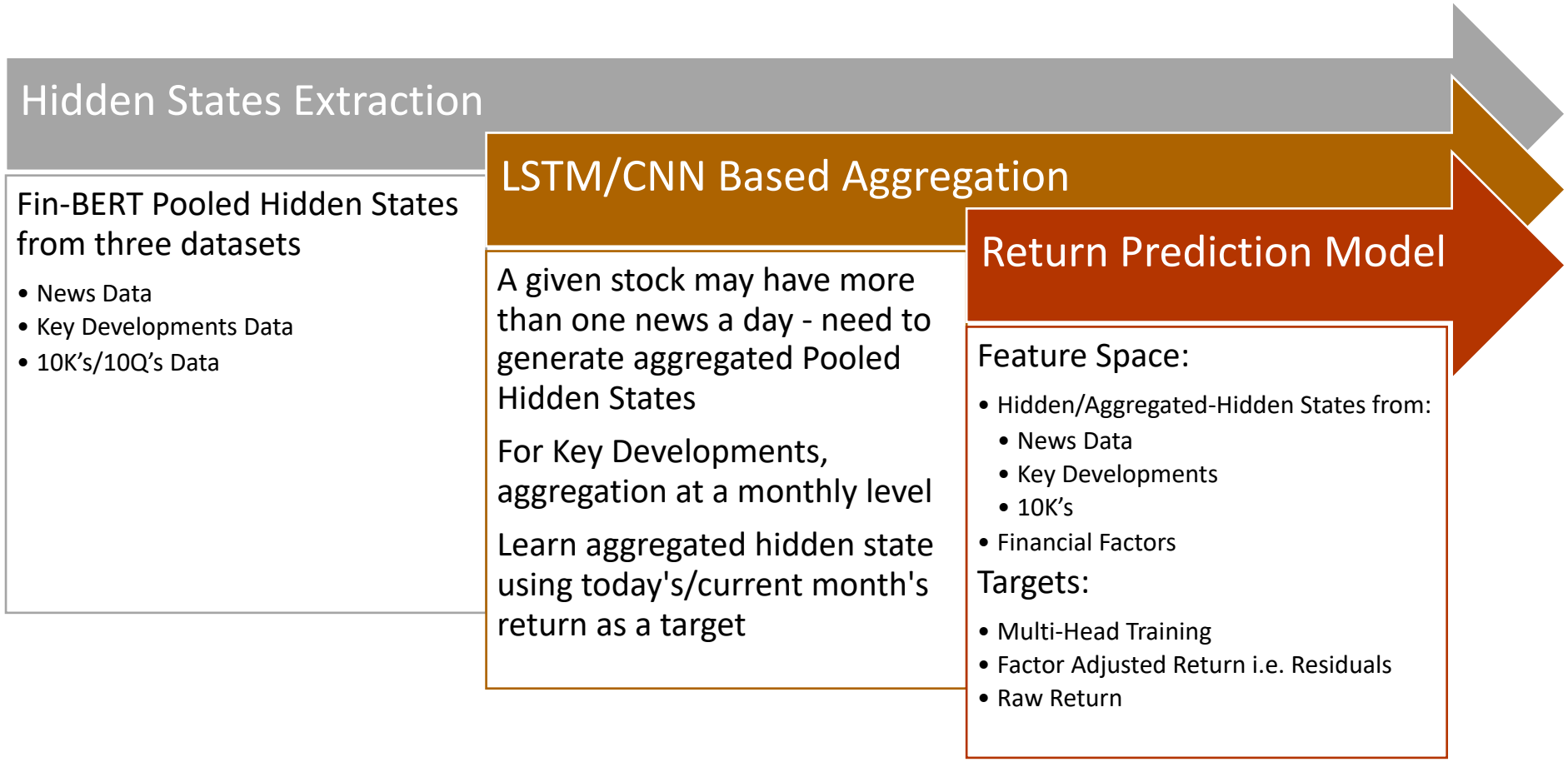
- Need prediction for all news in test set, over 2 million forward passes in BERT
- One lonely news in a small stock might dictate performance

Targets are noisy!

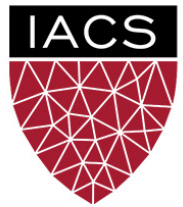
- Returns are just a soft measure of sentiment
- Easy to see model “forget” while training



Work in Progress



Thanks!



Data Storage

Google Drive

- Collaboration
- Mounts in Colab
- Unlimited free storage with Harvard email

mongoDB Atlas

- mongoDB cluster in cloud
- Efficient for text storage and retrieval

IACS

Data Management

Raw text data

- Meta data
- Processed feature vectors
- **Cluster level pre-processing**

Financial and returns data

- Efficient integration with zipline – financial modelling library by Quantopian

TensorFlow model pipelines

- **Storing preprocessed data for efficient training**

Intermediate storage

- JSON
- CSV
- PKL

IACS

Data Preprocessing

News/Text Data

- Sentiment Score from BERT
- Pooled hidden layer representation from BERT Model

10K's

- Extracting relevant items from 10K's – Risk Factors, MD&A, Market Risk
- Summary for each item using LSA due to long length – up to 150K words per item

Financial Data

- Financial data from Quandl ingested into Zipline for performance analysis

IACS

BERT vs. Fin-BERT – Positive

(“announced a multi-year agreement”)

BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (0.63)	positive	-0.33	[CLS] terra l ##y ##cos the global internet group , and google inc . developer of the largest performance - based advertising program , announced a multi - year agreement making context ##ually - targeted advertisements through the google ads ##ense (t ##m) program available on selected sites throughout the terra l ##y ##cos network . [SEP]

Fin-BERT

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
positive	positive (1.00)	positive	2.23	[CLS] Terra L ##y ##cos the global Internet Group , and Google Inc . developer of the largest performance - based advertising program , announced a multi - year agreement making context ##ually - targeted advertisements through the Google Ad ##S ##ense (T ##M) program available on selected sites throughout the Terra L ##y ##cos Network . [SEP]

IACS