# Stock Price Movement Using News Analytics

Ravi Prakash Singh, Sudhanshu Raj Singh, Manish Mehta, Aditya Aggarwal, Emily Huskins, Anna Riehle

## Summary


Stock price trend prediction is currently an active research area. It is a well-established fact that stock prices are driven by market sentiment. It is also reasonable to hypothesize that news has a role in shaping this sentiment.

Validating this hypothesis and making use of the news for stock price prediction can have tremendous benefits and utility in the finance world. While this field has attracted significant interest, a model with reasonable accuracy still evades the seekers. We aim to develop such a model, which would help investors to make better investment decisions.


## Our Data

Our data collection was primarily motivated by the challenge hosted by an investment management company, Two Sigma, on Kaggle. This challenge seeks to use news analytics to predict stock price performance. Our data spans from 2007-2016 and includes both Market Data provided by Intrinio and News Data from Thomson Reuters. The data contains information on around 3,500 stocks and includes features such as stock price, volume, and news sentiment.

| | Num. Rows | Num. Columns | GB |
|---|---|---|---|
| **Market Data** | 4,072,955 | 16 | 1.1152 |
| **News Data** | 9,328,749 | 35 | 4.7263 |

## Data Preprocessing

To avoid impact from the **Financial Crisis of 2008**, the data we used was limited to 2010-2016. Rows with blank headlines and rows with a delay of one day or more between when a news item was created and when it became publicly available were removed, since they provided little useful information. Rows with an urgency of 2 were also ignored as they made up less than 0.001% of the data.

| | Original Rows | Final Rows |
|---|---|---|
| **Market Data** | 4,072,955 | 2,946,738 |
| **News Data** | 9,328,749 | 6,987,537 |

## Feature Engineering

Most of the news features were engineered to follow the format:

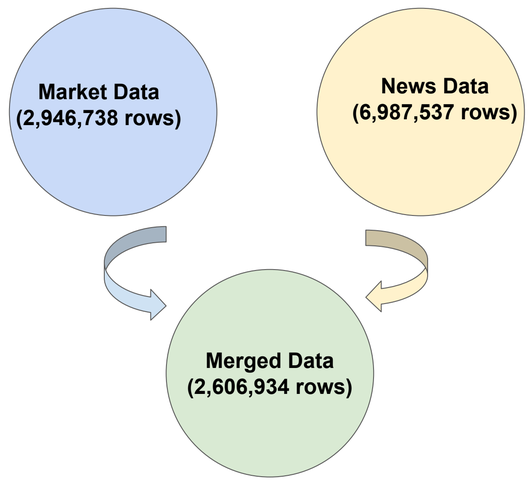
COLUMN NAME → *WordCount_3_1* ← URGENCY, SENTIMENT CLASS

**Urgency:** Differentiates between an alert (1) or an article (3)
**Sentiment Class:** Classification of overall news sentiment

- -1: Predominantly Negative Sentiment
- 0: Predominantly Neutral Sentiment
- 1: Predominantly Positive Sentiment

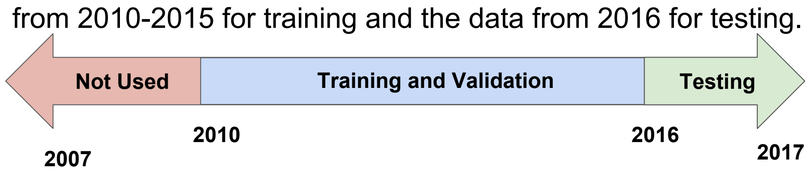
## Merging Data

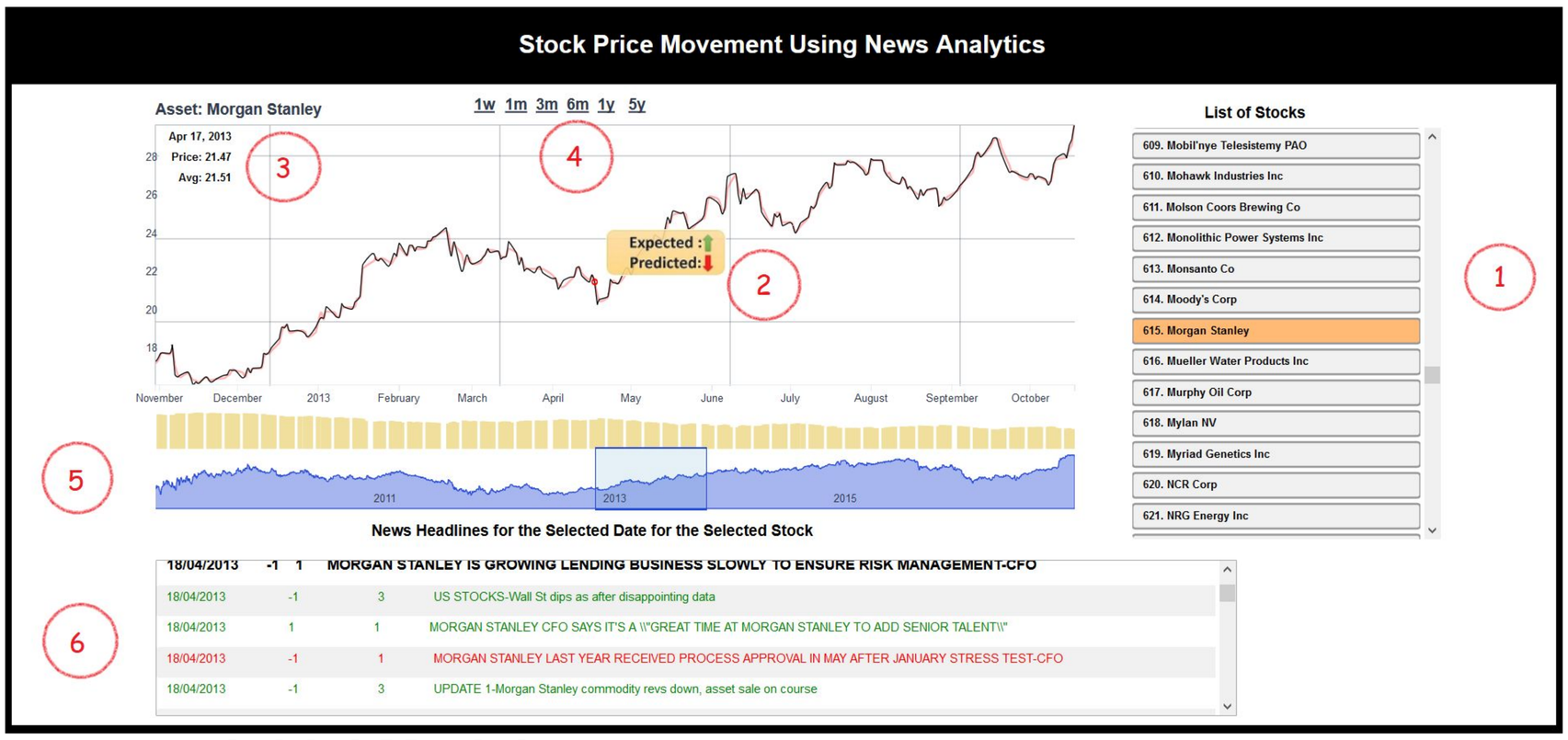
For each stock, we aggregated news data at the daily level and merged it with market data on date and stock code.




## Splitting Data

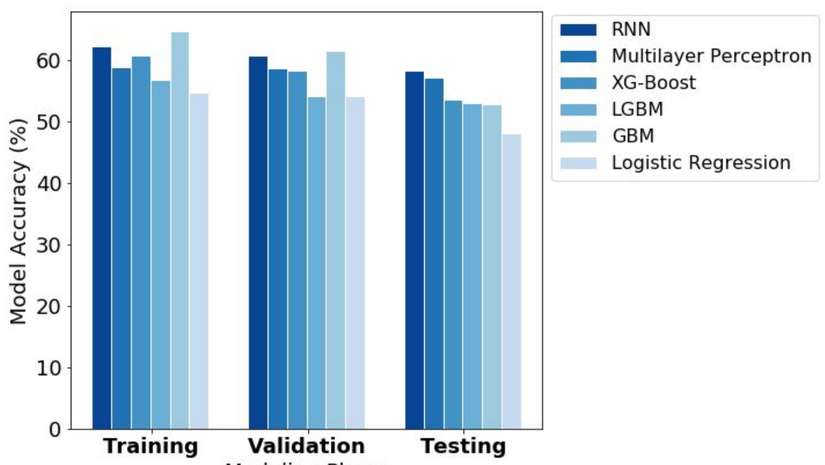
Our original data spanned from 2007-2016. We used the data from 2010-2015 for training and the data from 2016 for testing.




## Initial Models

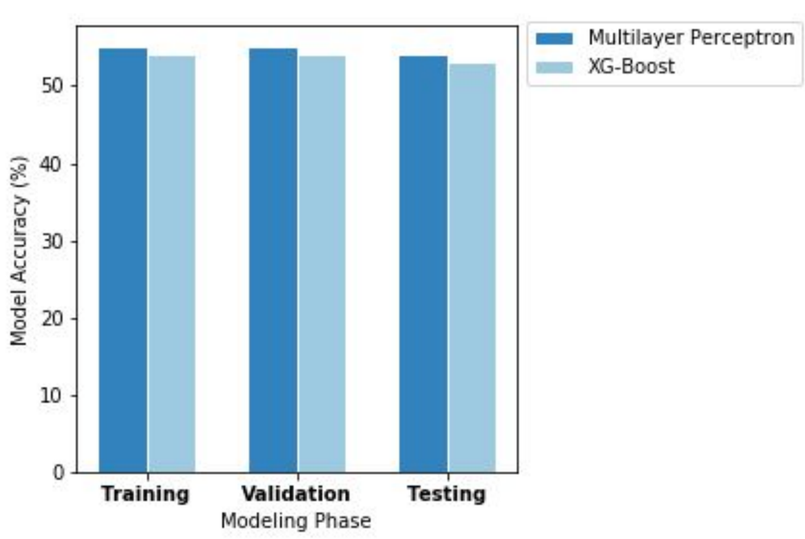
- **Recurrent Neural Network** (RNN)
- **Multilayer Perceptron**
- **Extreme Gradient Boosting Machine** (XG-Boost)
- **Light Gradient Boosting Machine** (LGBM)
- **Gradient Boosting Machine** (GBM)
- **Logistic Regression**




Primary models investigated were Neural Networks and Gradient Boosting Machines. These had higher accuracy than more traditional models like Logistic Regression. Neural Nets imitate workings of the human brain. GBMs are regularized decision trees providing increased control over model development.

| Deep Learning Models | Gradient Boosting Machines |
|---|---|
| **Recurrent Neural Network**<br>• Highest Testing Accuracy<br>• Second-highest Training and Validation Accuracy<br>• Typically used for Text/Speech Recognition | **XG-Boost**<br>• Accuracy similar to Neural Nets<br>• Provides greater control in building of trees<br>• Helpful for Regularized classification |
| **Multilayer Perceptron**<br>• More general than RNN with similar accuracies<br>• Helpful for Classification Prediction<br>• Recommended for Time Series data | **Light-Gradient Boosting Machine**<br>• Accuracy similar to XG-Boost<br>• Provides high speed for working with large data<br>• Negative stock price movement incorrectly predicted |

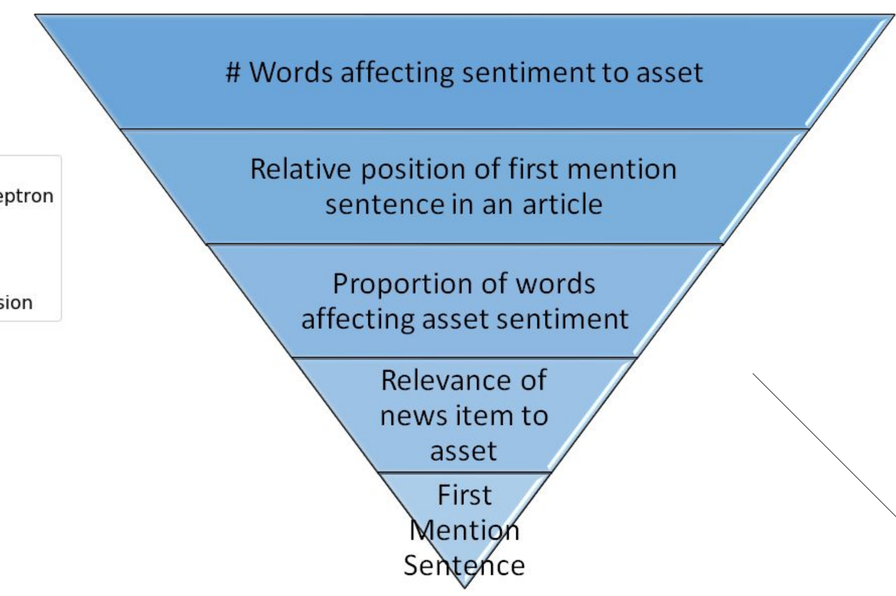
Given the time series nature of our data, MLP was used for further exploration among deep learning models. XG-Boost was further explored among Gradient Boosting Machines because it provided better classification of stock price movement.




After tuning, MLP and XG-Boost provided similar accuracy scores. XG-Boost was chosen due to its faster fit time and higher explainability, interpretability, and usability.

## Final Model: XG-Boost

The following are the most important news features for predicting stock price movement, which are in line with our intuition:

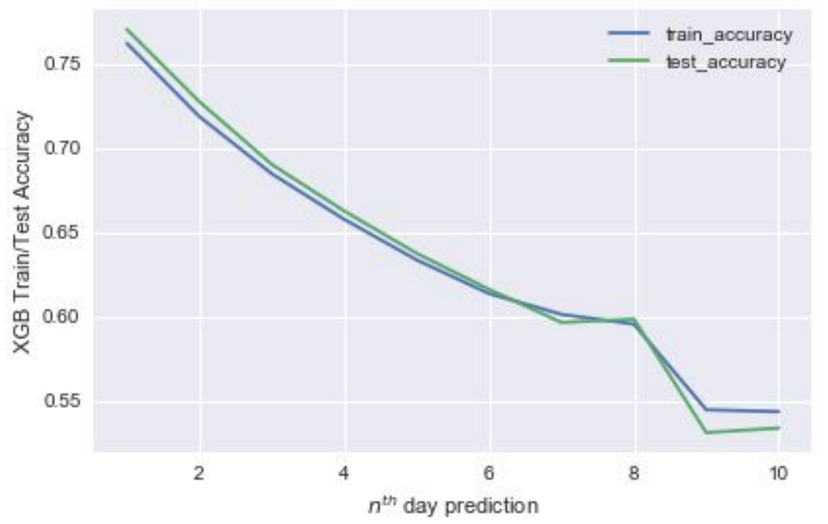



## Model Accuracy

Model had roughly equal percentages of predicted values for true positive, true negative, false positive, and false negative. True positive rate and true negative rate are provided below.

| TP Rate | TN Rate |
|---|---|
| 55.45% | 53.22% |

The models built predicted the stock price movement after 10 days. For analyzing the time-based effect of news, 10 different models predicting stock price movement each day from the day of observation until day 10 were developed. The following graph depicts the decay in accuracy over time:




For next-day prediction, accuracy is 75%. Accuracy decreases as we try to predict for longer time periods. This is intuitive because the most recent news will predict stock price movement most accurately.

## Conclusions

Currently, most research focuses on predicting stock price movement based solely on historical stock market data. Earlier research using this method has produced accuracies of around 78% for next-day prediction. Our model, using only news data, gives an accuracy of 75% for next-day prediction, verifying that news plays an important role in stock price movement. Additionally, since news data has an effect for longer durations than random fluctuations in market stock price data, this model could be useful to investors to make relatively better long-term investment decisions.