Stock Price Prediction Based on Sentiment Analysis on Social Media and News

Abeeku Bruce-Mensah brucemen@usc.edu
Christopher Imantaka imantaka@usc.edu
Haiwen Chen haiwenc@usc.edu
Xiaolin Cheng xcheng71@usc.edu
Minghui Wang minghuiw@usc.edu

Agenda

01 Introduction
- Problem statement
- General Area of Interest
- Why is it interesting?

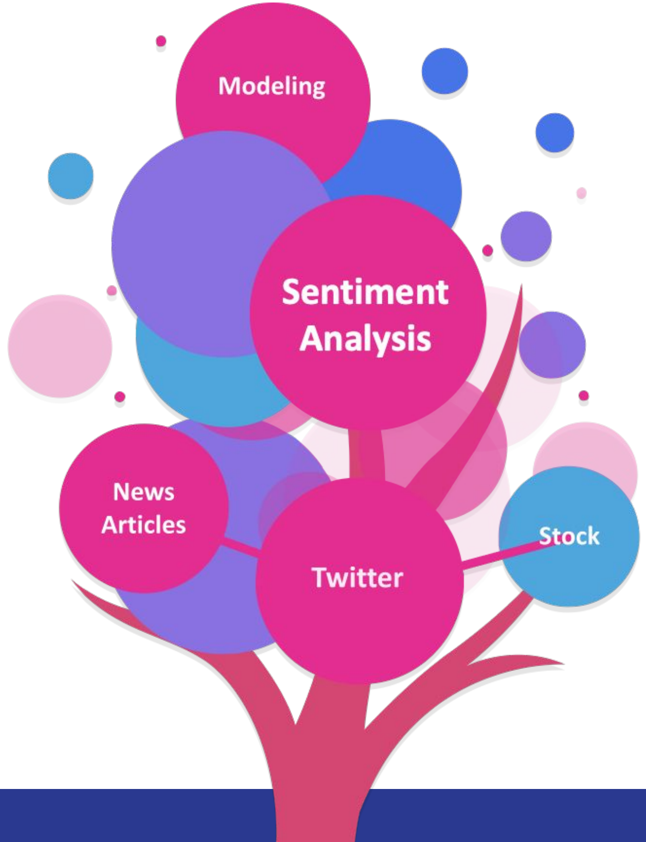
02 Data Collection
- Twitter
- News
- Stock Price
- Sentiment Analysis
    - Vader
    - LIWC

03 Model
- Data used
- The model of choice
- Features Analyzed
- Processed Data
- Trained the Model
- Predictions

04 Results
- AMC
- GameStop
- Nokia
- Interesting Find

05 Conclusions
- Theoretical Perspectives
- Comparison with other papers
- Implications, conclusions, limitations
- Division of Labor

Introduction - Problem Statement



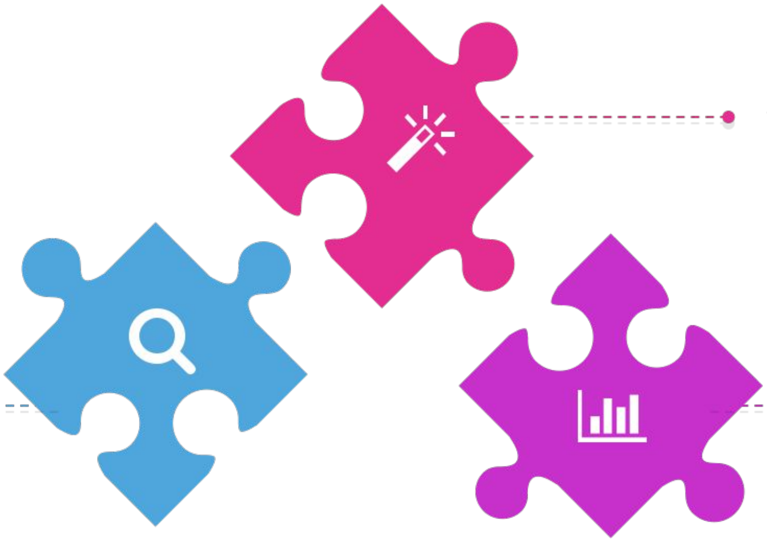
01

Find **correlation** between emotions on **Twitter/News** and **stock price** for specific stocks (GameStop, AMC, Nokia)

02

Upon finding that correlation, **predict** the same day return value based on sentiment analysis

Introduction - General Area of Interest



Emotion & Decision Making

We examined how emotions from social media and News affect investors' decision making

Emotion Recognition

We performed emotion analysis using various approaches form social media and News

Emotion Modeling

We predicted the stock prices based on emotion data

Introduction - Why is this interesting?

Recent Stock Market Spike

Recently certain stocks exploded as a result of social media movement.

Sentiment Analysis

Stock market is a good reflection of how people's emotions affect their investment decisions.

Data Science

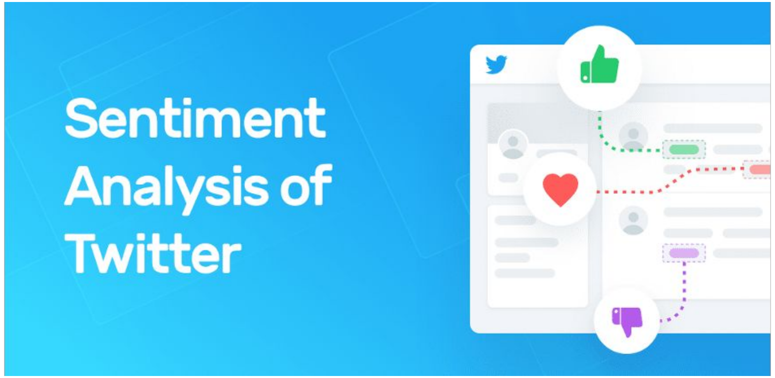
In the era of big data, it's always fun to find correlations from seemingly irrelevant areas.

Money!

Successfully building a highly accurate prediction model can generate revenue!

Data Collection - Twitter

- We filtered the irrelevant tweets using hashtags.
- We kept the emojis for sentiment analysis.
- We filtered for English tweets only.



Search By Keyword

We mainly used the names of the stocks as keywords to search for matching tweets.

Search By Date

We selected dates based on when the spike occurred and collected twitters before and after the spike.

Data Collection - Stock price data

- Started with Google Finance
- Yahoo finance was better programmatically

Data Collection - News Search

- Searched for credible news sources: Bloomberg, ny times, yahoo finance
- Searched for ease of dataset access with the various articles and pre-existing apis

- Nexis Uni
  - Better filtering
  - More easily scrapable
  - No existing API

- Google News
  - May include more diverse investor sentiments due to more articles
  - Readily available python API
  - Data was very noisy and require much post processing

Nexis Uni ended up being the better option

- More easily integratable due to less post processing
- Better filtering to enable a better investor sentiment analysis

Nexis Uni™

Data Collection - Sentiment Analysis

VADER (Valence Aware Dictionary for Sentiment Reasoning)

- It can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and much more.
- It works exceedingly well on social media type text, yet readily generalizes to multiple domains
- VADER can work with multiple domains.

Tweets	Results
One of my local GameStop stores is closing 😖	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
i remember when i wanted a ps4 and this dude was waiting outside of gamestop and he told me they didn't have any and i was instantly like nah you lying was in total defense mode	{'neg': 0.117, 'neu': 0.779, 'pos': 0.104, 'compound': -0.1298}
Thts why I didn't even bother with gamestop honestly, I was tired of them	{'neg': 0.275, 'neu': 0.57, 'pos': 0.155, 'compound': -0.3182}
I already have MK8 deluxe though 😭😭😭	{'neg': 0.437, 'neu': 0.563, 'pos': 0.0, 'compound': -0.8519}
I already have MK8 deluxe though	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Data Collection - Sentiment Analysis

(LIWC) Linguistic Inquiry and Word Count

- It calculates the percentage of words in a given text that fall into one or more of over 80 linguistic, psychological and topical categories indicating various social, cognitive, and affective processes.
- LIWC2015 dictionary

The largest movie chain operator has been caught up in the same retail-driven short squeeze that has pushed shares of GameStop Corp (NYSE:GME) into the stratosphere. StreetInsider said Wednesday that TD Ameritrade (NASDAQ:AMTD), the online broker, had put restrictions on trading in AMC and GME. The shares were also halted for trading on the stock exchange several times on Wednesday. Earlier this week, AMC's CEO Adam Aron said bankruptcy was off the table because the company was able to raise $917 million in cash from stock sales and debt deals since mid-December. Like other movie chains, AMC has struggled with pandemic-related business shutdowns and is trying to ride out the bad times, hoping moviegoers return to the theaters soon. Investing.com offers an extensive set of professional tools for the financial markets. Read more News on Investing.com and download the new Investing.com apps for Android and iOS! Load-Date: January 27, 2021

WC	Analytic	Clout	Authentic	Tone
159	98.26	69.26	52.56	36.86

affect	posemo	negemo	anx	anger	sad
3.14	1.89	1.26	0.63	0	0

focuspast	focuspresent	focusfuture
5.66	3.77	1.26

Model

- Data used
  - the closing price value of each day for AMC, Gamestop and Nokia as a predictive value
  - Sentiment analysis data got from vader and LIWC2015
- The model of choice
  - Linear regression
  - Reason : prediction will be concrete stock price for each day

Model

- Features selected:
  - For analysis from vader
    - One group of features:
      - Negative
      - Neutral
      - Positive
      - Compound
  - For analysis from LIWC2015
    - Two groups of features:
    - Psychological Processes
      - Affective process
      - Positive emotion
      - Negative emotion
      - Anxiety
      - Anger
      - Sadness
    - Drives
      - Drives
      - Affiliation
      - Achievement
      - Power
      - Reward

scores
{'neg': 0.0, 'neu': 0.722, 'pos': 0.278, 'compound': 0.9226}

Category	Abbrev	Examples
Psychological Processes		
Affective process	affect	happy, cried
Positive emotion	posemo	love, nice, sweet
Negative emotion	negemo	hurt, ugly, nasty
Anxiety	anx	worried, fearful
Anger	anger	hate, kill, annoyed
Sadness	sad	crying, grief, sad
Drives	drive	
Affiliation	affiliation	ally, friend, social
Achievement	achieve	win, success, better
Power	power	superior, bully
Reward	reward	take, prize, benefit
Risk	risk	danger, doubt

Model

- Preprocess
  - compress the data using 4 kinds of baselines
    - Mean
    - Median
    - Max
    - Min
  - normalize the stock value
- Train the model
  - Data:
    - Training data: all data except the last 10 data
    - Test data: the last 10 piece of data
  - 2 ways for each group of features
    - Cumulative (integrated)
    - Separately (train on separate feature)

```
#mean
amc_vader_df_mean = amc_vader_df.groupby(['Date']).agg({'neg':np.mean,'neu':np.mean,'pos':np.mean,'com':np.mean}).
#median
amc_vader_df_median = amc_vader_df.groupby(['Date']).agg({'neg':np.median,'neu':np.median,'pos':np.median,'com':np
#max
amc_vader_df_max = amc_vader_df.groupby(['Date']).agg({'neg':np.max,'neu':np.max,'pos':np.max,'com':np.max}).reset
#min
amc_vader_df_min = amc_vader_df.groupby(['Date']).agg({'neg':np.min,'neu':np.min,'pos':np.min,'com':np.min}).reset
```

```
#mean
reg_amc_mean = LinearRegression().fit(amc_merge_df_mean[['neg','neu','pos','com']][:-10], amc_merge_df_mean[['clos
#median
reg_amc_median = LinearRegression().fit(amc_merge_df_median[['neg','neu','pos','com']][:-10], amc_merge_df_median[
#max
reg_amc_max = LinearRegression().fit(amc_merge_df_max[['neg','neu','pos','com']][:-10], amc_merge_df_max[['close']
#min
reg_amc_min = LinearRegression().fit(amc_merge_df_min[['neg','neu','pos','com']][:-10], amc_merge_df_min[['close']
```

```
#Training models based on the amc data for four features(neg,neu,pos,com) separately
reg_amc_neg_mean = LinearRegression().fit(amc_merge_df_mean[['neg']][:-10], amc_merge_df_mean[['close']][:-10])

reg_amc_neu_mean = LinearRegression().fit(amc_merge_df_mean[['neu']][:-10], amc_merge_df_mean[['close']][:-10])

reg_amc_pos_mean = LinearRegression().fit(amc_merge_df_mean[['pos']][:-10], amc_merge_df_mean[['close']][:-10])

reg_amc_com_mean = LinearRegression().fit(amc_merge_df_mean[['com']][:-10], amc_merge_df_mean[['close']][:-10])
```

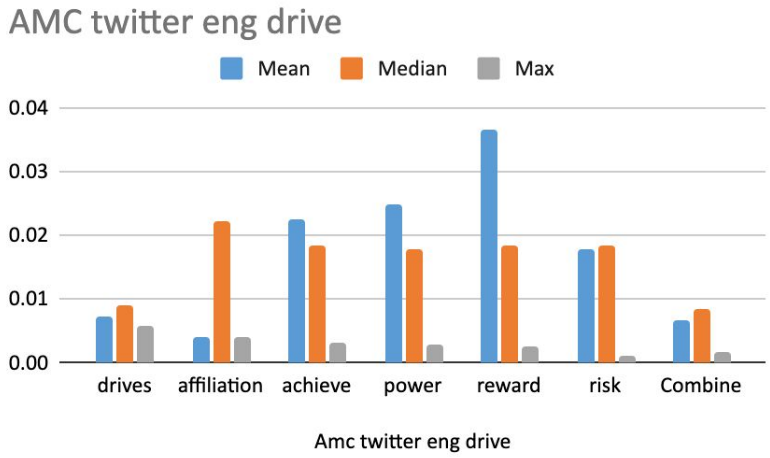
Model

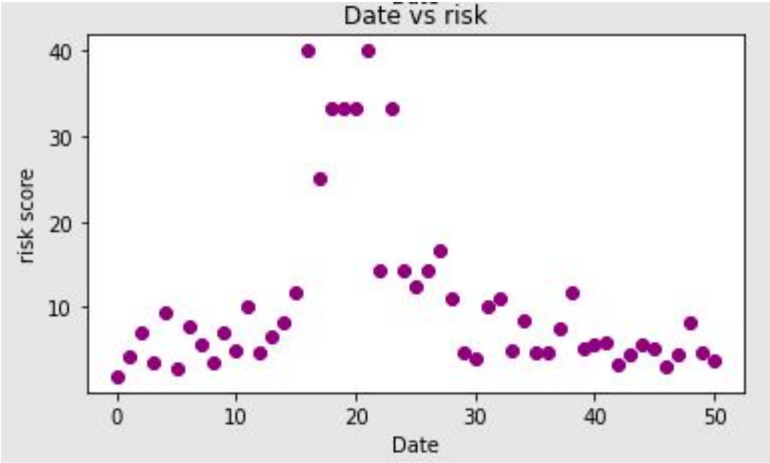
- Prediction
  - Metric
    - Mean squared error

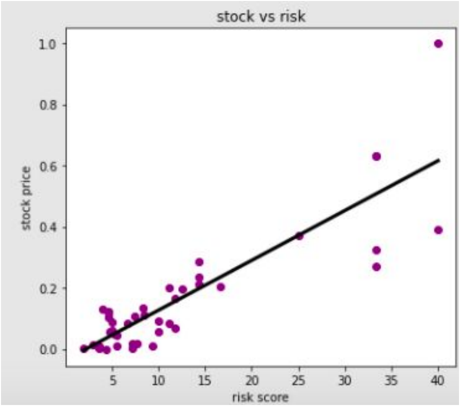
```
#This is the score of the prediction/ performance of the model
#mean
amc_score_mean = mean_squared_error(amc_merge_df_mean[['close']][-10:], pred_amc_mean)
#median
amc_score_median = mean_squared_error(amc_merge_df_median[['close']][-10:], pred_amc_median)
#max
amc_score_max = mean_squared_error(amc_merge_df_max[['close']][-10:], pred_amc_max)
#min
amc_score_min = mean_squared_error(amc_merge_df_min[['close']][-10:], pred_amc_min)
[amc_score_mean,amc_score_median,amc_score_max,amc_score_min]
```

AMC Result

- LIWC analysis and Twitter data
- Max baseline
- Risk feature
- MSE of 0.00101

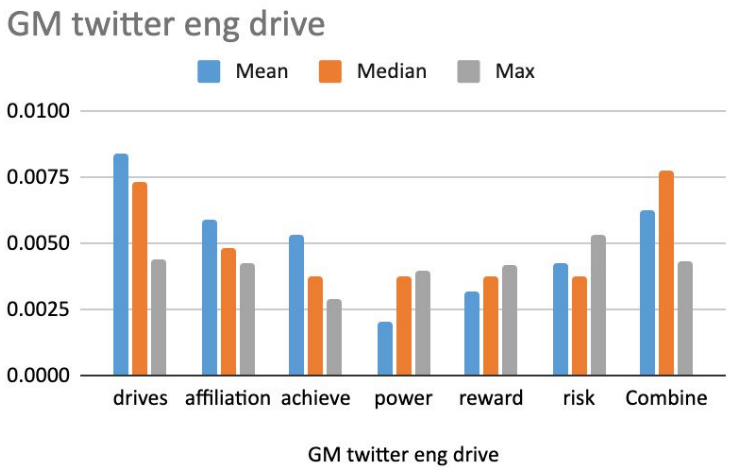


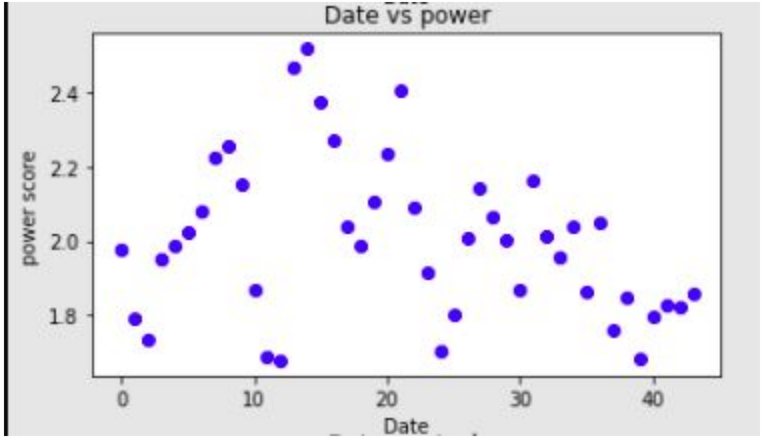


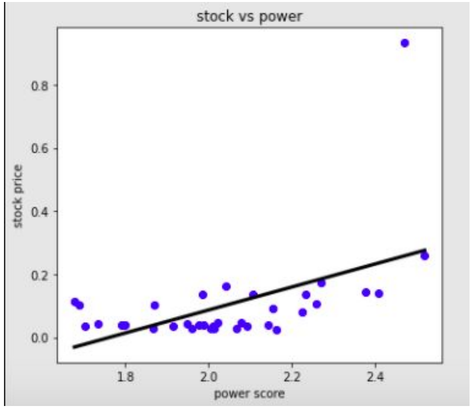


Gamestop Result

- LIWC analysis and Twitter data
- Mean baseline
- Power feature
- MSE of 0.00205



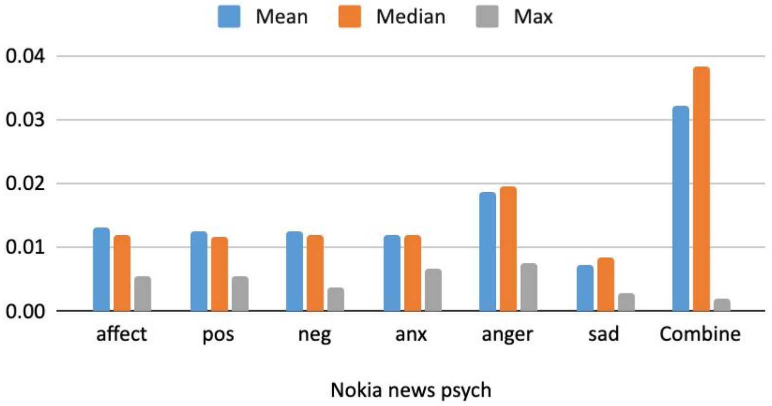


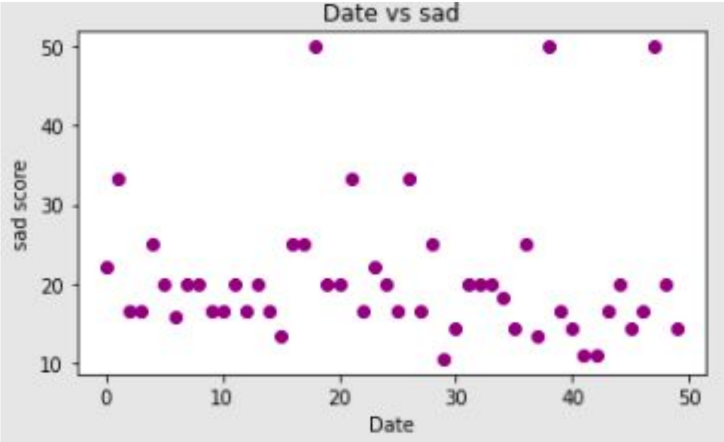


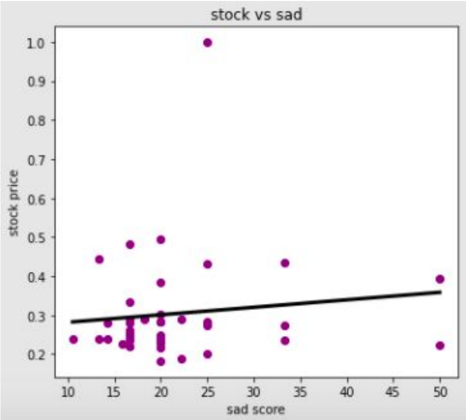
Nokia Result

- LIWC analysis and News data
- Max baseline
- Sad feature
- MSE of 0.00285

Nokia news psych

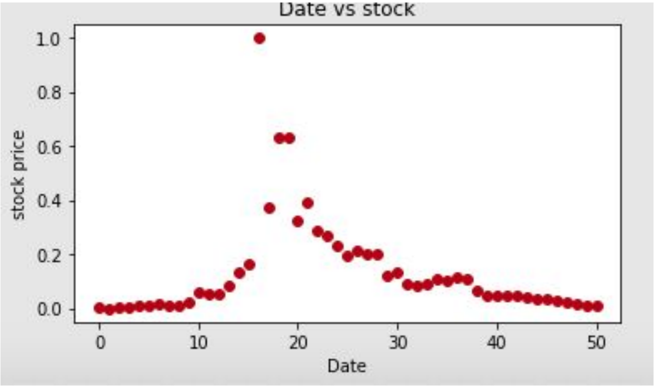


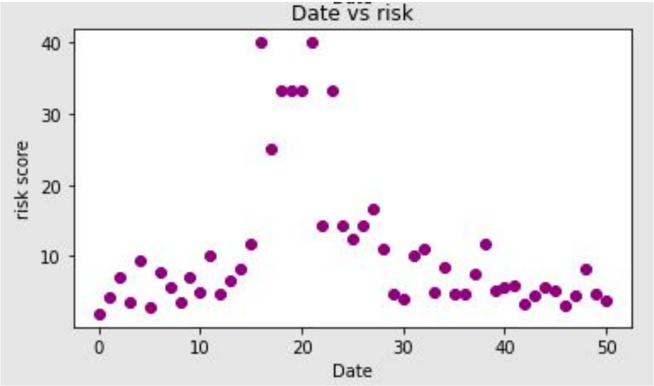


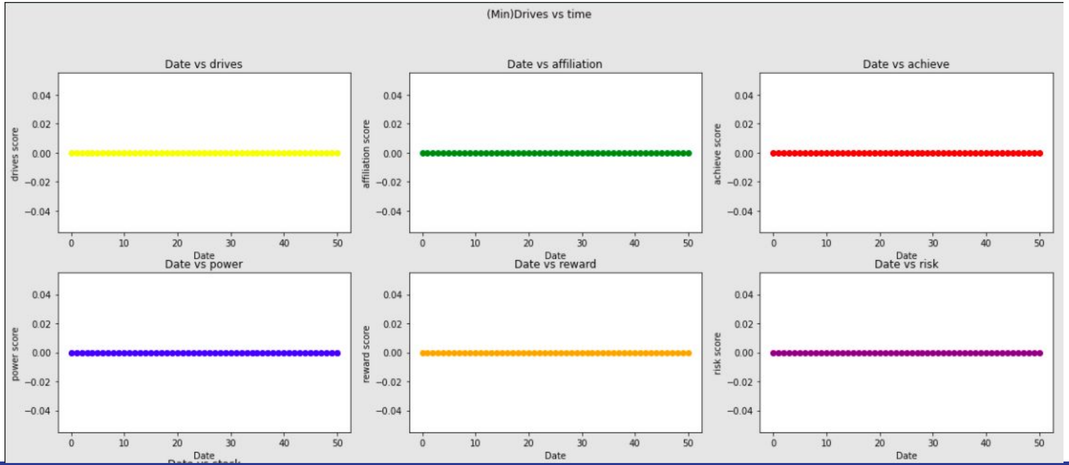


Interesting Find

- Minimum baseline is not useful
- Max baseline is the most accurate
- High correlation = high prediction

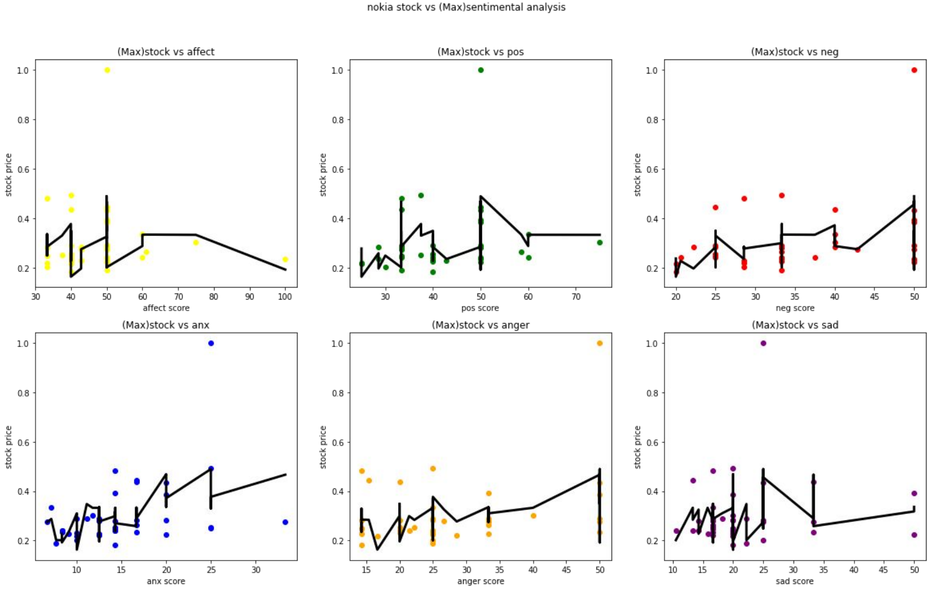


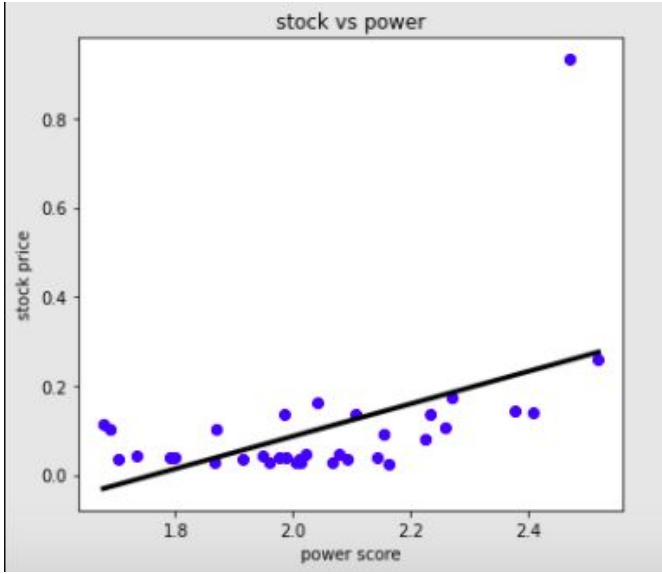




Flaw in the model

- Unable to accurately predict fluctuation
- Works well with short time frame
- Unable to interpret combine feature prediction





Theoretical Perspectives

- In class we learned about sentiment analysis and determining affect through text
- People have already been able to determine useful applications:
  - Business intelligence
  - Informing consumers
  - Opinion polling
  - Prediction
  - Health
- Our goal was to expound upon prediction

Comparison with other papers

- *Sentiment analysis of Twitter data for predicting stock market movements*
  - N-gram, word2vec
  - Strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets
- *Can facebook predict stock market activity?*
  - Facebook's Gross National Happiness index (GNH: a measure of happiness based on the sentiment analysis from facebook statuses in a nation) an effective measure of investor sentiment.
  - Change in GNH is directly related to change in stock price & is a good measure of investor sentiment.
- *Trading on twitter: The financial information content of emotion in social media.*
  - Evaluated all S&P 500 stocks with retweet and follower data
  - Posts with more retweets and accounts with more followers tend to increase the accuracy for predicting the same day and future returns

Conclusions & Implications

- Use the emotional polarity and intensity of social media tweets to predict the same day stock market price--compare MSE
- The max baseline of sentiment analysis scores (the most polarized emotions among tweets) has the best predictability.
- Compared to VADER, LIWC2015 had better performance when predicting stock market price.
- Future directions: incorporating more features such as the number of followers and the speed of information dissemination to construct a better model

Roles & Team members

- Haiwen Chen - Introduction, extract and process Twitter Data, comparare our model with others
- Abeeku Bruce-Mensah - Extract stock market data and news article for gamestop, amc, and nokia
- Minghui Wang - Sentiment Analysis (LIWC & VADER)
- Christopher Imantaka & Xiaolin Cheng - Train models, make predictions and analyze the results

References

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (pp. 1345-1350). IEEE.

Sul, H., Dennis, A. R., & Yuan, L. I. (2014, January). Trading on twitter: The financial information content of emotion in social media. In *2014 47th Hawaii International Conference on System Sciences* (pp. 806-815). IEEE.

Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24-54.

Umar, Z., Gubareva, M., Yousaf, I., & Ali, S. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. *Journal of Behavioral and Experimental* Finance, 100501.

Vicki Liu, Carmen Banea, Rada Mihalcea, "Grounded emotions", *Affective Computing and Intelligent Interaction (ACII) 2017 Seventh International Conference on*, pp. 477-483, 2017.

Karabulut, Yigitcan. “Can Facebook Predict Stock Market Activity?” SSRN Electronic Journal, 2012, doi:10.2139/ssrn.2017099.

Chen, Mu-Yen, et al. “Modeling Public Mood and Emotion: Stock Market Trend Prediction with Anticipatory Computing Approach.” Computers in Human Behavior, vol. 101, 2019, pp. 402–408., doi:10.1016/j.chb.2019.03.021.

Q & A