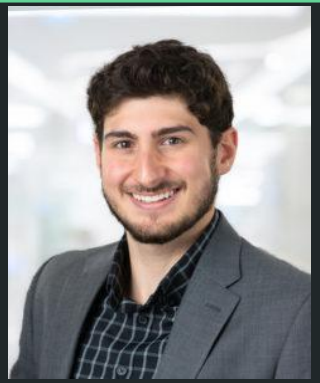
Company Sentiment Analysis and Performance Modeling



Team Members

Alexander Bell
MS in Computer Science

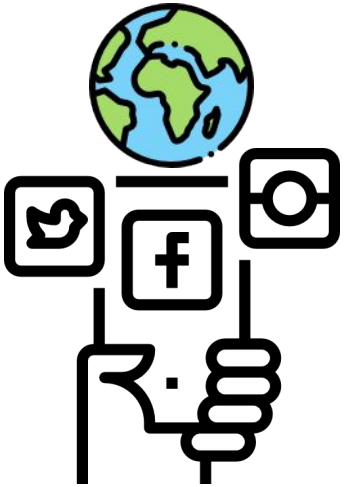
Talia Andrews
BS/MS in Computer Science

Nicholas Chantre
MS in Computer Science

Richard Plunkett
MS in Applied Stats

Motivation and Background

Project Motivation



Social Media Use: 62.3%

Social media is an echo chamber.

Examples: GameStop and Twitter

GameStop's "Twin Peaks" (High Intra-day Share Price – Jan 4 2021 to Mar 12 2021)

➔ Optimistic Reddit posts lead to surge in GameStop stock price

➔ Hate speech on Twitter leads to lower ad revenue

Project Goals

- ➔ Understanding the relationship between company performance and social media
- ➔ Using social media to predict future performance

Performance Metric

- ➔ Stock prices
  - ◆ January-December 2022
  - ◆ Target variable: Close Price

Related Work

Social Media Sentiment Analysis

- Nyugen, H., Calantone, R., & Krishnan, R. (2019). Influence of Social Media Emotional Word of Mouth on Institutional Investors' Decisions and Firm Value. Management Science, 66(2).
- Butt, S., Sharma, S., Sharma, R., Sidorov, G., & Gelbukh, A. (2022). What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses. Computers in Human Behavior, 107345.

Project Pipeline

Data Acquisition

Retrieving:

- ➔ Social media posts
- ➔ Amazon stock prices

Data Preprocessing

- ➔ Filtering posts for Amazon-related content
- ➔ Cleaning post content

Sentiment Analysis

- ➔ Assigning sentiment to each social media post

Merging Datasets

- ➔ Sentiment data + stock prices

Performance Modeling

- ➔ Modeling the relationship between sentiment and stock price

Data Collection

id	user	title	score	date	url	body
806	u/[deleted]	Best...	23	12/31	https://www...	Upvotes for non...
1053	u/AutoModerator	Januar...	29	12/31	https://www....	***Bold***the...

Modules/Libraries Used

pandas

scikit learn

Hugging Face

BERTopic

fast.ai

Data Preprocessing

- Cleaning the text body of posts
  - Removing stop words
  - Removing special characters
  - Truncating/padding to max length

Sentiment Analysis

Electra Results

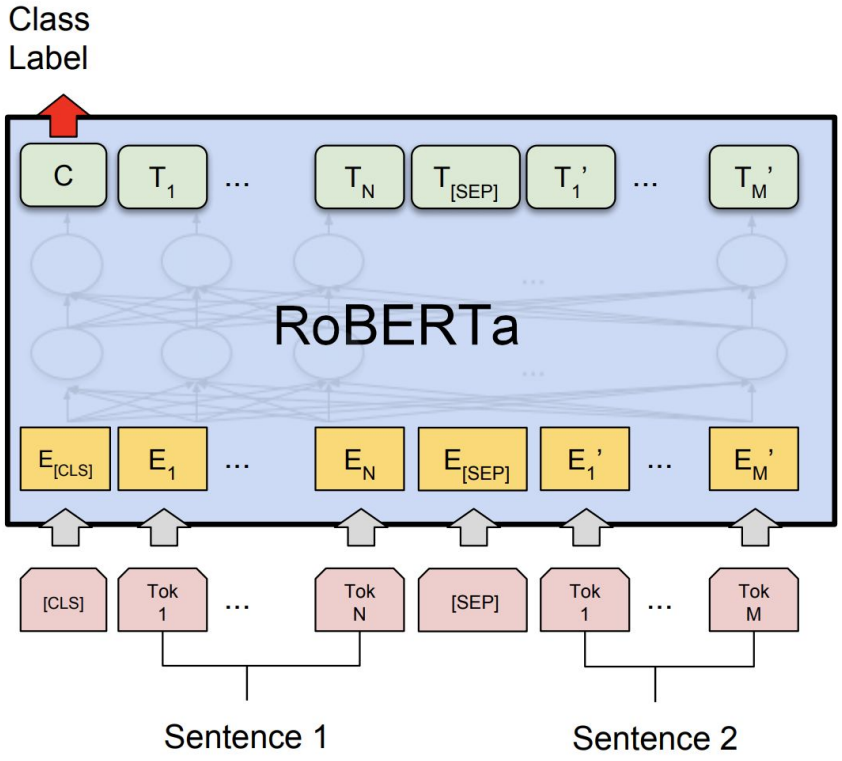
- ➔ Amazon Labeled Dataset
- ➔ (500 positive vs 500 negative)
- ➔ Accuracy: 20 %

```python
import pandas as pd
import torch
from transformers import AutoTokenizer, AutoModelForSequenceClassification

model_name = "google/electra-base-discriminator"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name)

#Opening csv with reddit data retrieved from API
#df = pd.read_csv(path + "RS_2022_Amazon_with_subreddits_no_stopwords.csv")
df = pd.read_csv(path + "amazonLabeledDataSets.csv")
```

```python
[ ] def electra_sentiment(text):
        max_length = 1200
        if len(text) > max_length:
          text = text[:max_length]

        encoded_text = tokenizer(text, return_tensors='pt', truncation=True, max_length=512)
        outputs = model(**encoded_text)
        predictions = outputs.logits
        probabilities = torch.softmax(predictions, dim=-1)
        return probabilities[:, 1].item()

    def electra_sentiment2(text):
        # Tokenize the text
        encoded_text = tokenizer(text, return_tensors='pt', truncation=True, max_length=512)

        # Get predictions from the model
        outputs = model(**encoded_text)
        predictions = outputs.logits.argmax(dim=-1)

        # Return the predicted sentiment label
        return predictions.item()
```

RoBERTa/SiEBERT Results

- 93.2% average accuracy across 15 datasets
- Reddit Labeled Dataset
  - Accuracy: 78.52%
- Amazon Reviews Labeled Dataset:
  - Accuracy: 46.8%



https://www.labellerr.com/blog/roberta-a-robustly-optimized-bert-pretraining-approach/

XLNet Results

Pretrained model based on Cornell Sentiment:

Reported Loss: 0.3771

Reported Accuracy: 0.8833

Reported F1: 0.8793

Tested Accuracy: 0.8110

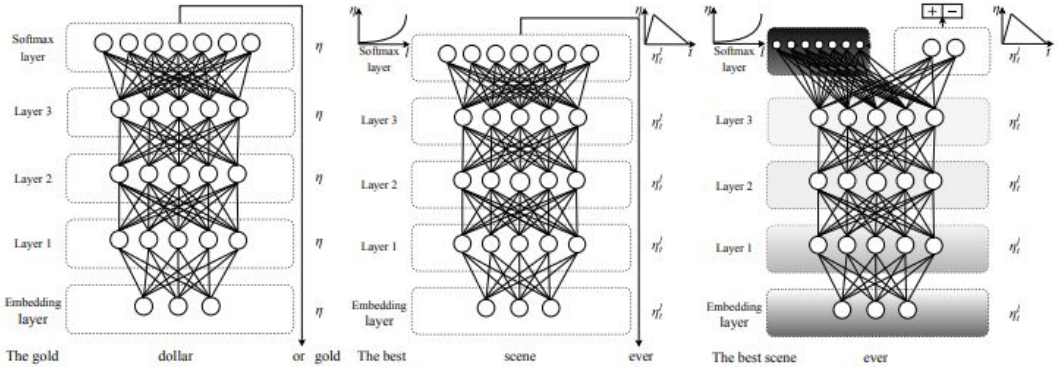
Tested F1: 0.8915

Tested Precision: 0.7716

XLNet trained on Cornell sentiment, tested on Amazon reviews		
	Predicted Positive:	Predicted Negative:
Actual Positive:	429	62
Actual Negative:	127	382

ULMFiT Results

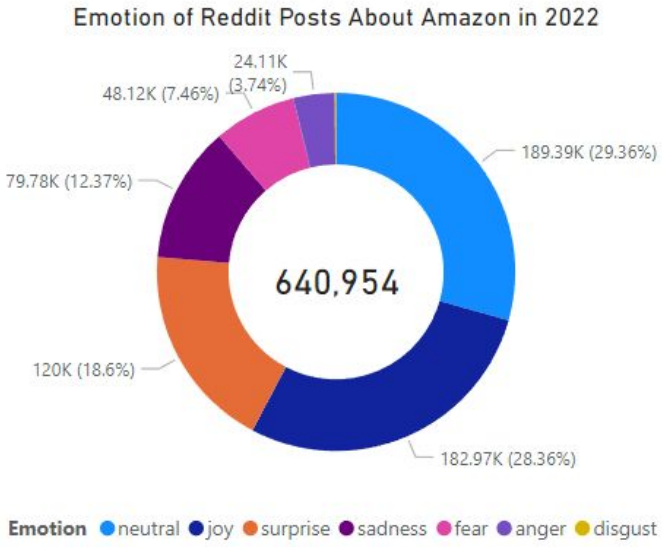
- Chosen for:
  - Power on small amounts of labeled data
  - Performance on limited hardware
- Test set:
  - Accuracy: 77.6%
  - Precision: 84.3%
  - Recall: 87.1%
  - F1 Score: 85.7%



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Distil-RoBERTa Results

- Chosen for:
  - Availability of pretrained model for emotion classification
  - Previously used in social media analyses
- Used to assign emotion to 640K Reddit posts

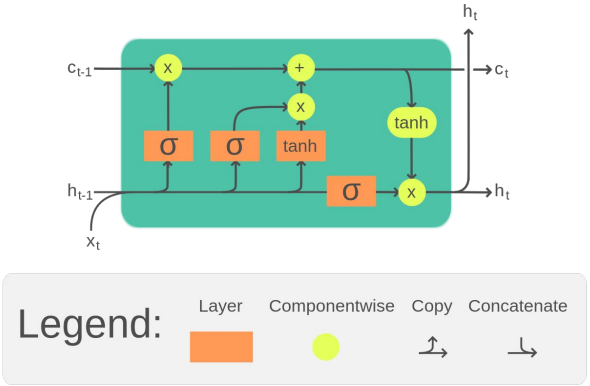


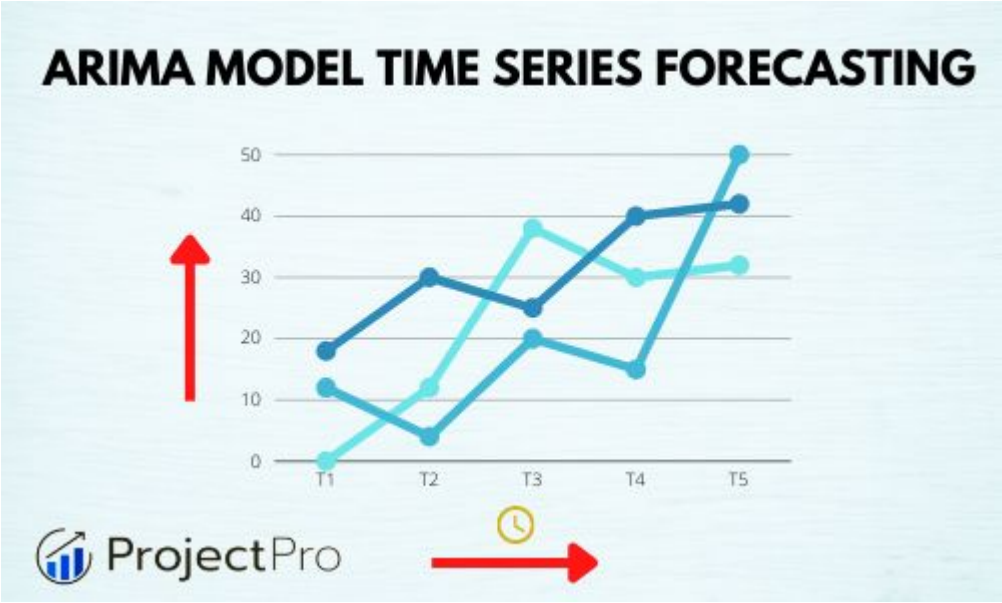
Model: J Hartmann/Emotion English DistilRoBERTa Base

Performance Modeling

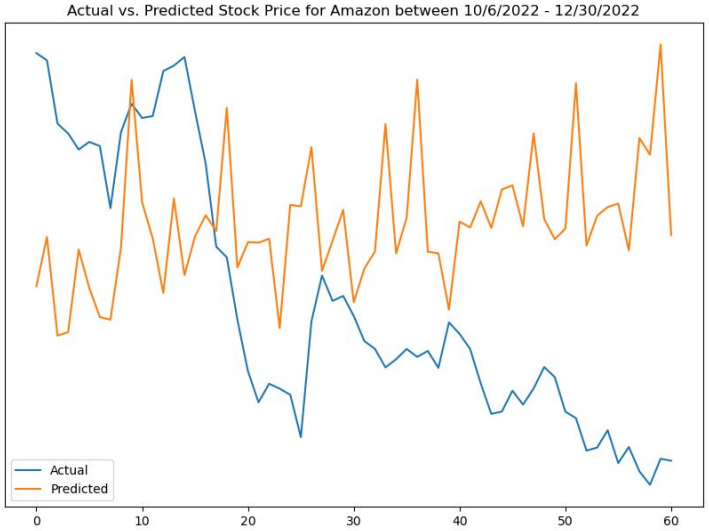
Performance Modeling

- ➔ LSTMs (Long Short-Term Memory)
- ➔ ARIMA
- ➔ ARIMA-GARCH





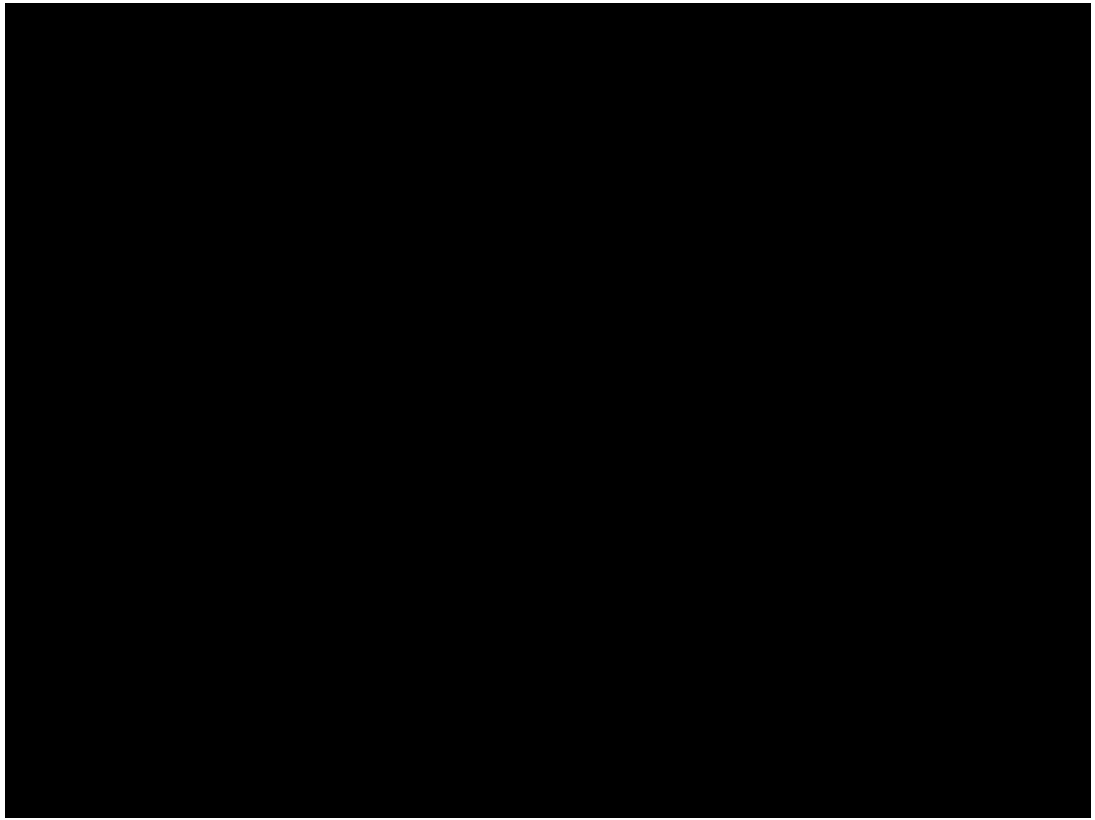
LSTM Network (with emotion dataset) Results



- ➔ Using all 640K posts with emotion labels
- ➔ Features: # of posts and total score (upvotes) by date and emotion
- ➔ High error (MAE, MSE, etc.) on testing set

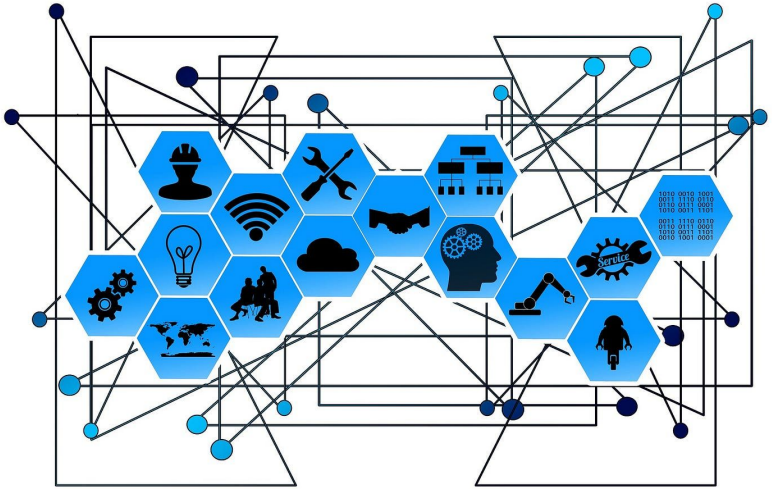
MAE	MSE	RMSE	Norm. MAE	Norm. MSE	Norm. RMSE
14.8	386.4	19.7	0.22	5.60	0.29

Demo



Lessons Learned

- ➔ A company's performance is a result of many different complex factors beyond social media sentiment
- ➔ Representation of the diversity of data points matters. Sentiment is not one dimensional
- ➔ Relevant training matters



Future Work

- Accessing different social platforms
  - Major news media companies
- Proper filtering of relevant data clearly possessing a significant impact (engagement)
- Integration of new features/data points to capture complex relationship
- Upgrading computational resources to achieve accurate modeling

Thank you!
Questions?