Stock market Prediction based on Daily News Headlines Summary

Shaohang Hao, Weikun Hu,
Ji Peng and Ruiyu Zeng

CSE 538

2019/12/2

Boeing 737 Max crash revelations could cost shareholders $53 billion

BY STEPHEN GANDEL
OCTOBER 21, 2019 / 5:19 PM / MONEYWATCH

Fitbit surges 17% after Google agrees to buy the company for $2.1 billion (FIT)

Daniel Strauss
Nov. 1, 2019, 10:08 AM

SHARE

FITBIT (FIT) STOCK NYSE

ADD
SHARE

7.21 USD 1.00 (16.10%) Pre-market 09:17:25 AM EDT BTT

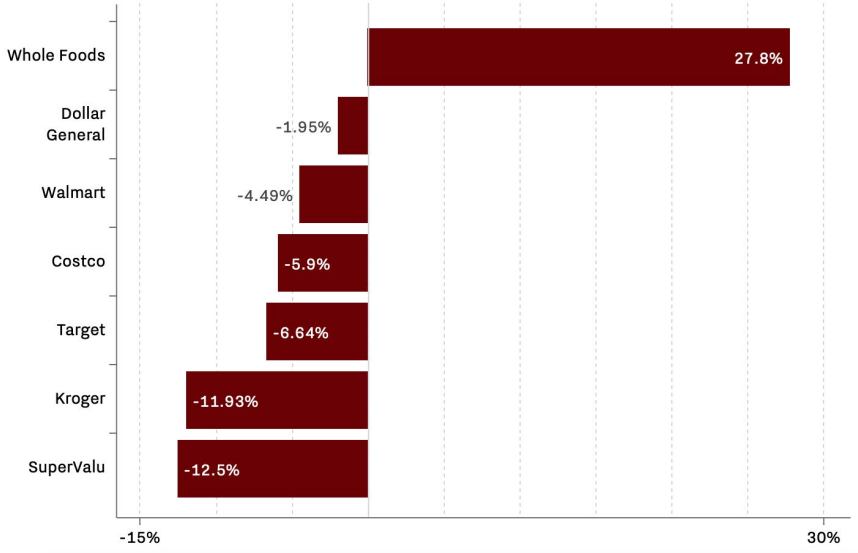
6.18 USD 0.32 (5.46%) Official Close 10/31/2019 NYSE

Prev. Close	6.18	Market Cap (USD)	1.40 B	Day Low	Day High	52 Week Low	52 Week High
Open	-	Volume (Qty.)	-	-	-	2.81	6.96

INTRADAY 1W 1M 3M 6M YTD 1Y 3Y 5Y 10Y MAX INDICATORS CHART OPTIONS

Look what happened to grocery stocks after Amazon announced it's buying Whole Foods

Grocery chain share price percentage change on Jun. 16



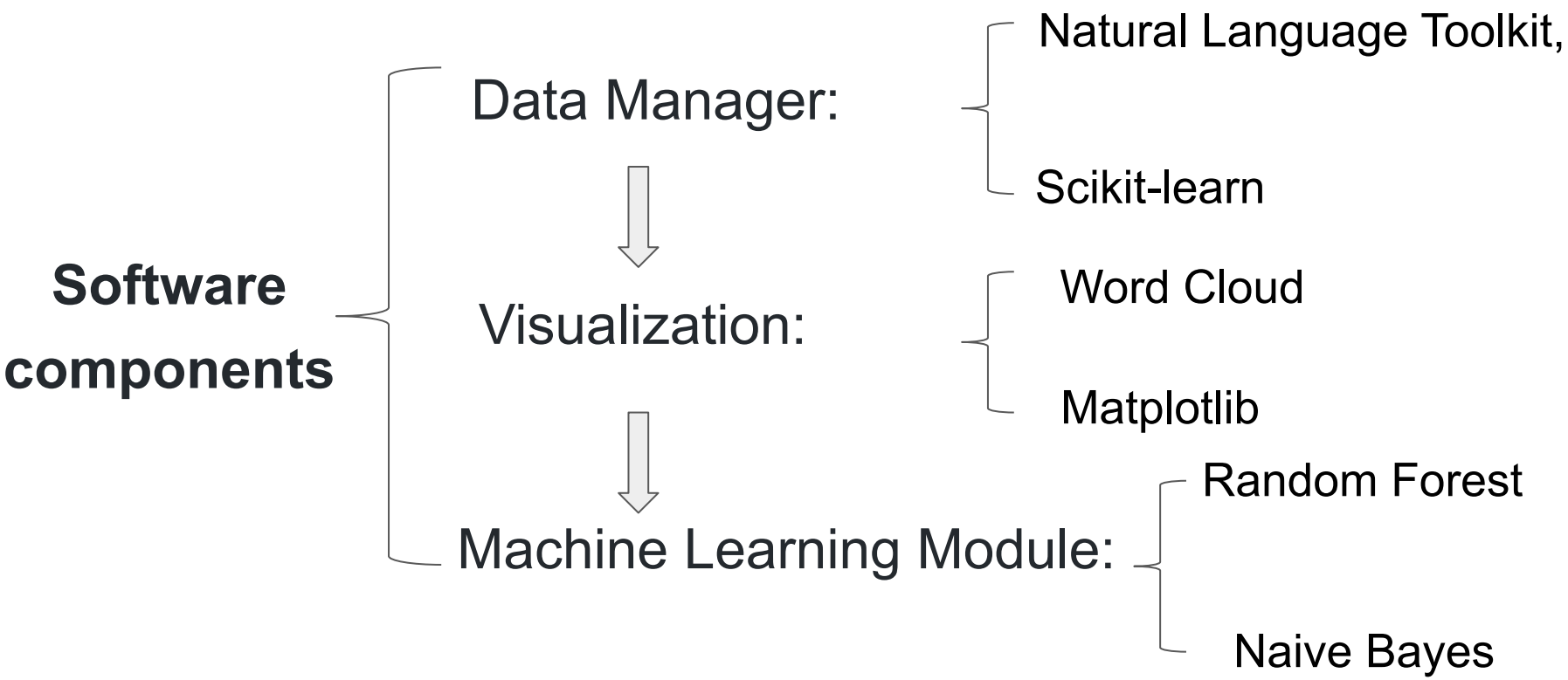
Background

Use Daily News to Predict Stock Market Performance

- **News data** was obtained from Reddit WorldNews Channel (/r/worldnews). Top 25 headlines were voted by reddit users for a single date. (Range: 2008-06-08 to 2016-07-01)
- **Stock data**: Dow Jones Industrial Average (DJIA) is used as the label to supervise model training . (Range: 2008-08-08 to 2016-07-01)
- Training Set: Data from 2008-08-08 to 2014-12-31 (80%)

  Test Set: The following two years data (from 2015-01-02 to 2016-07-01). (20%)
- Accuracy will be used as the evaluation metrics

Project Structure: Components Specification



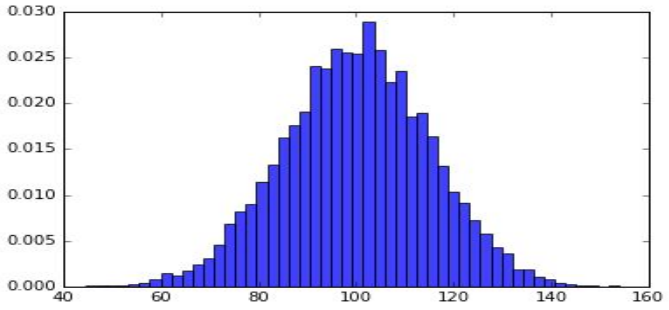
Python Libraries used in this project

NLTK (Natural Language Toolkit): Preprocessing: tokenization, stopwords, stemming

Scikit-learn: Word embedding (Vectorization): TfidfVectorizer,CountVectorizer

Matplotlib: Primarily used for 2D visual representation of data distribution

Word Cloud: Size of each word indicates its frequency or importance.



Machine Learning Module

Naive Bayes

```
# Word embedding for training and testing set
tfidf = TfidfVectorizer(min_df=0.1, max_df=0.7, max_features = 200000, ngram_range = (1, 1))
tfidf_train = tfidf.fit_transform(trainheadlines)
tfidf_test = tfidf.transform(testheadlines) #
print(tfidf_train.shape)
print(tfidf_test.shape)
```

```
(1611, 529)
(378, 529)
```

```
advancedmodel = MultinomialNB(alpha=0.01)
advancedmodel = advancedmodel.fit(tfidf_train, train["Label"])
preds = advancedmodel.predict(tfidf_test)
acc=accuracy_score(test['Label'], preds)
```

```
acc # the accuracy score of the naive bayes model where no stemming and processing is applied to t
he training and testing set
```

```
0.5132275132275133
```

Random Forest

```
advancedmodel = RandomForestClassifier()
advancedmodel = advancedmodel.fit(advancedtrain, train["Label"])
advancedtest = advancedvectorizer.transform(testheadlines)
preds6 = advancedmodel.predict(advancedtest)
acc6 = accuracy_score(test['Label'], preds6)
```

```
/Users/pj/miniconda3/lib/python3.7/site-packages/sklearn/ensemble/
e default value of n_estimators will change from 10 in version 0.2
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
print('RF 1 accuracy: ', acc6)
```

```
RF 1 accuracy:  0.5370370370370371
```

Design: Components Specification

Interaction with the use cases

Function: Stock price prediction;

Inputs: Daily top 25 news headlines;

Outputs: Visualization of stock price trend;

Final Results: A prediction on the following day stock price between “0” to “1” to represent its trending.

Demo

Lesson Learned

Every steps to solve a natural language processing problem with machine learning technology

- Word embedding (Vectorization)
- Different machine learning models: Logistic Regression, Random Forest, Naive Bayes
- Data Visualization.
- Evaluation metrics for machine learning.

Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp+tn}{tp+fp+tn+fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp+fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn+fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp+fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp+tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2*p*r}{p+r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp*tn}$	This metric is used to maximize the *tp* rate and *tn* rate, and simultaneously keeping both rates relatively balanced

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

Future Work

- Possibly improve performance of models by preprocessing text data with natural language toolkit

- Besides using the News to predict same-day stock market, probe the influence of News to second-day stock market(News after stock closing, delay of the reddit voting)

Thank You!

Happy Holidays!!!