



# Forecasting Stock Prices with Public Sentiment



# Project goals

- We are interested in analyzing the relationship between public sentiment and stock price movement for Apple, Microsoft, and Google
- Why is this important?
  - Profit generation
  - Market participation



# Data

- Data from 3 sources:
  - NLTK sentiment scores from:
    - Reddit
    - NYT + Guardian
  - Historical stock closing prices
- Focusing on Apple, Microsoft, Google
- NLTK Vader sentiment

Negative Score	0 to 1 (probability that response is negative)
Neutral Score	0 to 1 (probability that response is neutral)
Positive Score	0 to 1 (probability that response is positive)

# Reddit Submission Examples

Posted by u/[redacted] r/apple

**Apple Failed to Cope with a Hacker**

[redacted] r/apple

I would agree on that. To me, both EarPods and AirPods fit perfectly, never had a problem. Friend of mine has trouble with both, but can fit all these other silicone "one-size-fits-all-except-me-apparently" earphones without a problem.

After Apple Watch, AirPods is my most recommended Apple product!

[redacted] r/google

wow, google amazes me around every corner

[redacted] r/microsoft

Hey [redacted] unfortunately your submission has been removed because it seems to be a support request. Please review the [\[r/microsoft rules\]](#) (<http://www.reddit.com/r/microsoft/about/sidebar>) before posting. The sidebar contains links to contact us if you feel this removal was an error

\*I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](/message/compose/?to=[r/microsoft](#)) if you have any questions or concerns.\*

[redacted] r/apple

Google news is way better than Apple news

# Reddit + NLTK Examples

Posted by u/[redacted] r/apple  
**Apple Failed to Cope with a Hacker**

0.398, 0.602, 0.0

[redacted] r/apple  
I would agree on that. To me, both EarPods and AirPods fit perfectly, never had a problem. Friend of mine has trouble with both, but can fit all these other silicone "one-size-fits-all-except-me-apparently" earphones without a problem.

0.067, 0.624, 0.308

After Apple Watch, AirPods is my most recommended Apple product!

[redacted] r/google  
wow, google amazes me around every corner

0.0, 0.417, 0.583

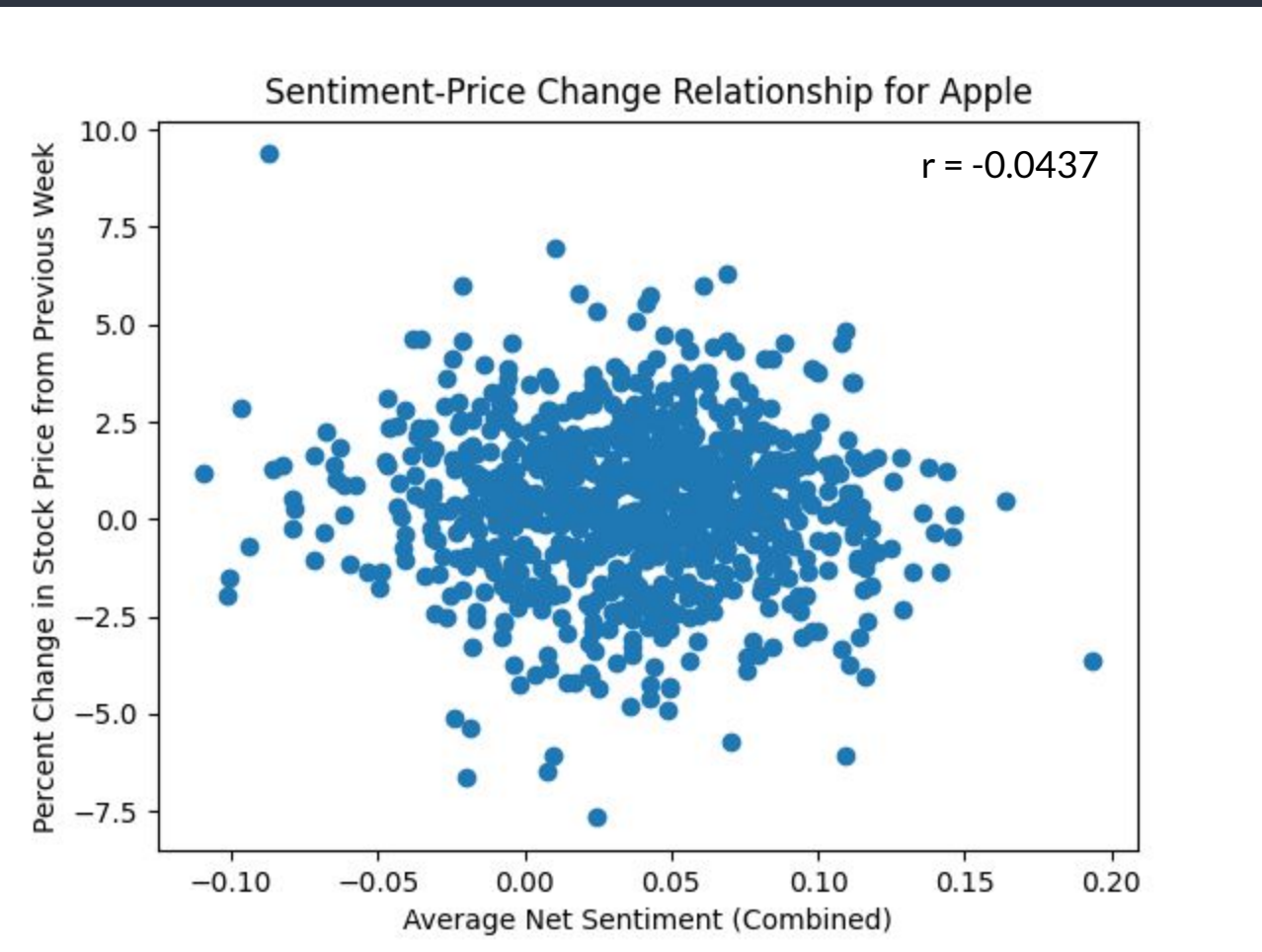
[redacted] r/apple  
Google news is way better than Apple news

0.0, 0.707, 0.293

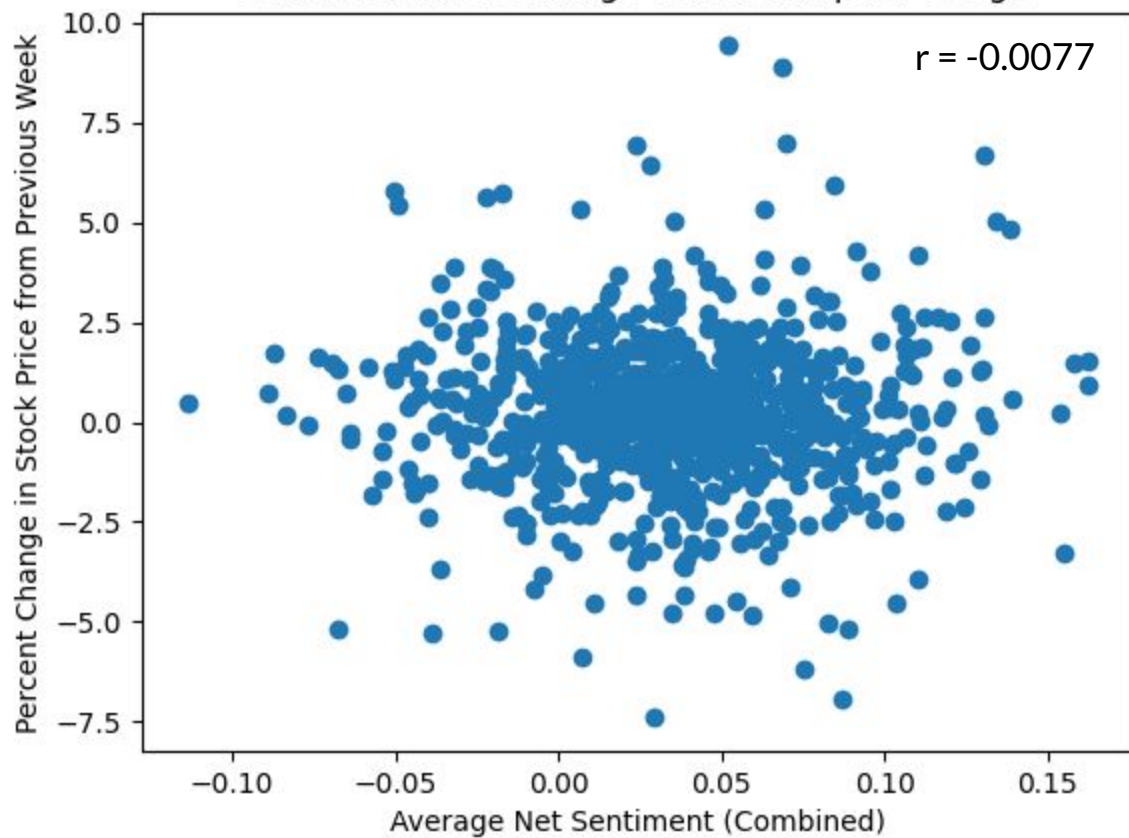
[redacted] r/microsoft  
Hey [redacted] unfortunately your submission has been removed because it seems to be a support request. Please review the [\[r/microsoft rules\]](#) (<http://www.reddit.com/r/microsoft/about/sidebar>) before posting. The sidebar contains links to contact us if you feel this removal was an error

0.076, 0.816, 0.108

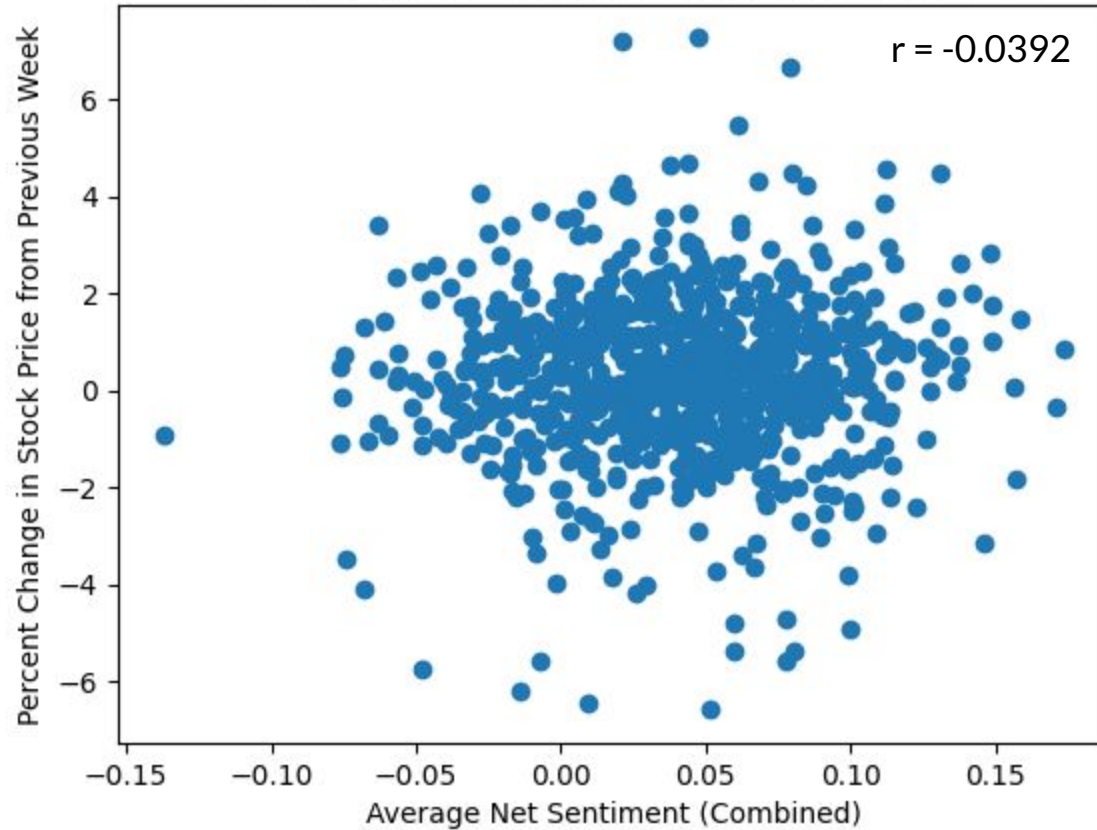
\*I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](/message/compose/?to=[r/microsoft](#)) if you have any questions or concerns.\*



Sentiment-Price Change Relationship for Google



Sentiment-Price Change Relationship for Microsoft







# Pearson Analysis

- R-Squared Values
  - Apple: 0.001913
  - Google: 0.000059
  - Microsoft: 0.001534
- Adding more windows
- Trade-offs of adding more independent variables
  - Likely leads to improvement of r-squared value, as well as improvements to the model
  - Opens up possibilities of overfitting



# Autoregression

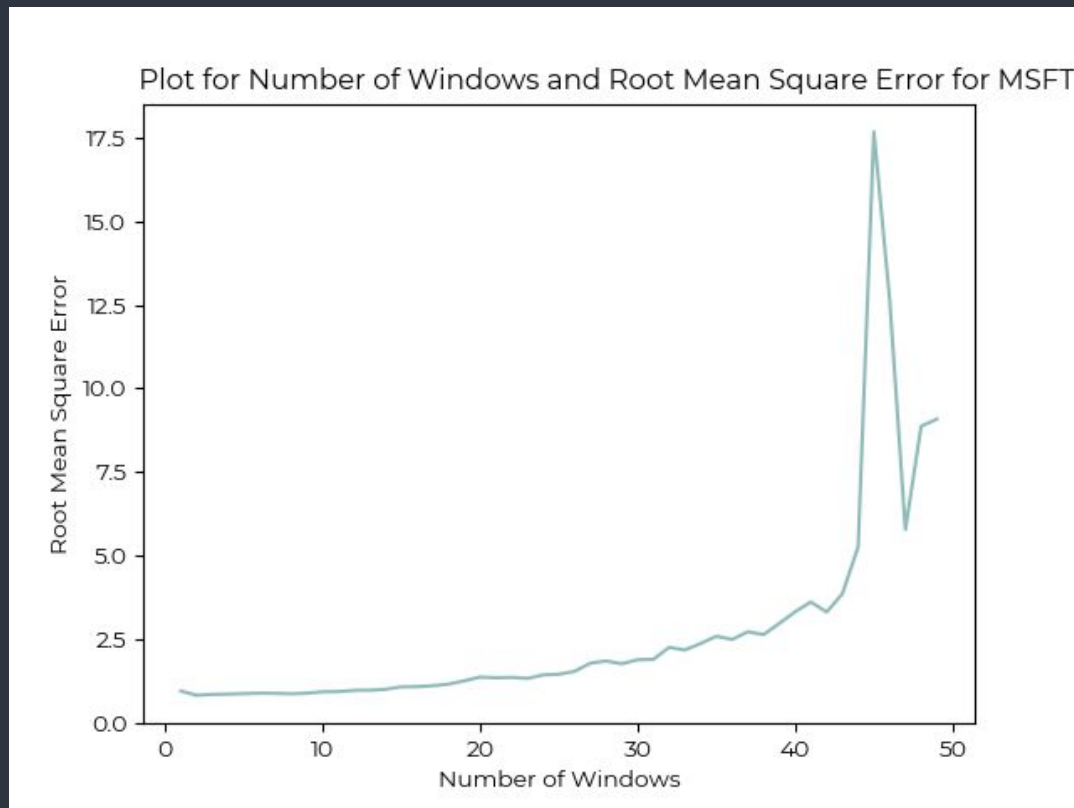


# Goal

- Determine whether an autoregression can be made using the sentiment values separately and price
  - Features:
    - Price
    - Average Positive Sentiment from Reddit
    - Average Neutral Sentiment from Reddit
    - Average Negative Sentiment from Reddit
    - Average Positive Sentiment from the Guardian
    - Average Neutral Sentiment from the Guardian
    - Average Negative Sentiment from the Guardian
    - Average Positive Sentiment from NYT
    - Average Neutral Sentiment from NYT
    - Average Negative Sentiment from NYT
- Look at different windows/amount of lag time to determine which gives the optimal fit for the model

- 
- Price
  - Average Positive Sentiment from Reddit
  - Average Neutral Sentiment from Reddit
  - Average Negative Sentiment from Reddit
  - Average Positive Sentiment from the Guardian
  - Average Neutral Sentiment from the Guardian
  - Average Negative Sentiment from the Guardian
  - Average Positive Sentiment from NYT
  - Average Neutral Sentiment from NYT
  - Average Negative Sentiment from NYT

# RMSE Values vs Windows Size



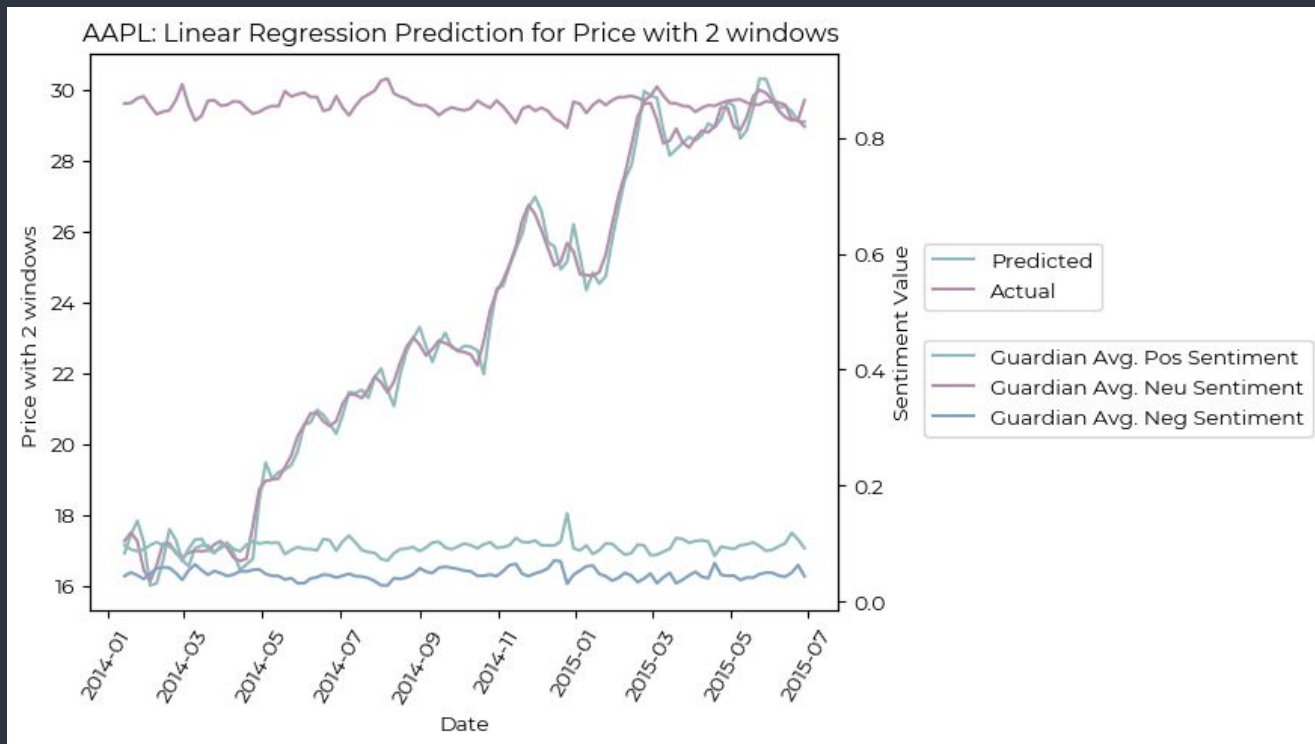
# Model Result



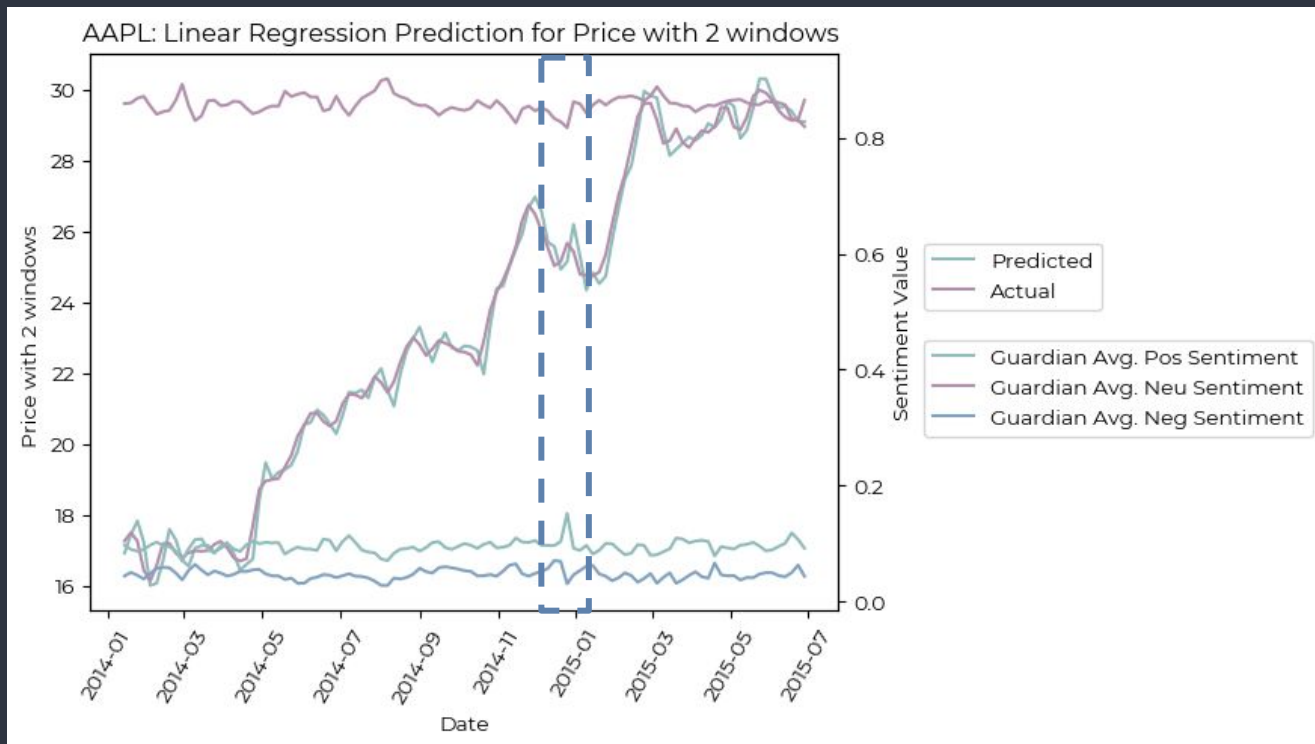
RMSE: 0.352

$R^2$ : 0.994

# Model Analysis

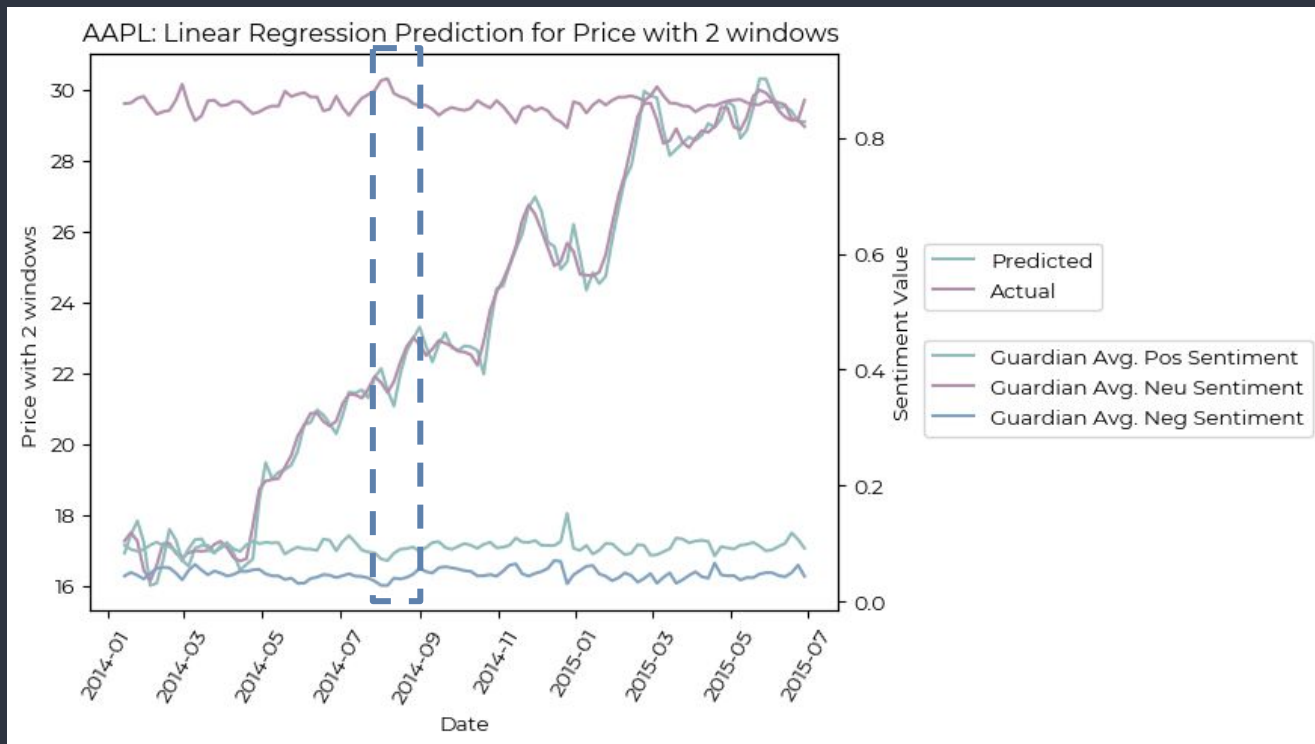


# Model Analysis





# Model Analysis





# Feature Analysis

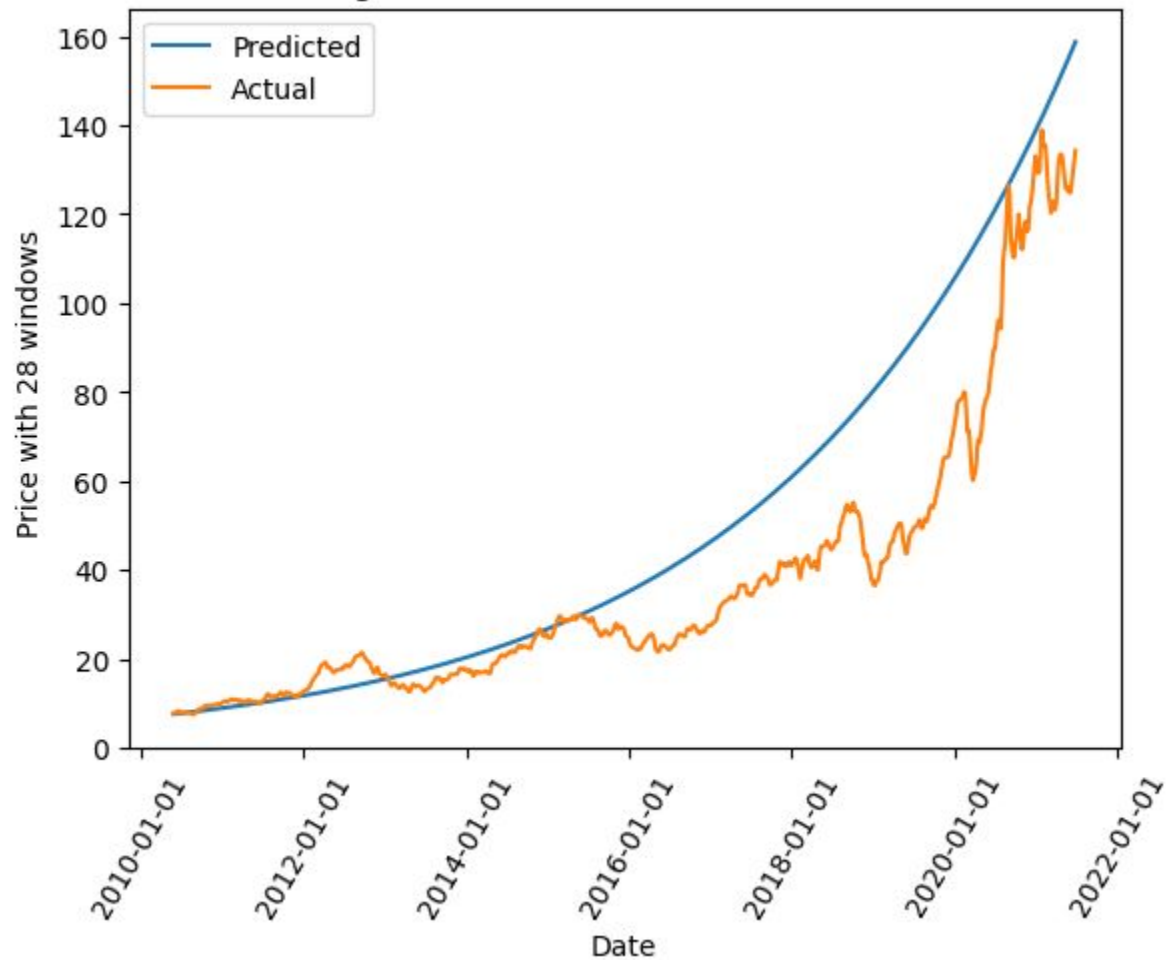




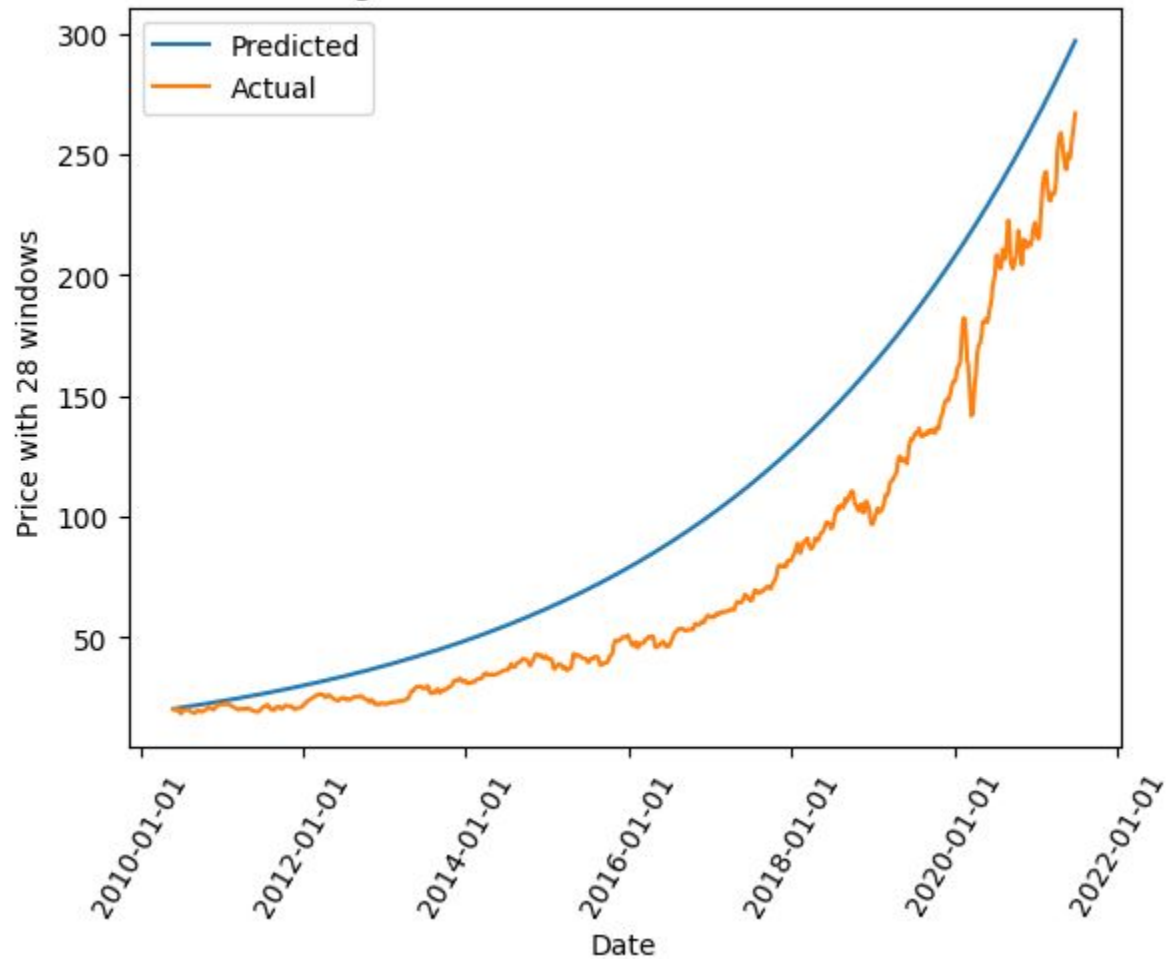
# Baseline Model

- Baseline Model:
  - An autoregression that looked at the previous 28, 10-day sliding windows and made predictions about stock price using **Average Closing Price**.

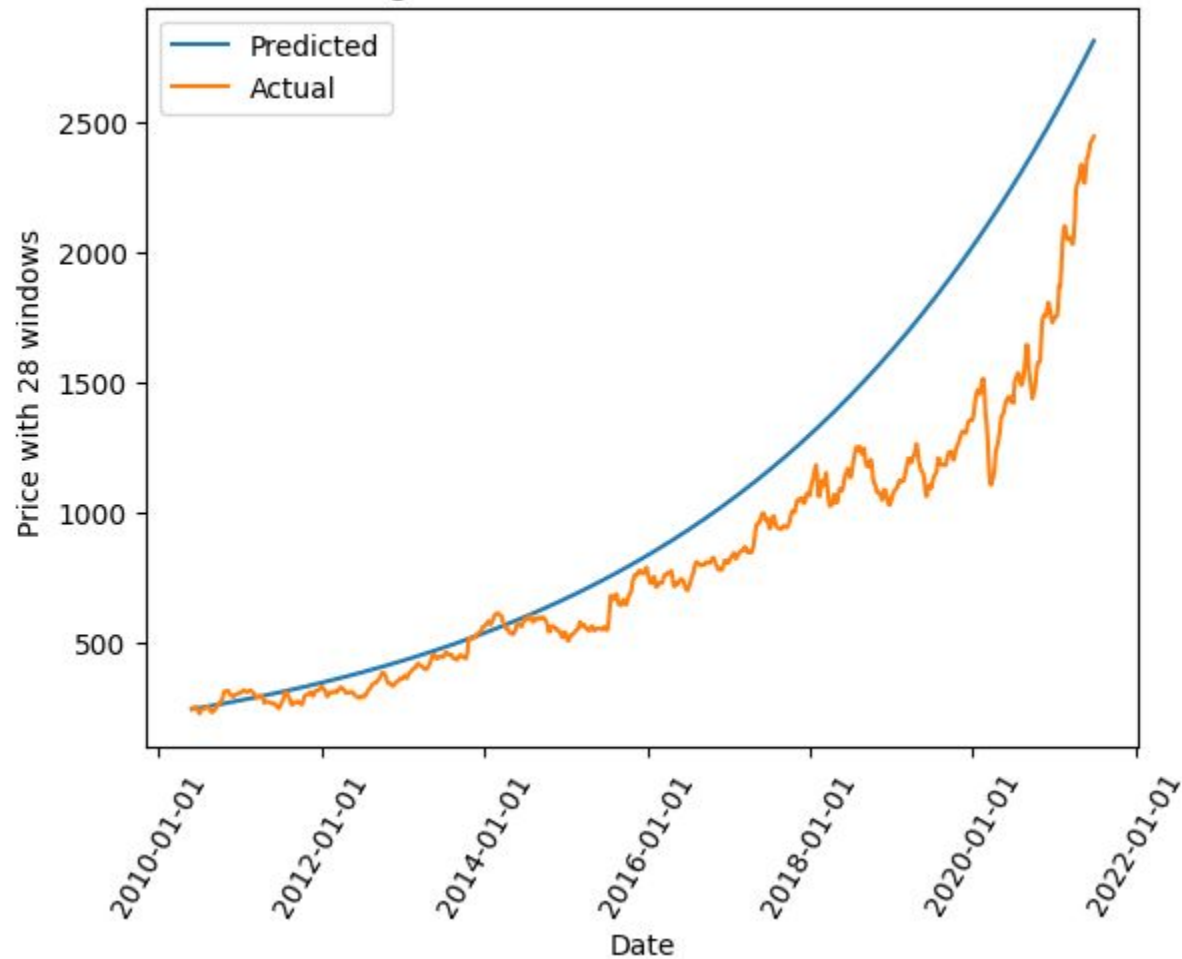
AAPL: Autoregression Prediction for Price with 28 windows



MSFT: Autoregression Prediction for Price with 28 windows



GOOGL: Autoregression Prediction for Price with 28 windows





# Results

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



# Baseline Results

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



# Model Improvement with new features

- Improved Model:
  - Added additional features to the autoregression keeping everything else constant from baseline to see results of adding sentiment to our model.
- Goals:
  - See if there were benefits to isolate positive, neutral, and negative sentiment
  - See if there were benefits to isolate Reddit, Guardian, or NYT data
  - See if neutral sentiment is as useful as more polarized(pos,neg) sentiment



# Improved Results

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26

## Isolating positive, neutral, negative

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26

## Isolating positive, neutral, negative

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment 3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



## Isolating by Source

features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



## Isolating by Source


features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment 3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



# Takeaways

- Overall, adding sentiment produces a noticeable improvement our models accuracy
- There is not a clear combination of sentiment features that produces the most accurate model.



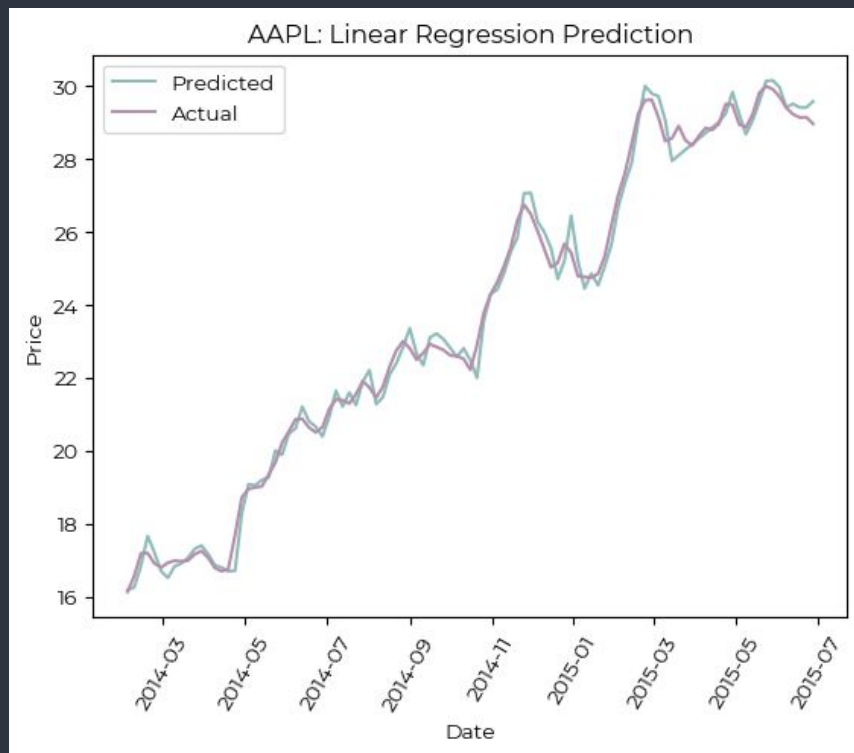


features	AAPL	MSFT	GOOGL
price only(baseline)	\$19.26	\$34.85	\$355.61
price w/ pos sentiment 3 sources	\$12.48	\$13.20	\$171.32
price w/ neg sentiment 3 sources	\$12.74	\$6.28	\$153.18
price w/ neu sentiment 3 sources	\$12.26	\$11.06	\$145.78
price w/ all sentiment 3 sources	\$10.37	\$7.11	\$185.61
price w/ pos neg sentiment 3 sources	\$11.00	\$7.84	\$176.28
price w/ all reddit sentiment	\$11.12	\$16.70	\$153.26
price w/ all guardian sentiment	\$15.69	\$16.14	\$237.89
price w/ all nyt sentiment	\$14.44	\$15.41	\$114.26



Final Model

# Model Results



RMSE: 0.345

$R^2$ : 0.994



# Conclusion & Takeaways

- Our modelling results demonstrate feasibility of price forecasting and related profit-generating methods
- Necessary areas of improvement for more comprehensive and accurate modelling
  - Data
  - Complex modelling techniques
  - Incorporation of more independent variables