Predicting a stock's closing price from news sentiment

By Anthony

Introduction

- Build NLTK Model to predict the news sentiment
- Scrape news articles of a stock and run NLTK Model
- Build an ML Model to predict a stock's closing price by 12PM

NLTK Model

- Acquired two datasets with labeled sentiment of financial news articles
- Used regex to tokenize words, lower capitalization, remove stop words and reduce to stem words.
- TF-IDF function to transform the data to reflect how important a word is in the collection of descriptions in each sentiment rating.

NLTK Models

Model	Training Accuracy	Test Accuracy
LinearSVC	95.73%	61.42%
LinearSVC Tuned	95.73%	61.42%
KNN	64.30%	55.38%
MultinomialNB	60.04%	55.91%
RandomForestClassifier	100%	61.42%
SVM SVC	100%	61.42%
Gradient Boosting	65.52%	60.37%
Adaboosting Classifier	54.92%	50.39%
XGB Classifier	76.31%	61.94%
Voting Classifier	**83.39 %**	**64.57%**

NLTK Model Example

"Amazon beat on earnings and revenues—here's what three experts are saying about the stock now Amazon beat on earnings and revenues—here's what five experts are saying about the stock now. Amazon cloud-computing division said revenue jumped 45 percent in the fourth quarter, as the company continued to cement its lead over Microsoft and Google. Sales at Amazon Web Services climbed to $7.43 billion from $5.11 billion a year ago, topping the $7.29 billion consensus estimate among analysts polled by FactSet. AWS revenue represented 10 percent of total quarterly sales at Amazon."

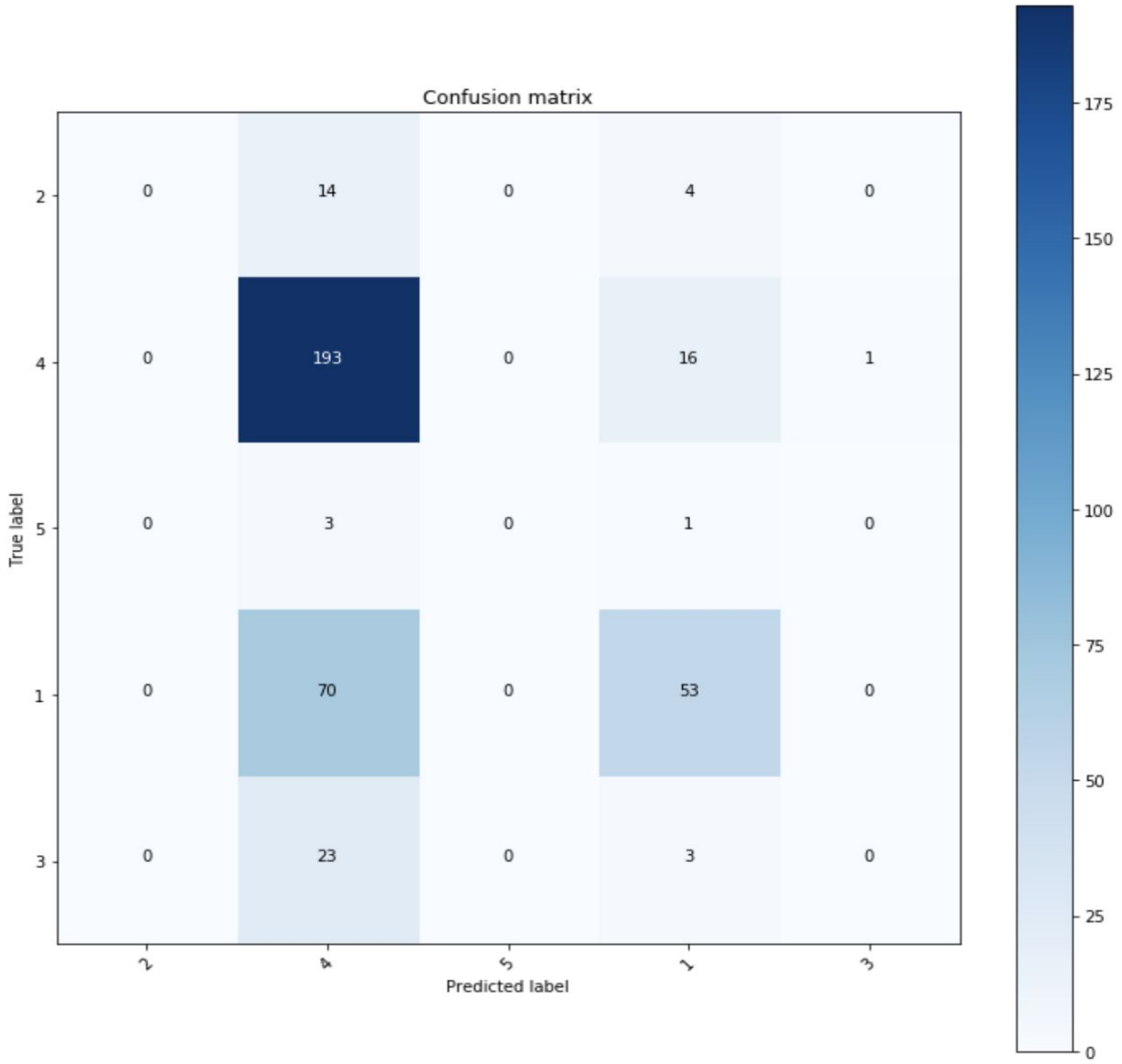
```
vc_predict_app(x)
```

```
4
```

Negative	Weakly Negative	Positive	Weakly Positive	Neutral
1	2	3	4	5

NLTK Best Model

Confusion Matrix



Scraping Data & API Calling

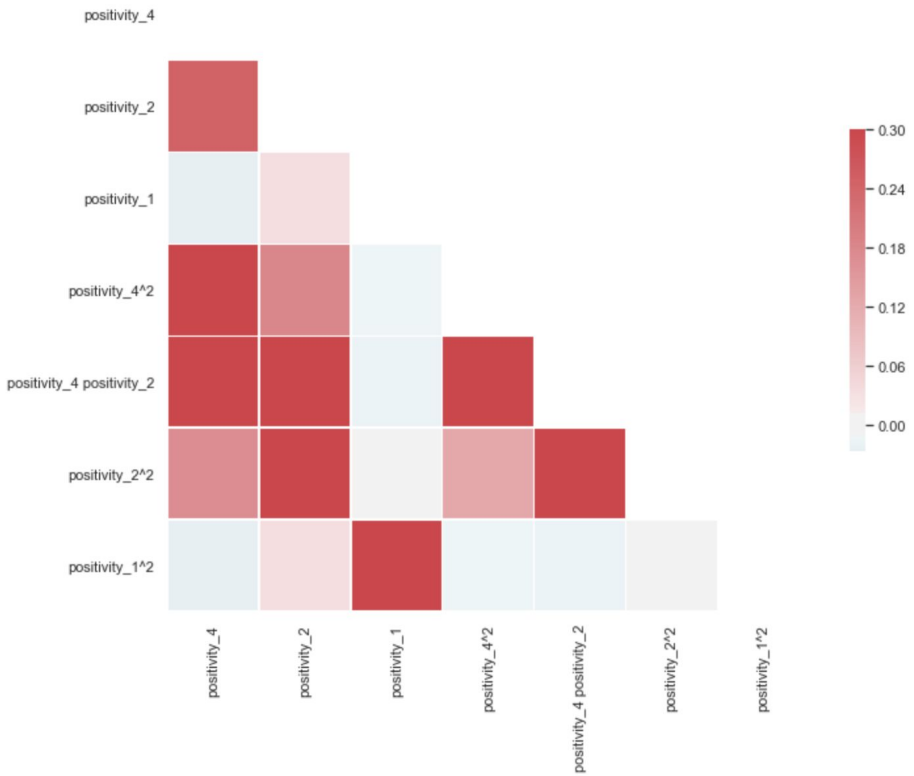
- Used Beautifulsoup and Selenium to scrape 1500 articles from CNBC of Microsoft Corp
- Used Alpha Advantage API to retrieve opening and closing prices of Microsoft Corp
- Run NLTK Model on each to article to get the predicted sentiment rating

Models Before Feature Engineering

Model	Training Accuracy	Test Accuracy
LinearSVC	52.76%	54.64%
KNN	51.03%	42.27%
Log Regression	52.76%	54.64%
RandomForestClassifier	56.90%	53.61%
SVM SVC	56.90%	53.61%
Gradient Boosting	54.83	53.61%
Adaboosting Classifier	56.21%	**55.67%**
XGB Classifier	55.86%	**55.67%**
Voting Classifier	56.55%	**55.67%**

Feature Engineering & Selection

up	positivity_4	positivity_3	positivity_2	positivity_1	positivity_5
True	7	0	4	0	0
True	8	0	3	0	0
False	3	0	7	0	0



Models After Feature Engineering

Model	Training Accuracy	Test Accuracy
LinearSVC	**53.31%**	**58.70%**
Log Regression	56.90%	55.67%
Adaboosting Classifier	56.90%	55.67%

Mean Value of the Target Variable: **53.23%**

Best Model Accuracy: **58.70%**

Result: The model was **9.08%** better at predicting the closing price of a stock than random guessing.

Conclusion & Thank You

- Get a larger dataset to Improve NLTK Model
- Improve ML Models
- Add sentiment of news articles of the general economy