Sentiment Analysis of Twitter Data for Predicting Stock Price

TEAM BOXELDER

WEI XI: XIWEI26@RAMS.COLOSTATE.EDU

XIANG LI: RAVNO@RAMS.COLOSTATE.EDU

Introduction

- Many factors that may influence the stock price
- Public opinions really matters
- Relation between tweets and price change

It’s a big data problem because…

- Nearly 3.5 million active users of twitter
- People comment on all different kind of things

Our approach

Dataset:

- Trademark List
  - 4 companies(Apple, Amazon, Intel, Microsoft)
  - From official website
- Historical stock price dataset
  - 4 companies(Apple, Amazon, Intel, Microsoft)
  - Yahoo finance

```
Date,Open,High,Low,Close,Adj Close,Volume
2009-06-01,19.495714,19.998571,19.428572,19.907143,17.84543,113124900
2009-06-02,19.855715,20.191429,19.764286,19.927143,17.863356,114055900
2009-06-03,20.0,20.158571,19.867144,20.135714,18.050327,141299900
2009-06-04,20.018572,20.597143,20.005714,20.534286,18.407618,137658500
2009-06-05,20.758572,20.914286,20.45857,20.667143,18.526722,158179000
2009-06-08,20.545713,20.604286,19.918571,20.549999,18.421707,232913100
```

- Twitter7 dataset --from Stanford Large Network Data Collection
  - 476 million tweets, 25GB
  - three parts: time, user and tweets.

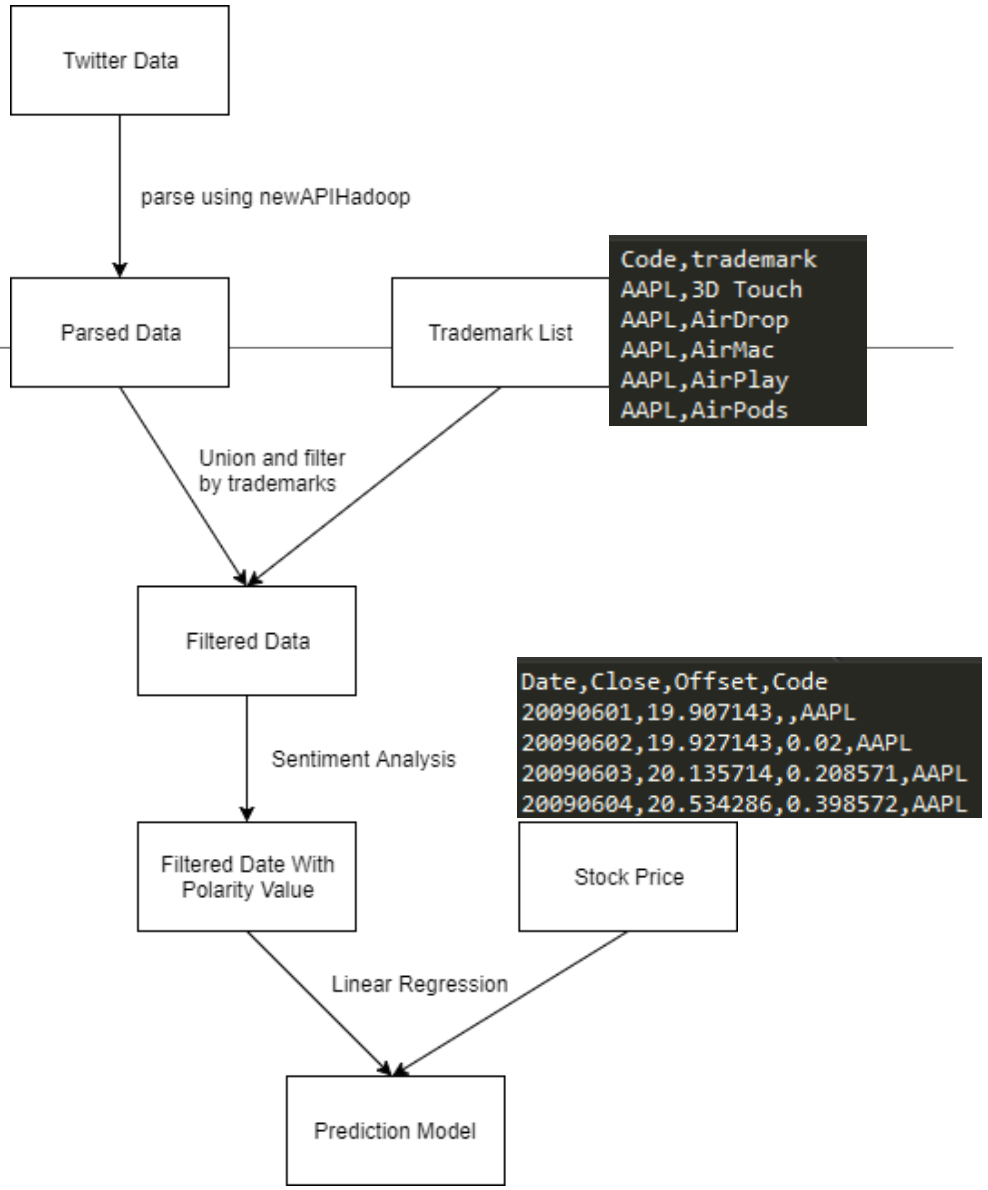
```
T       2009-11-30 21:03:12
U       http://twitter.com/janeadarby
W       Gettin my good good perm!

T       2009-11-30 21:03:13
U       http://twitter.com/philuponem
W       @LovelyLittleLey lmaooo that's a compliment in my eyes thnx
```

Our approach

Methodology:

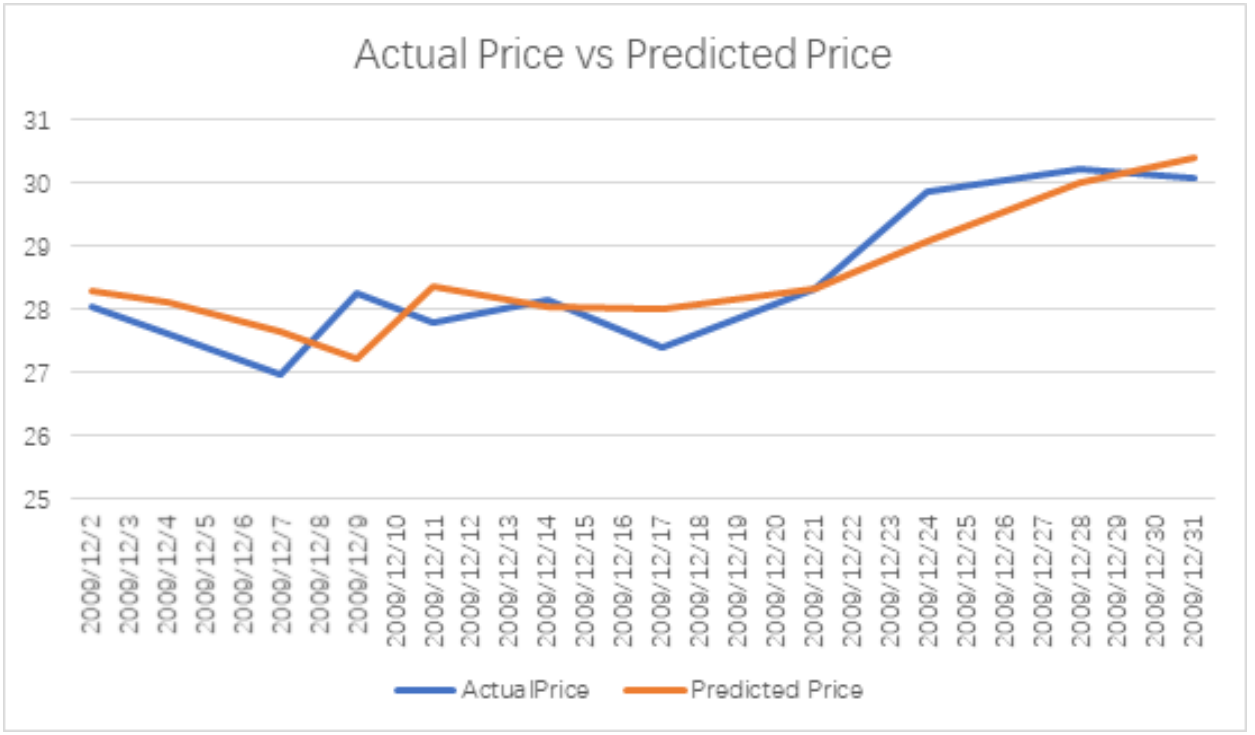
- Load data as Dataframes in Spark
- MapReduce to parse data
  - newAPIHadoop in spark
  - customInputFormat
- Stanford nlp to calculate sentiment value
  - Use user defined function (udf) to extract column value
- Spark Mllib to build model



Final Software

- Input:
  - A stock company name
  - Dates (Start Date & End Date)
- Output:
  - A csv file contains stock date, Code, Close Price, Predicted Price and Difference
  - A Graph contains both actual and predicted stock price
  - Price trends line

```
Stock Date,Code,Close,Predicted Price,Diff
20091202,AAPL,28.032858,28.301181767182122,0.26832387228466104
20091203,AAPL,28.068571,28.172936004266877,0.10436491356863442
20091204,AAPL,27.617144,28.11941073127754,0.5022671002960948
```



Result and Evaluation

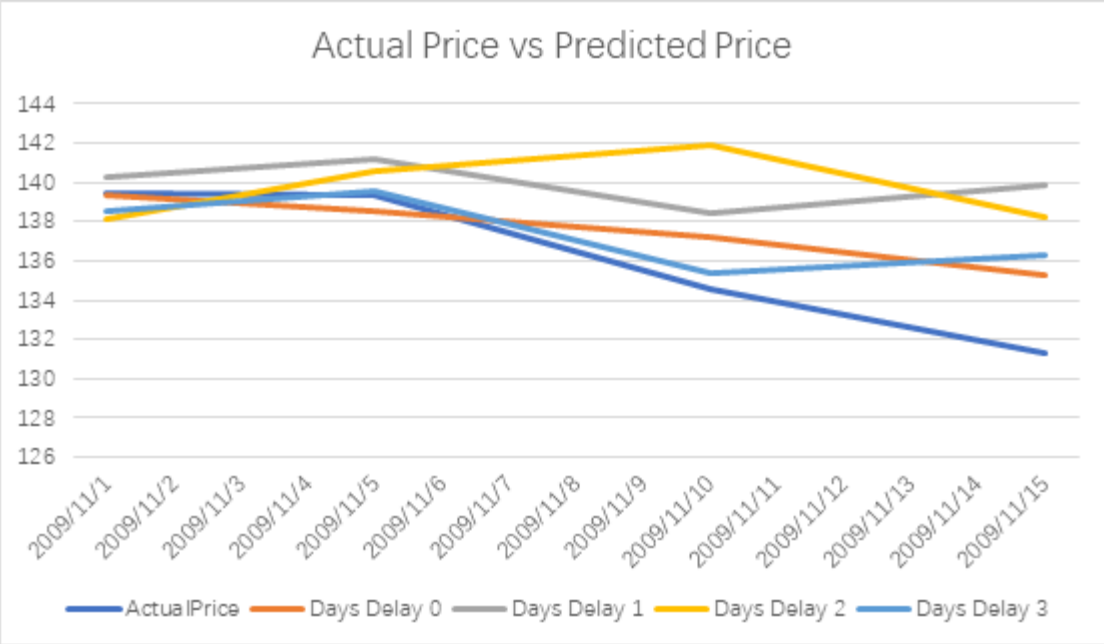
- Coefficient of Determination
  - How well a linear regression model fits the data (lies between 0 – 1)
  - It can always be increased by adding more variables into the model

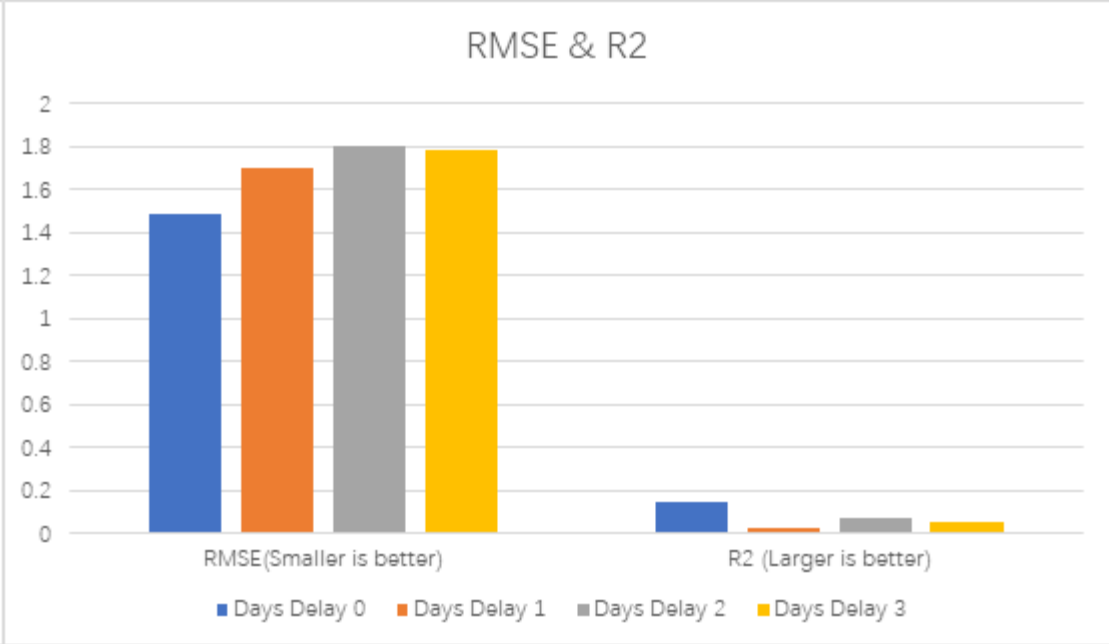
$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},$$

- Root Mean Square Error

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}.$$

Result and Evaluation





Conclusion

- There exists some correlations between stock price and tweets
  - Positive consensus has a greater possibility to cause the stock price go up
  - Vice versa

- Possible improvements
  - Analysis on twitter users
    - Set individual weight for users
    - User that has more followers may impact the stock price more (e.g. Donald Trump)
  - Trademark list
    - Run analysis to remove unrelated tweets and get most popular trademarks
  - More features to train
    - Ex. Opening price and volume

Q & A