# Stock Market Prediction Using Natural Language Processing

-Arnav Dahal
-Hassan Pasha
-Manoj Kumar Gunasekaran
-Niharika Balachandra
-Sathvik Raju
-Yi-Huan Chen

# INTRODUCTION:

- Natural Language Processing : Attempting to discover patterns and ability to manipulate the human language by a computer.
- Stock Market Prediction is one of the most famously researched areas that takes the help of Machine learning to predict the rise and fall of a stock based on past data.

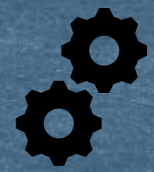
# DATASET

- The data set in consideration is a combination of the world news and stock price shifts available on Kaggle.
- There are 25 columns of top news headlines for each day in the data frame.
- Data ranges from 2008 to 2016 and the data from 2000 to 2008 was scrapped from Yahoo finance.
- Labels are based on the Dow Jones Industrial Average stock index.
- Class 1→ the stock price increased.
- Class 0→ the stock price stayed the same or decreased.

# Data Wrangling

- The data has a lot of stopwords. (Words like a, the, you doesn't help in predicting a stock!)
- Convert all the words to lowercase.
- Remove punctuation marks and numbers.
- Combine all the top 25 News headline into one single list of words per day.

**The Data has been Processed!**



# Words → Vectors

- **CountVectorizer** helps to tokenize and determine the frequency of the words.

- Then **fit_transform** is applied on the above object to obtain a sparse matrix of word counts.

# MODEL Logistic Regression:

```
(3975, 46002)
             precision    recall  f1-score   support

          0       0.83      0.80      0.82       186
          1       0.81      0.84      0.83       192

avg / total       0.82      0.82      0.82       378

0.822751322751
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 149 | 37 |
| 1 | 30 | 162 |

**1 gram model Accuracy 82.275%**

```
(3975, 584289)
             precision    recall  f1-score   support

          0       0.85      0.85      0.85       186
          1       0.86      0.86      0.86       192

avg / total       0.86      0.86      0.86       378

0.857142857143
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 159 | 27 |
| 1 | 27 | 165 |

**Bi-gram model Accuracy 85.714%**

```
(3975, 969254)
             precision    recall  f1-score   support

          0       0.92      0.76      0.84       186
          1       0.80      0.94      0.87       192

avg / total       0.86      0.85      0.85       378

0.851851851852
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 142 | 44 |
| 1 | 12 | 180 |

**Tri-gram model Accuracy 85.185%**

# MODEL : Random Forests

```
(3975, 46002)
             precision    recall  f1-score   support

          0       0.90      0.77      0.83       186
          1       0.81      0.92      0.86       192

avg / total       0.85      0.85      0.85       378

0.846560846561
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 144 | 42 |
| 1 | 16 | 176 |

**1 gram model**
**Accuracy 84.465%**

```
(3975, 584289)
             precision    recall  f1-score   support

          0       0.93      0.77      0.84       186
          1       0.81      0.95      0.87       192

avg / total       0.87      0.86      0.86       378

0.859788359788
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 143 | 43 |
| 1 | 10 | 182 |

**Bi-gram model**
**Accuracy 85.978%**

```
(3975, 969254)
             precision    recall  f1-score   support

          0       1.00      0.70      0.82       186
          1       0.77      1.00      0.87       192

avg / total       0.89      0.85      0.85       378

0.851851851852
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 130 | 56 |
| 1 | 0 | 192 |

**Tri-gram model**
**Accuracy 85.185%**

# MODEL: LINEAR SVM

```
(3975, 46002)
             precision    recall  f1-score   support

          0       0.83      0.81      0.82       186
          1       0.82      0.83      0.83       192

avg / total       0.82      0.82      0.82       378

0.822751322751
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 151 | 35 |
| 1 | 32 | 160 |

**1 gram model**
**Accuracy 82.275%**

```
(3975, 584289)
             precision    recall  f1-score   support

          0       0.83      0.86      0.85       186
          1       0.86      0.83      0.85       192

avg / total       0.85      0.85      0.85       378

0.846560846561
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 160 | 26 |
| 1 | 32 | 160 |

**Bi-gram model**
**Accuracy 84.656%**

```
(3975, 969254)
             precision    recall  f1-score   support

          0       0.90      0.78      0.83       186
          1       0.81      0.91      0.86       192

avg / total       0.85      0.85      0.85       378

0.846560846561
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 145 | 41 |
| 1 | 17 | 175 |

**Tri-gram model**
**Accuracy 84.656%**

# MODEL: SVM(GAUSSIAN KERNEL)

```
(3975, 46002)
             precision    recall  f1-score   support

          0       1.00      0.70      0.82       186
          1       0.77      1.00      0.87       192

avg / total       0.89      0.85      0.85       378

0.851851851852
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 120 | 66 |
| 1 | 0 | 192 |

**1 gram model**
**Accuracy 85.185%**

```
(3975, 584289)
             precision    recall  f1-score   support

          0       1.00      0.70      0.82       186
          1       0.77      1.00      0.87       192

avg / total       0.89      0.85      0.85       378

0.851851851852
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 130 | 56 |
| 1 | 0 | 192 |

**Bi-gram model**
**Accuracy 85.185%**

```
(3975, 969254)
             precision    recall  f1-score   support

          0       1.00      0.65      0.78       186
          1       0.74      1.00      0.85       192

avg / total       0.87      0.83      0.82       378

0.825396825397
```

| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 130 | 56 |
| 1 | 0 | 192 |

**Tri-gram model**
**Accuracy 82.539%**

# MODEL: NAÏVE BAYES

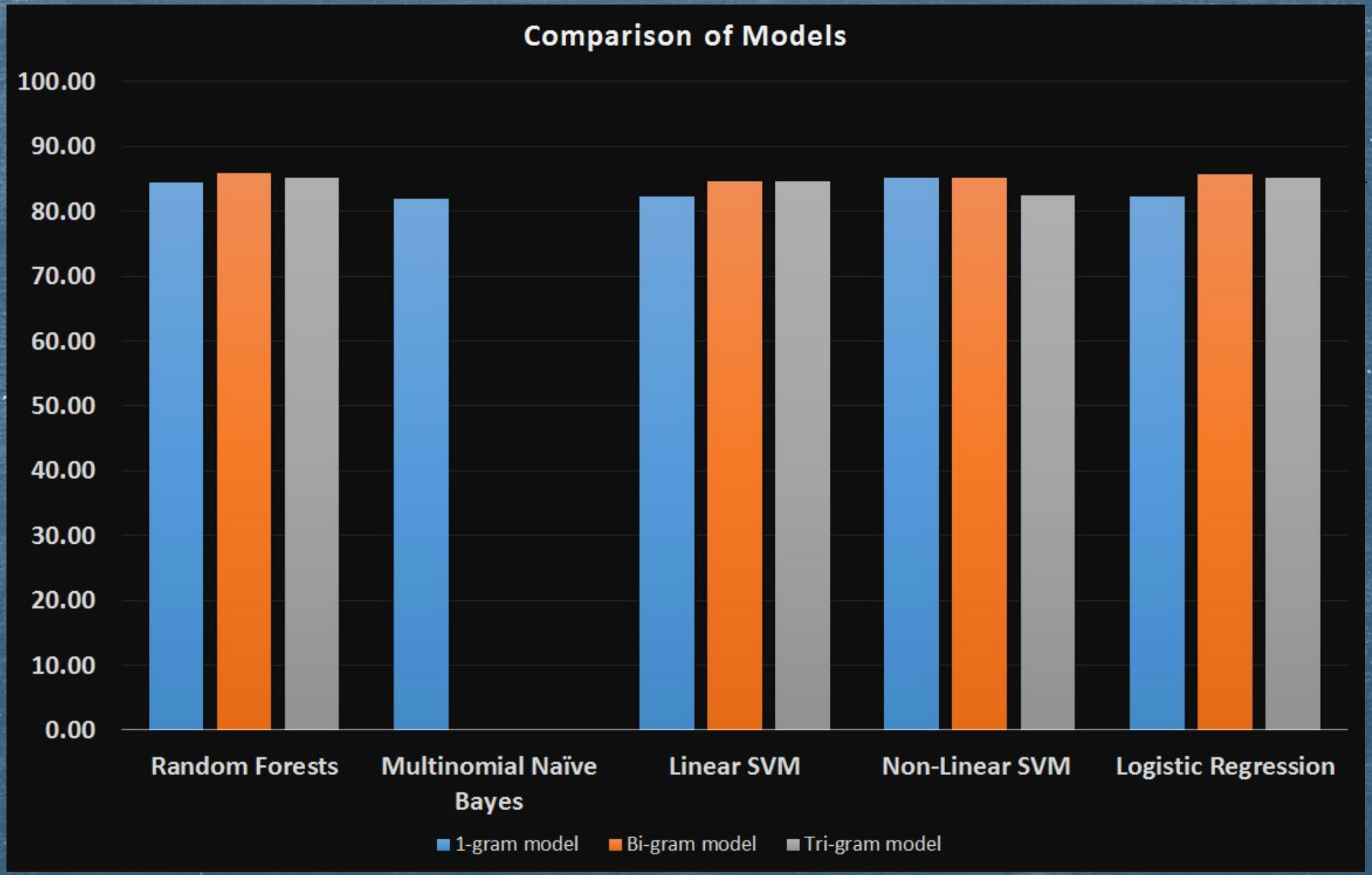
```
(3975, 46002)
             precision    recall  f1-score   support

          0       0.81      0.83      0.82       186
          1       0.83      0.81      0.82       192

avg / total       0.82      0.82      0.82       378

0.820105820106
```

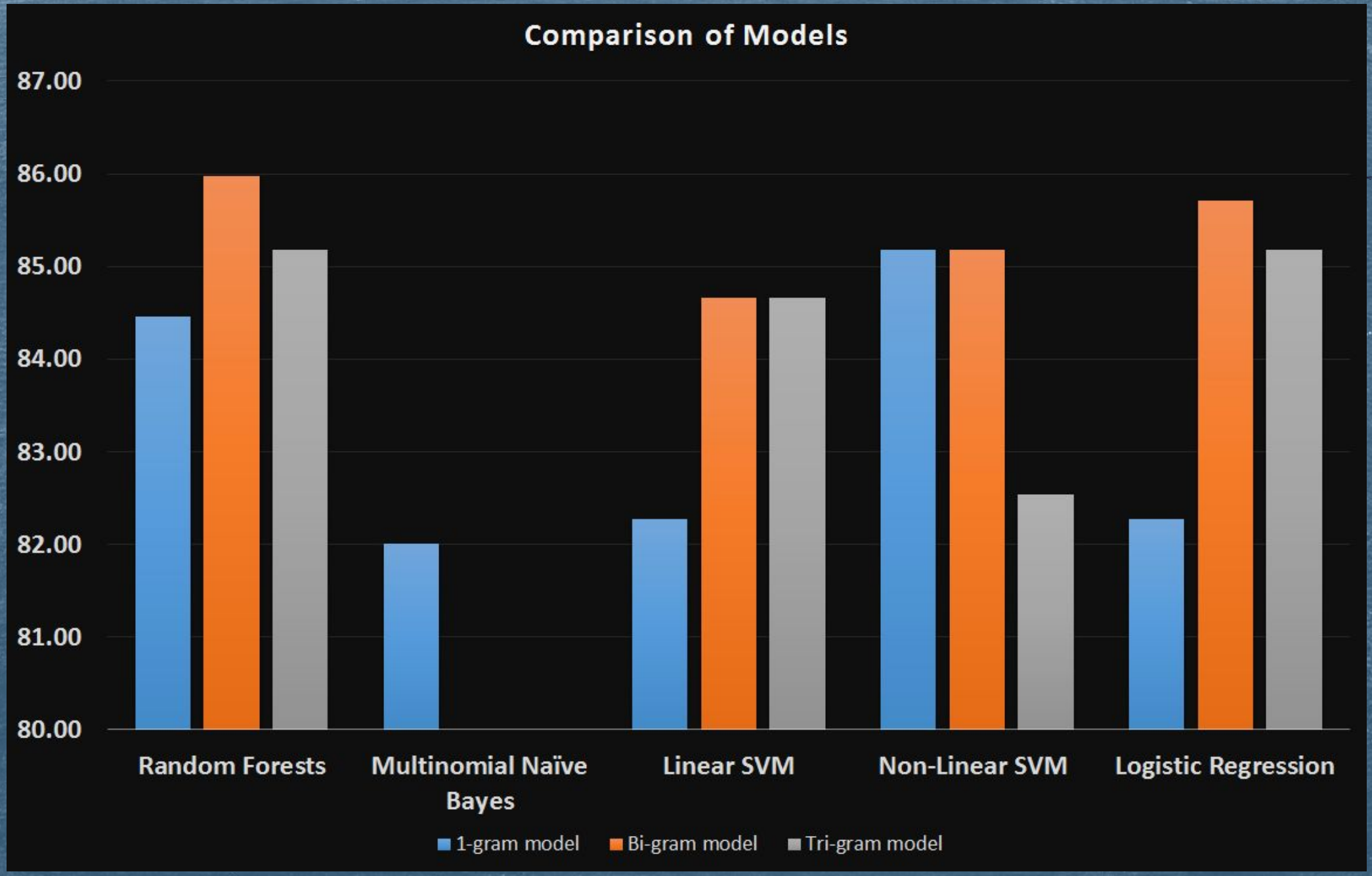
| Predicted | 0 | 1 |
|---|---|---|
| Actual | | |
| 0 | 155 | 31 |
| 1 | 37 | 155 |

1 gram model – Accuracy 82.0105%

Bi-gram model didn't want to execute on our computer

```
MemoryError:
```


Comparison of Models
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
Random Forests
Multinomial Naïve Bayes
Linear SVM
Non-Linear SVM
Logistic Regression
1-gram model
Bi-gram model
Tri-gram model


Comparison of Models
87.00
86.00
85.00
84.00
83.00
82.00
81.00
80.00
Random Forests
Multinomial Naïve Bayes
Linear SVM
Non-Linear SVM
Logistic Regression
1-gram model
Bi-gram model
Tri-gram model

# Conclusion

- Random forests had highest accuracy on the a bi-gram model as shown in the chart. The prediction accuracy was 85.97%.

- Using Natural Language Processing techniques, we were able to accurately predict the stock market trends 85% of the time.

# Questions?