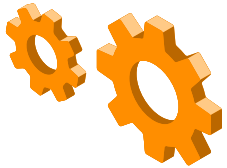
Stock Price Prediction with News and Historical Prices

Prof. Shayan Shams

Group 2 - Hui, Maria, Manyu

Table of contents



01 **Introduction**
Background & Motivation

02 **Methodology**
Project workflow

03 **Literature Review**
Insights and relevance to our project

04 **Dataset Exploration**
- Dataset description
- Text mining
- Numeric visualization

05 **Experiment & Result**
Model Experiment & results

06 **Discussion & Improvement**
Result discussion and future improvement

01

Introduction

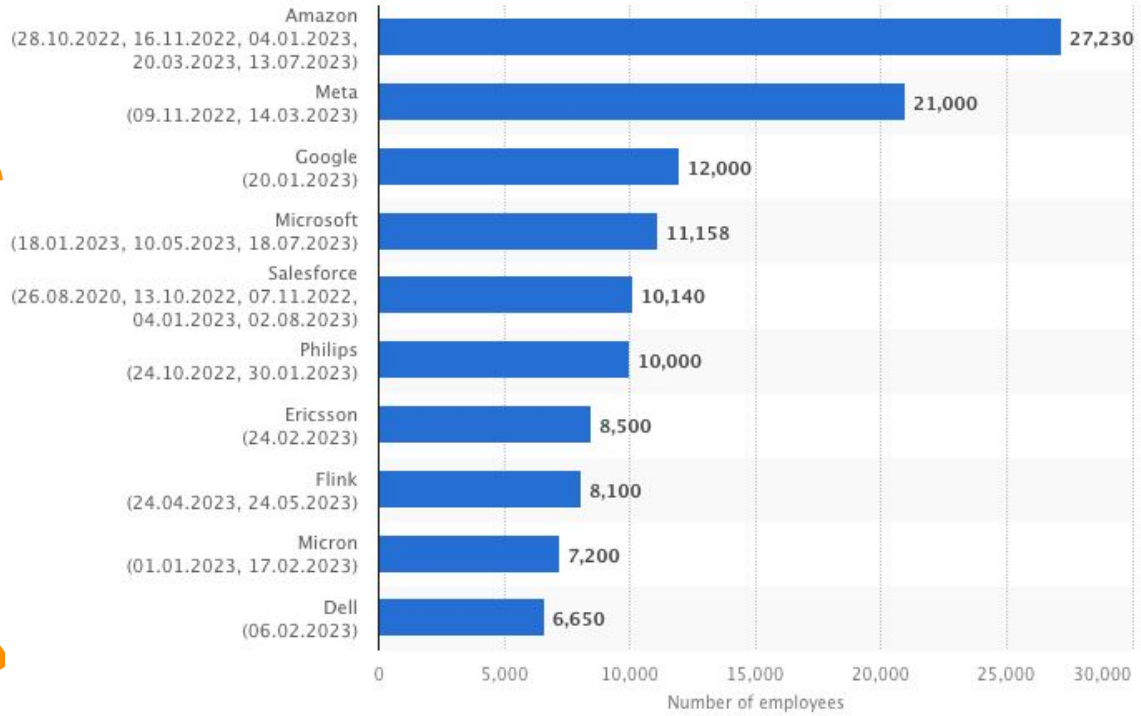
1.1 Background

In 2022, the tech-heavy Nasdaq 100 Index (NASDAQ:QQQ) is down ~28%, and many tech stocks are down more than 50-90%+.

Source: https://seekingalpha.com/article/4560496-2023-forecast-tech-stocks

1.1 Background

Number of Tech Laid Off Worldwide by Company (2020-2023)

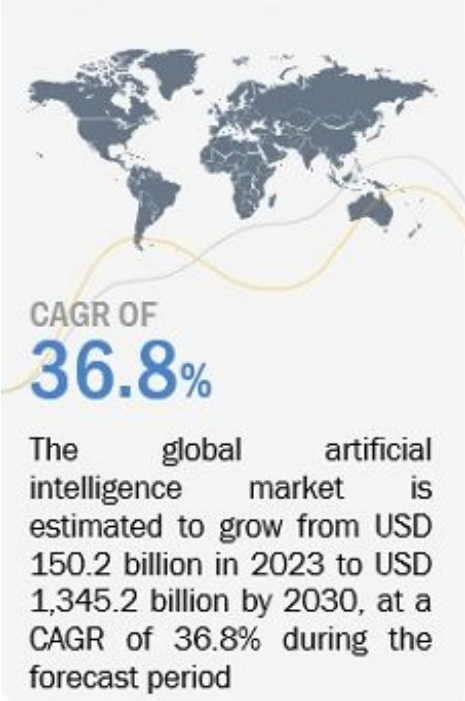


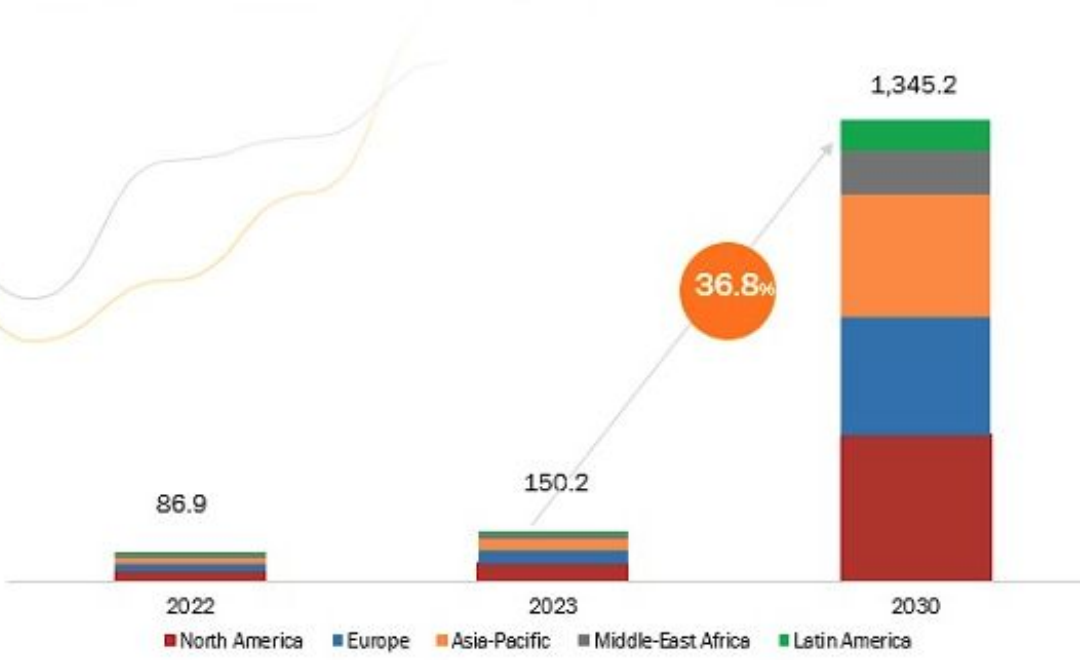
Since covid-19, tech layoffs continues, peaked at Jan 2023.

Source: https://www.statista.com/statistics/1127080/worldwide-tech-layoffs-covid-19-biggest/

1.1 Background

Artificial Intelligence Market Global Forecast To 2030 (USD BN)





Source: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-market-74851580.html

1.2 Motivation

Findings based on background:

1. Great change in tech sector results in heavy discussion on media, such as news, functions as external factor influencing the stock market.
2. Traditional stock price prediction methods might be biased in today's situation.

Goal of the project:

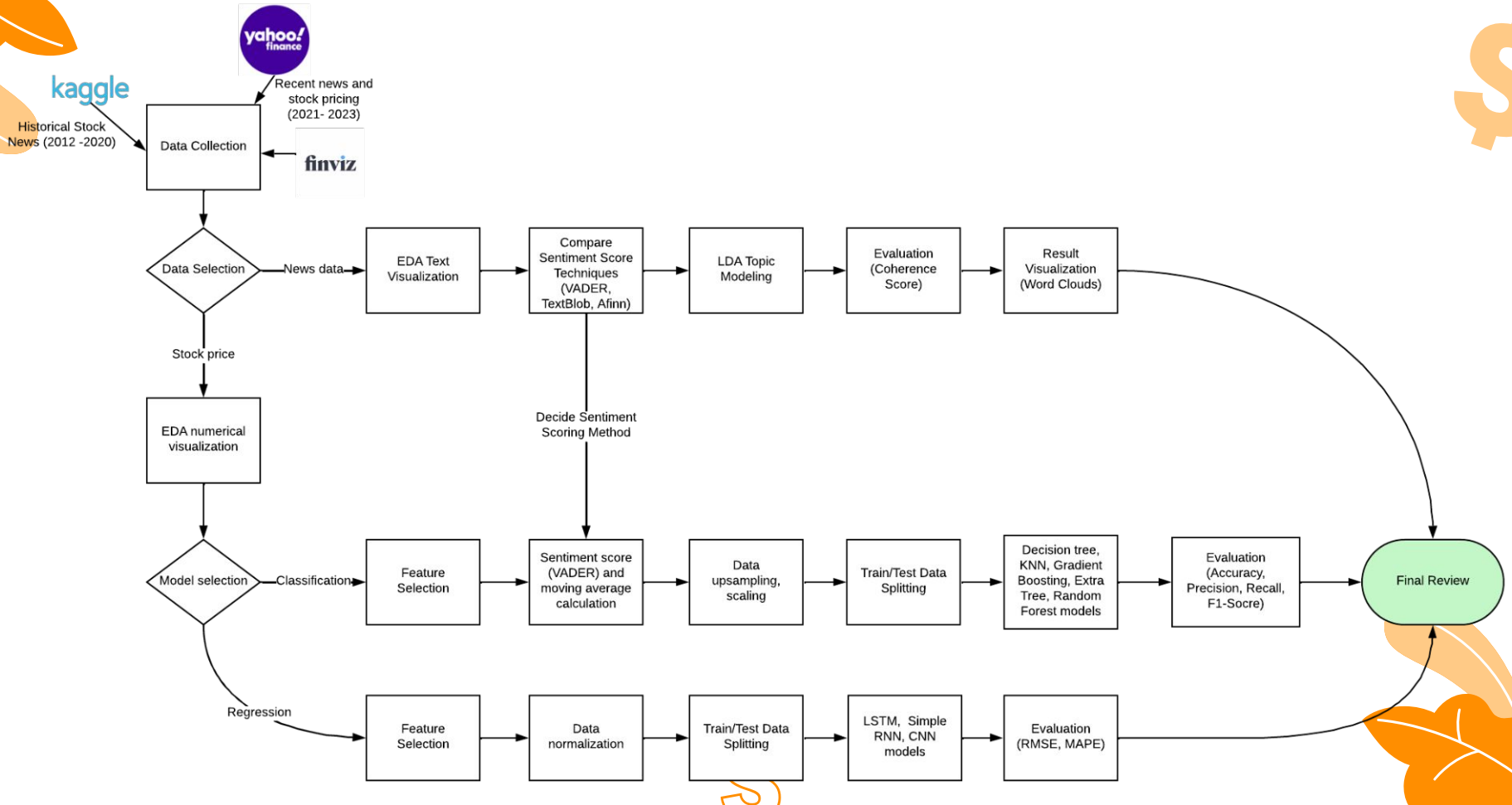
Combine historical stock price data and the influence of recent media,

Build models to predict the stock price of tech sector more accurately.

02

Methodology

Project Methodology Flowchart



03
Literature Review

3 Literature Review

	Paper, Author, Year	Summary
Text Mining	*Stock prediction by integrating sentiment scores of financial news and MLP-Regressor: A machine learning approach (Junaid et. al., 2023)*	- Authors compared 3 sentiment scoring algorithms (VADER, TextBlob, and FLAIR) combined with MLP-Regressor for news based stock price prediction. - Although VADAR is most common one, TextBlob delivered best performance with MLP-Regressor.
	Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms (Wei-Chao et al., 2022)	- Authors compared four text representation methods (TF-IDF, Word2Vec, ELMo and BERT) and three learning algorithms (SVM, CNN and LSTM) to find the best combination for stock price prediction. - CNN+Word2vec and CNN+EBRT were the most effective combination.
Deep Learning	*Stock price prediction using LSTM, RNN and CNN-sliding window model (Selvin et al., 2017)*	- The Authors compared traditional time series models like AR, ARMA, and ARIMA with deep learning models like LSTM, RNN and CNN to predict stock prices. - The error rate of using deep learning models are 10 times less. - Deep learning model is good at analyzing the interaction and pattern of the data and handling non linear data like stock price.
	Stock market prediction using LSTM recurrent neural network (Moghar and Hamiche, 2020)	The Authors experiment LSTM, a subtype of RNN to predict stock price of GOOG and NKE due to its capability to learn long-term dependencies and address the vanishing gradient problem prevalent in traditional RNNs, making it suitable predicting financial time series, where long-term dependencies are often present.

3 Literature Review (Cont.)

	Paper, Author, Year	Summary
Classification	*Predicting the direction of stock market prices using random forest* (Luckyson et al., 2016)	The authors treat the stock market forecasting problem as a classification problem and use machine learning algorithms to predict stock returns. They use technical indicators such as Relative Strength Index (RSI) and stochastic oscillator as inputs to train their model, which is an ensemble of multiple decision trees. The authors also mention the use of other algorithms such as Support Vector Machines (SVM), Neural Networks, and Logistic Regression in the field of stock market prediction. Best works random forest

04

Dataset Exploration

4.1 Dataset Description

1.Historical Dataset

- Kaggle
- Apple
- 2012-2020
- Stocks and News

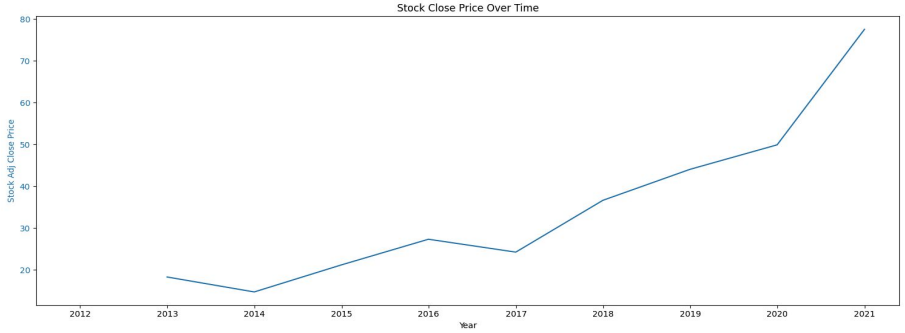
2.Scrapped Dataset

- 01/01/2021 - 11/26/2023
- Stocks
- News only for November
- Finviz and Yahoo
- FAANG companies
- Adding label (Classification)

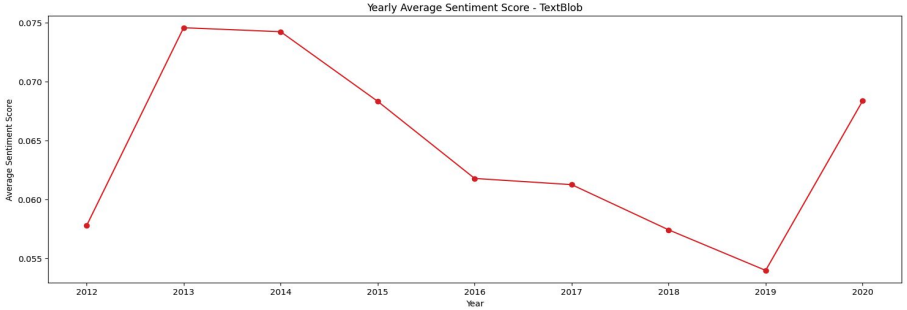
4.2 Text Data- Historical Data

VADER (Valence Aware Dictionary and sEntiment Reasoner) is the optimal sentiment scoring technique.

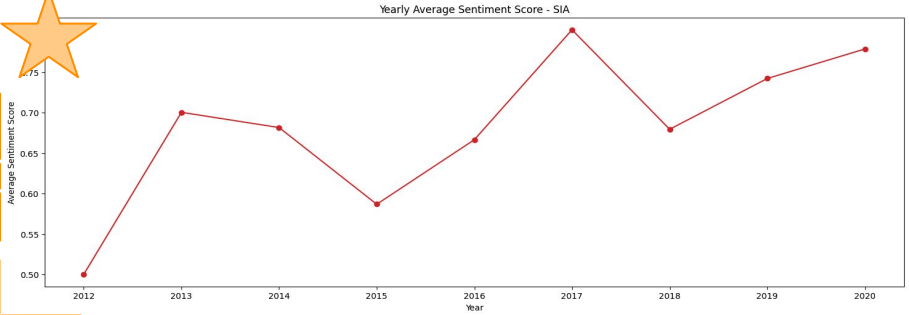
Stock Price Over Time



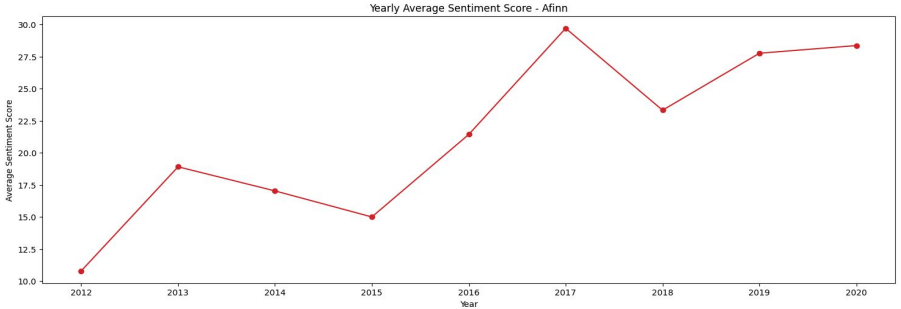
TextBlob Sentiment Score Over Time



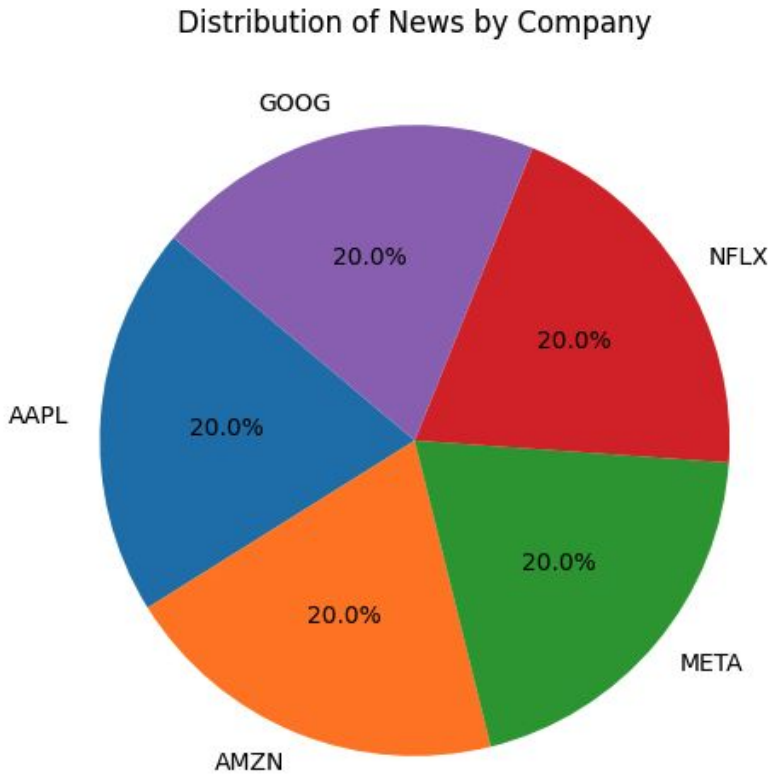
VADER Sentiment Score Over Time



Afinn Sentiment Score Over Time

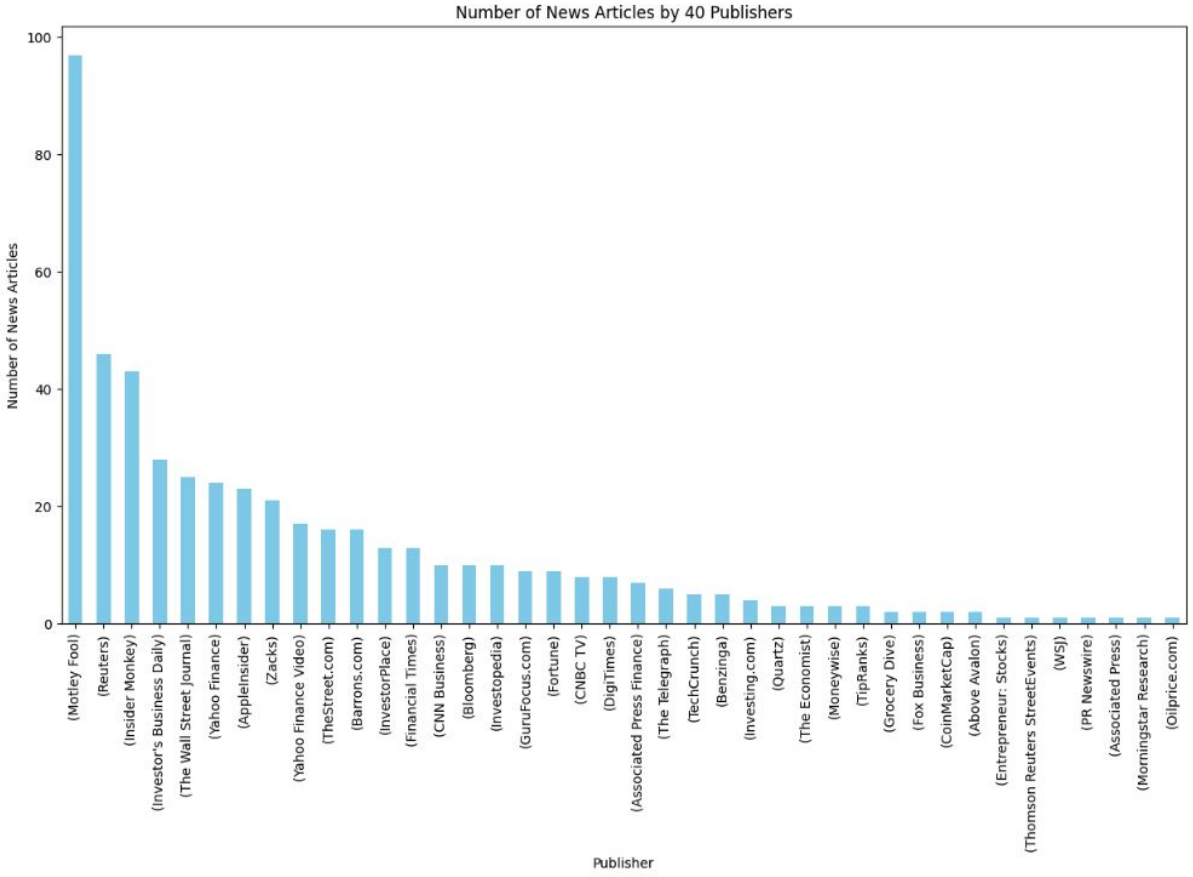


4.2 Text Data - Scraped Data



500 News published in 11/2023
(11/01/2023 - 11/26/2023)
about “FAANG” big tech companies.

4.2 Text Data - Scraped Data



News sourced from 40 different publishers, including news, blog, etc.

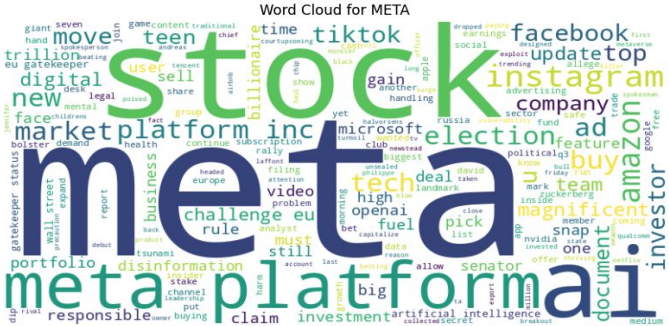
4.2 Text Data - Scraped Data

Word Cloud of All Scraped Data

4.2 Text Data - Stratify by Company

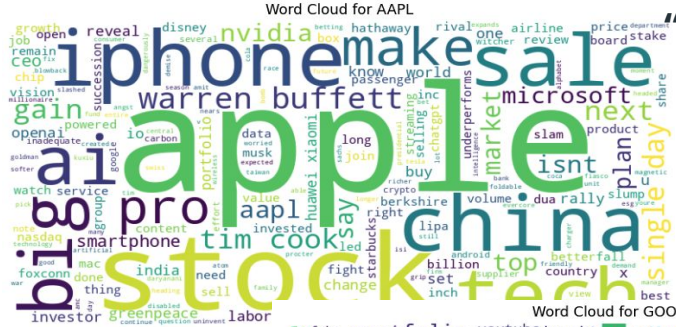
“F”

Word Cloud for META



“A”

Word Cloud for AAPL



“A”

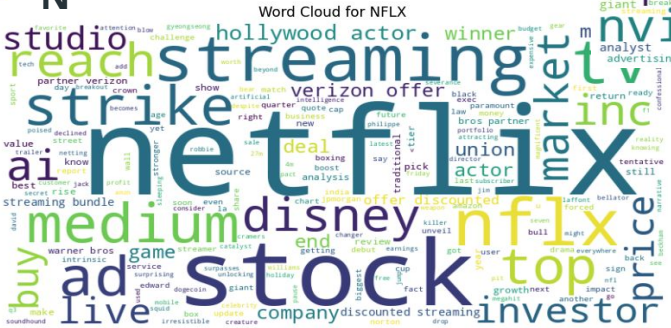
Word Cloud for AMZN

“G”

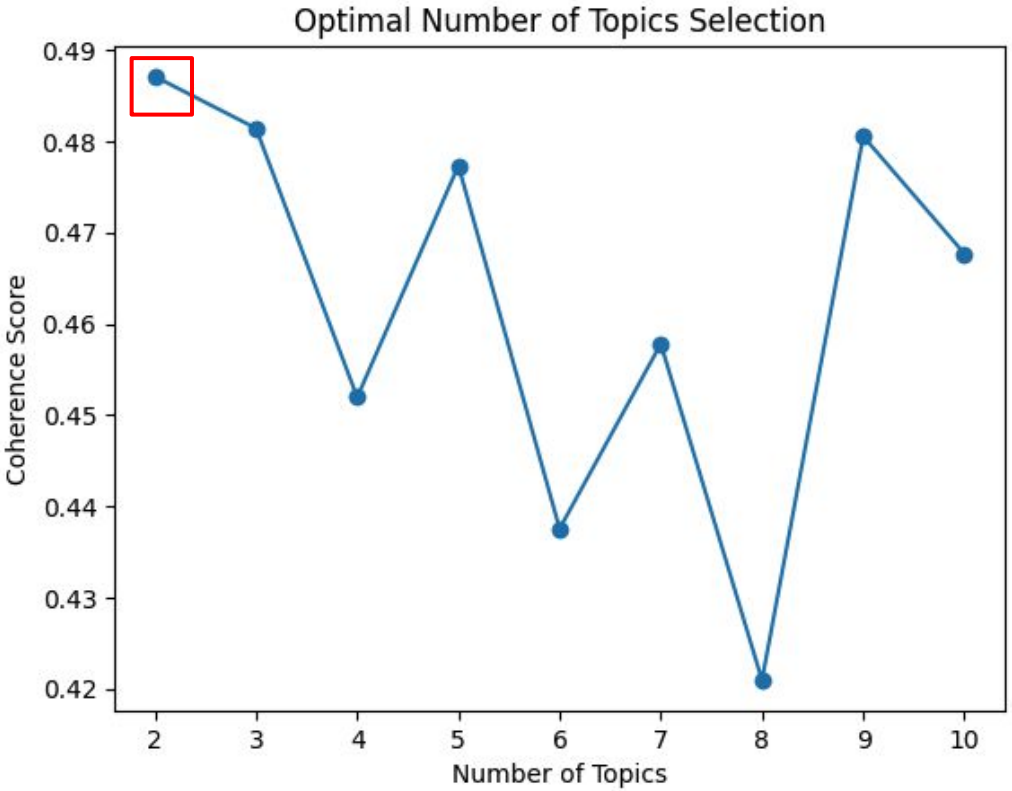
Word Cloud for GOOG

“N”

Word Cloud for NFLX



4.2 Text Mining - LDA Topic Modeling



The optimal number of topics is 2

4.2 Text Mining - LDA Topic Modeling

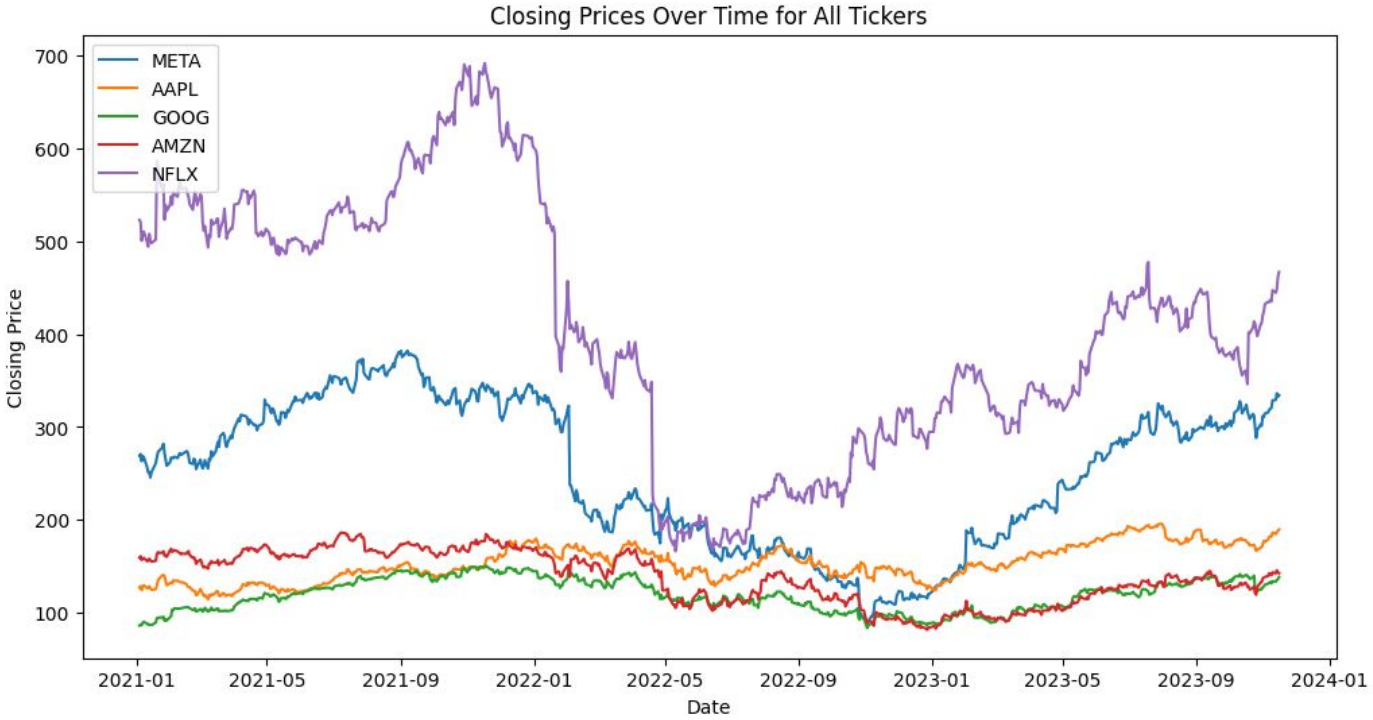
Topic 1: 0.058*"stock" + 0.018*"amazon" + 0.014*"ai" + 0.012*"meta" + 0.012*"buy" + 0.010*"tech" + 0.009*"top" + 0.008*"market" + 0.006*"big" + 0.006*"pick"

Topic 2: 0.021*"apple" + 0.017*"netflix" + 0.011*"amazon" + 0.009*"meta" + 0.007*"streaming" + 0.007*"strike" + 0.007*"black" + 0.006*"ai" + 0.006*"friday" + 0.006*"update"

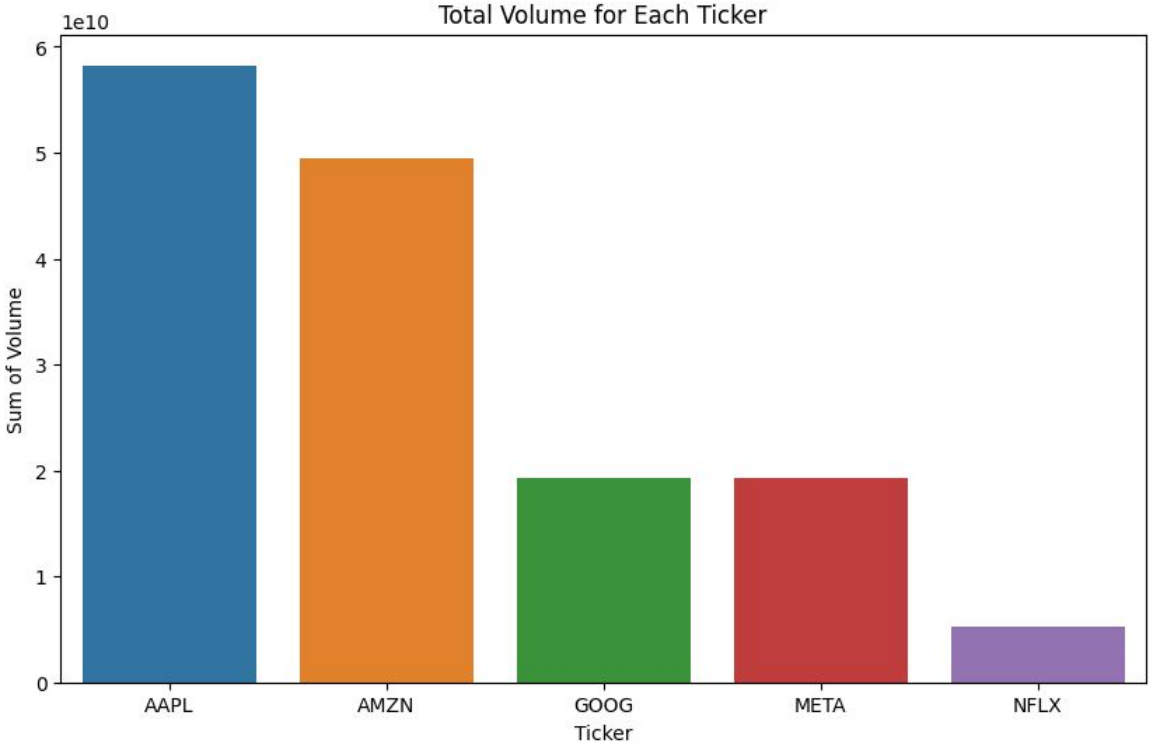
Topic 1: Talks about AI and Tech

Topic 2: Talks about Black Friday

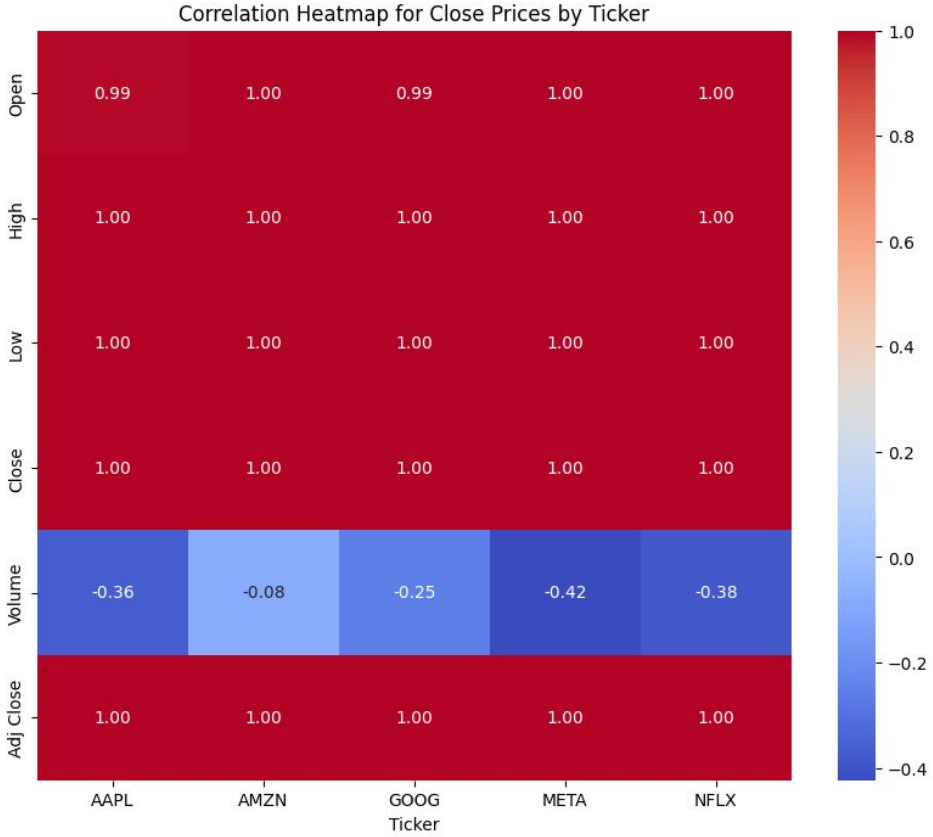
4.3 Numerical Data Visualization



4.3 Numerical Data Visualization (Cont.)



4.3 Numerical Data Visualization(Cont.)



Open, **High**, **Low**, **Adj Close** have strong correlation to **Close price (target value)** that can be used as feature values.

05
Experiment and Results

5.1 Deep Learning Models

Experiment Steps

1. **Data Selection:** Stock price of FAANG companies ranging from 1/1/2021 to 11/17/2023.
2. **Feature Selection:** Open, High, Low, Adj Close , **Target Value:** Close
3. **Data Normalization:** Normalize training data between 0 and 1 to avoid overfitting
4. **Data Splitting:** Data before 1/1/2023 is training; data after 1/1/2023 is testing
5. **Create sequences** for time series data
6. **Modeling:** Build LSTM, Simple RNN and CNN
7. **Training:** Run training and testing data of each ticker on 3 models
8. **Evaluation:** Compare RMSE, MAPE the actual v.s predicted price plots

Results - Error Comparison

Error Result of Deep Learning Models

Model	Ticker	RMSE	MAPE
LSTM	META	0.0304	8.81%
	AAPL	0.0504	13.51%
	AMZN	0.03752	14.89%
	NFLX	0.0188	10.30%
	GOOG	0.053	11.99%
Simple RNN	META	0.0307	10.34%
	AAPL	0.0487	10.74%
	AMZN	0.03753	13.86%
	NFLX	0.0190	12.32%
	GOOG	0.056	12.95%
CNN	META	0.0405	11.16%
	AAPL	0.0723	14.86%
	AMZN	0.04689	20.14%
	NFLX	0.02711	15.54%
	GOOG	0.0735	19.97%

MAPE	Interpretation
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

Source: Lewis (1982, p. 40)

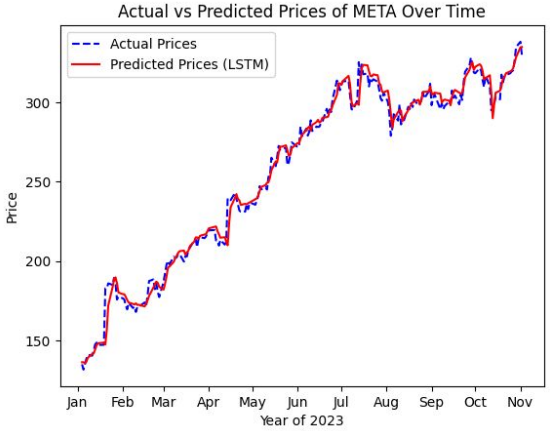
1. RNN models outperformed CNN models for the training on all tickers, with LSTM performed the best most of the time, and Simple RNN also performed quite well.

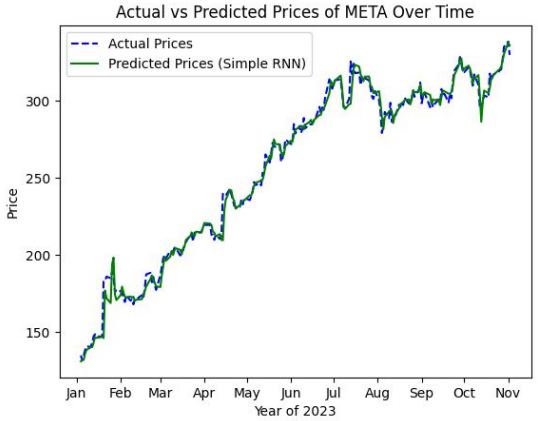
2. CNN tended to perform worst among 3 models on all tickers.

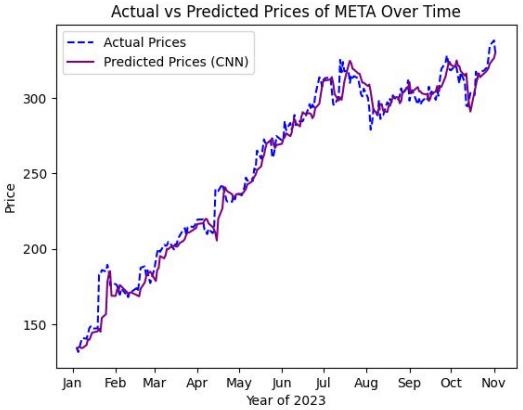
Results - Actual Price v.s Predicted Price Plot

LSTM | **Simple RNN** | **CNN**

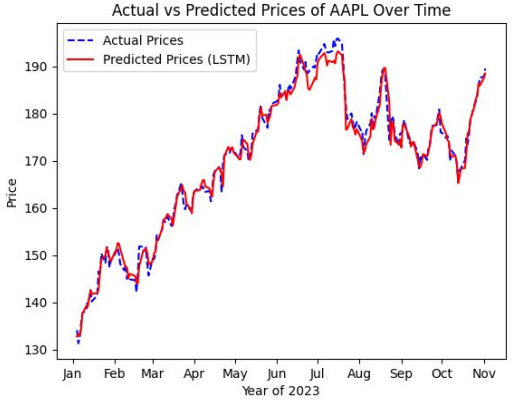
META

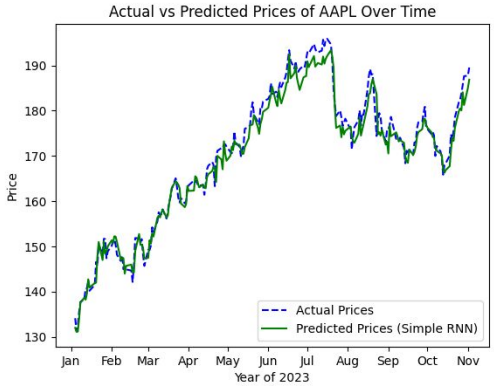


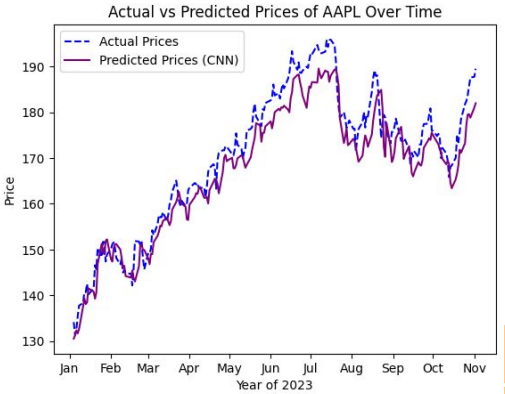




AAPL



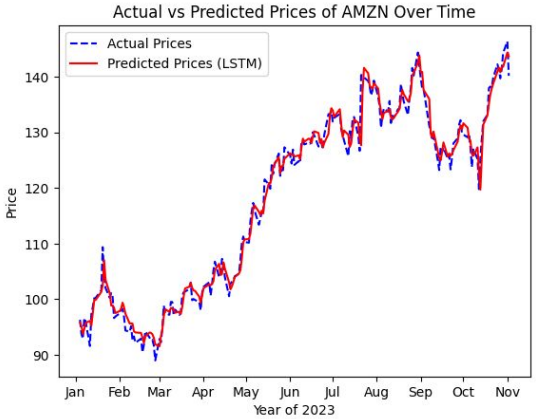


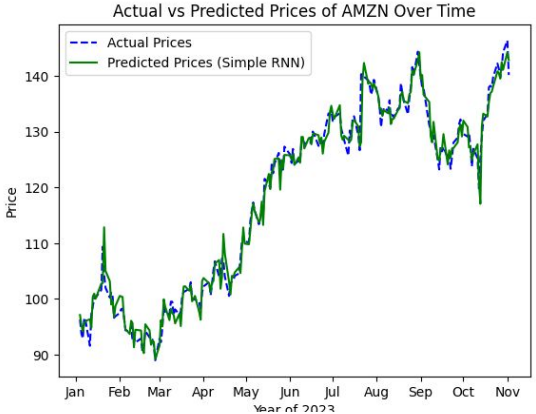


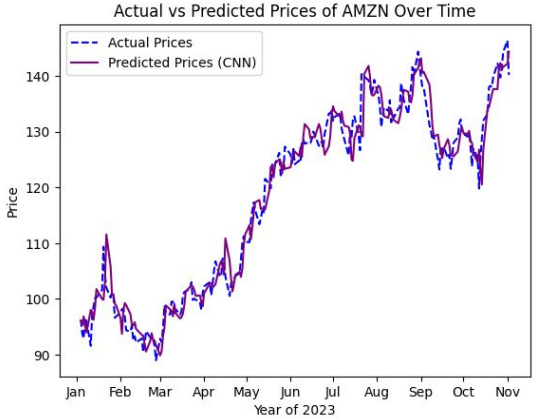
Results - Actual Price v.s Predicted Price Plot

LSTM | **Simple RNN** | **CNN**

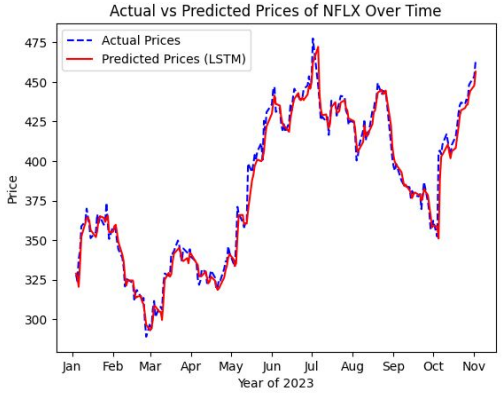
AMZN

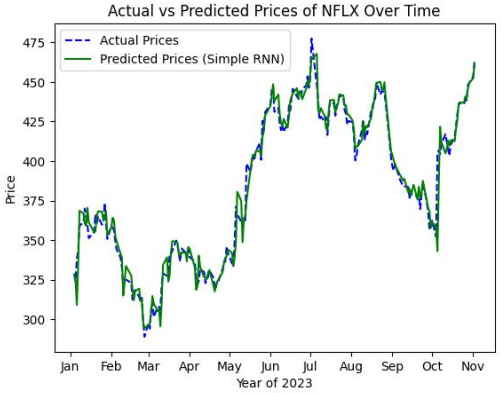


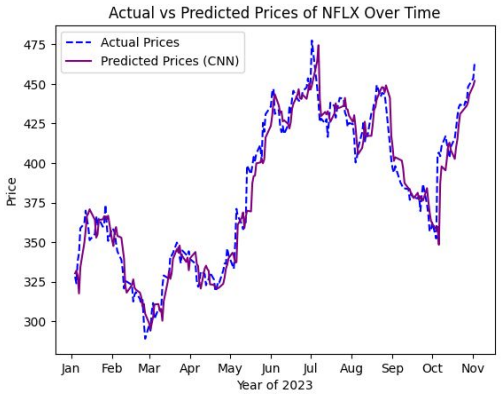




NFLX



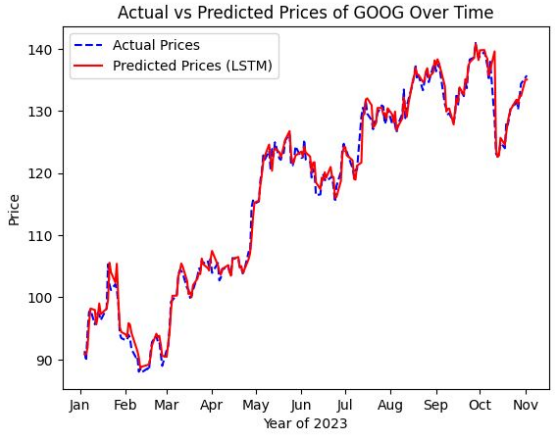




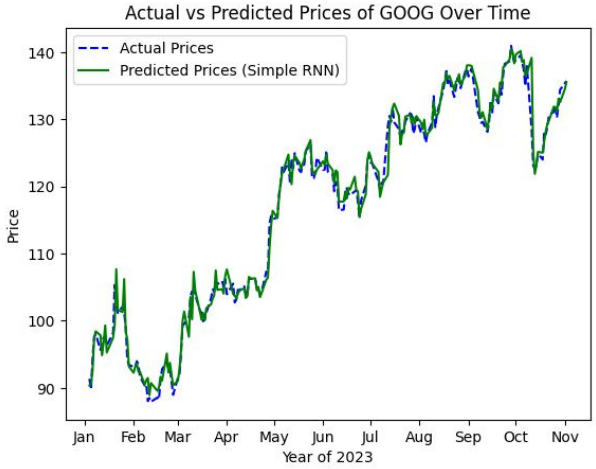
Results - Actual Price v.s Predicted Price Plot

GOOG

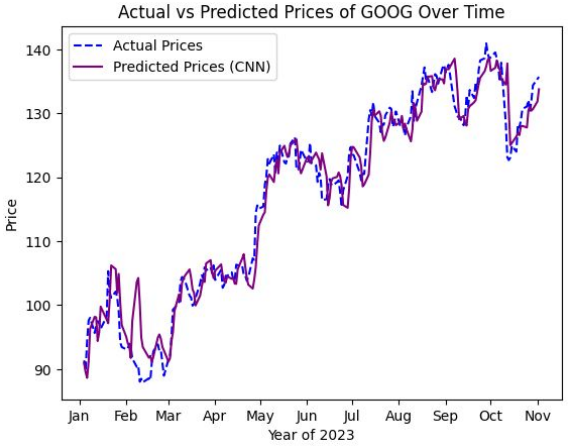
LSTM



Simple RNN



CNN



5.2 Classification

5.2 Classification

1. **Data Collection:** Scraping Data From Yahoo and Finviz
2. **Creating New Feature:** Compare previous close price with next close price and add 'label' column
3. **Feature Engineering:** Moving Average 5 and 10 window, and compound score
4. **One Hot Encoding and Feature Selection**
5. **Data Scaling:** Normalization
6. **Balancing:** Upsampling
7. **Modeling and Training:** Decision Tree, KNN, Gradient Boosting, Extra Trees, Random Forest
8. **Evaluation:** Compare Accuracy,Precision, Recall, F1-Score

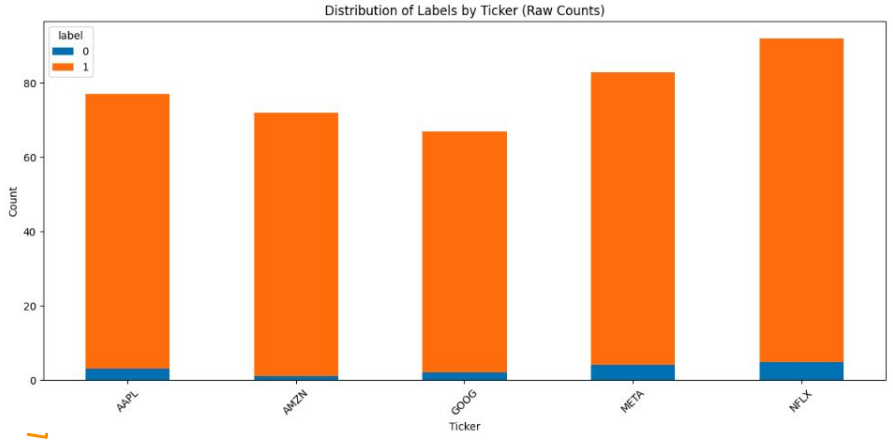
Before Transformation

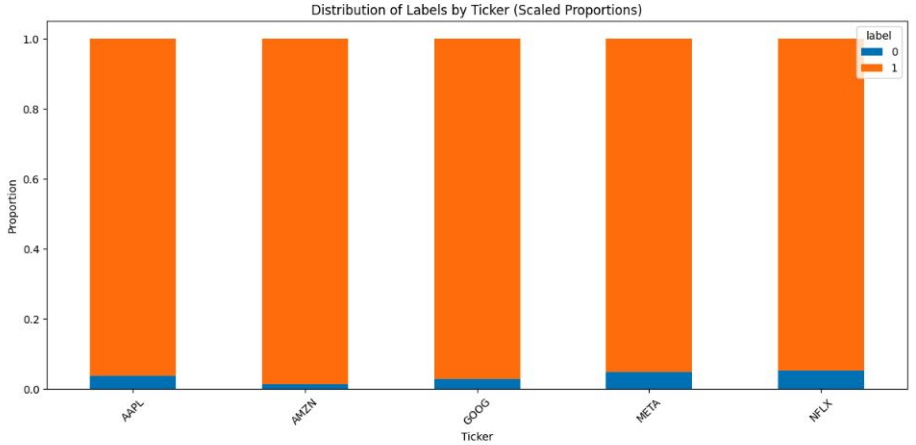
	ticker	Open	High	Low	Close	Adj Close	Volume	date	publish_time	title	publisher
0	AAPL	189.889999	191.910004	189.880005	191.449997	191.449997	46505100	2023-11-20	10:30PM	How to share AirTag in iOS 17	(AppleInsider)
1	AAPL	189.889999	191.910004	189.880005	191.449997	191.449997	46505100	2023-11-20	09:30PM	Apple eyes increased iPhone production amid Im...	(DigiTimes)
2	AAPL	189.889999	191.910004	189.880005	191.449997	191.449997	46505100	2023-11-20	09:01PM	iPhone 16 Pro to get 120 mm camera says Kuo, a...	(AppleInsider)
3	AAPL	189.889999	191.910004	189.880005	191.449997	191.449997	46505100	2023-11-20	08:01PM	CEO of Fortnite game maker casts Google as a '...	(Associated Press Finance)
4	AAPL	189.889999	191.910004	189.880005	191.449997	191.449997	46505100	2023-11-20	06:34PM	Epic Games Sweeney Takes Aim at Androids Fake ...	(Bloomberg)

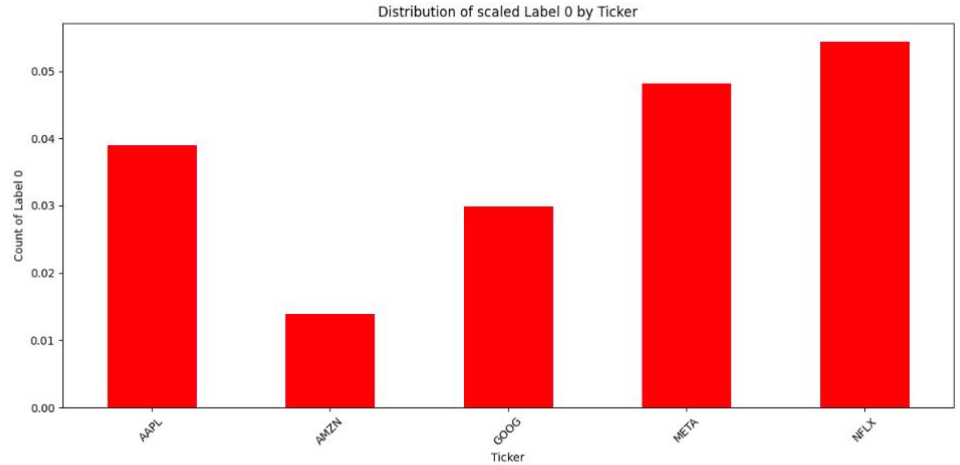
After Transformation

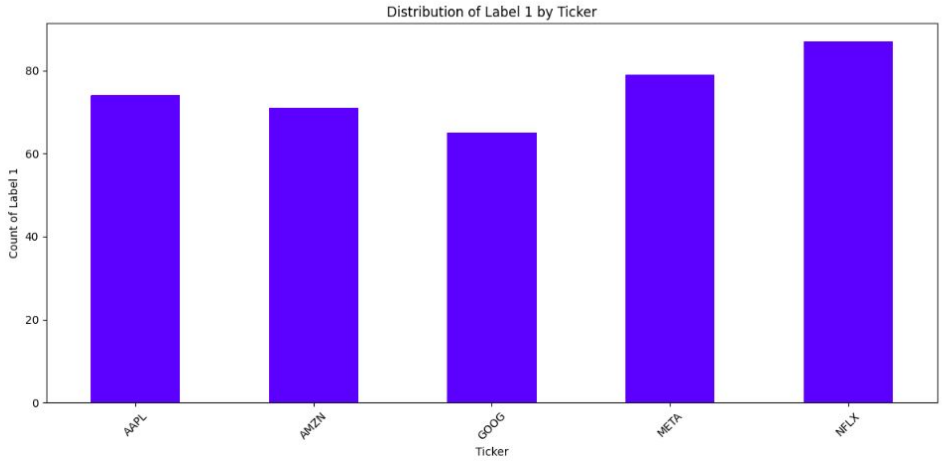
	Open	High	Low	Close	Adj Close	Volume	compound	Close_Pct_Change	MA_5	MA_10	ticker_AAPL	ticker_AMZN	ticker_GOOG	ticker_META	ticker_NFLX
0	-0.597807	-0.594869	-0.587178	-0.590060	-0.590060	1.500805	0.566098	-0.053149	-0.583963	-0.576042	1	0	0	0	0
1	-0.597807	-0.594869	-0.587178	-0.590060	-0.590060	1.500805	1.491158	-0.053149	-0.583963	-0.576042	1	0	0	0	0
2	-0.597807	-0.594869	-0.587178	-0.590060	-0.590060	1.500805	-0.336400	-0.053149	-0.583963	-0.576042	1	0	0	0	0
3	-0.597807	-0.594869	-0.587178	-0.590060	-0.590060	1.500805	-1.842291	-0.053149	-0.583963	-0.576042	1	0	0	0	0
4	-0.597807	-0.594869	-0.587178	-0.590060	-0.590060	1.500805	-1.789848	-0.053149	-0.583963	-0.576042	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
386	1.706318	1.693270	1.716669	1.708228	1.708228	-1.202189	0.353584	-0.053149	1.736814	1.769129	0	0	1	0	0
387	1.706318	1.693270	1.716669	1.708228	1.708228	-1.202189	0.280409	-0.053149	1.736814	1.770401	0	0	1	0	0
388	1.706318	1.693270	1.716669	1.708228	1.708228	-1.202189	0.206014	-0.053149	1.736814	1.771673	0	0	1	0	0
389	1.706318	1.693270	1.716669	1.708228	1.708228	-1.202189	-0.336400	-0.053149	1.736814	1.772945	0	0	1	0	0
390	1.706318	1.693270	1.716669	1.708228	1.708228	-1.202189	-0.336400	-0.053149	1.736814	1.772945	0	0	1	0	0

Imbalance Data









Results

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.493671	0.943741	0.493671	0.602438
1	KNN	0.746835	0.879999	0.746835	0.804711
2	Gradient Boosting	0.493671	0.943741	0.493671	0.602438
3	Extra Trees	0.518987	0.918233	0.518987	0.629904
4	Random Forest	0.518987	0.918233	0.518987	0.629904

06

Discussion & Improvement

6. Discussion & Future Improvement

- Use tools like **crontab** to scrap more data
- Integrate other websites for news
- Use not only titles but articles too
- Advanced Natural Language Processing (NLP) Techniques:
  Using LLMs for summarizing Articles

07

Contribution & Reference

Contribution Table

Name	Contribution
Hui Yun	Literature review, pricing data collection via yahoo api, numerical data visualization, deep learning model experiment and evaluation
Maria Hovhannisyan	Literature review, scrapping news and stock data from yahoo and finviz, data EDA, Classification model building and evaluation
Manyu Zhang	Background study, literature review, sentiment scoring techniques comparison, text EDA, LDA topic modeling

Reference

[1]Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest.

[2] Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock prediction by integrating sentiment scores of financial news and MLP-Regressor: A machine learning approach. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050923000868

[3] Moghar, A., & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 170, 1168–1173. https://doi.org/10.1016/j.procs.2020.03.049

[4] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). https://doi.org/10.1109/icacci.2017.8126078

[5] Wei-Chao Lin, Chih-Fong Tsai, Hsuan Chen, Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms, Applied Soft Computing (2022), https://doi.org/10.1016/j.asoc.2022.109673

[6] https://seekingalpha.com/article/4560496-2023-forecast-tech-stocks

[7] https://www.statista.com/statistics/1127080/worldwide-tech-layoffs-covid-19-biggest/

[8] https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-market-74851580.html

THANKS
Q & A

Prof. Shayan Shams

Group 2 - Hui, Maria, Manyu