CSCI 5622 Graduate Machine Learning Final Project

Using Media Data to Predict Stock Fluctuations

Mary Letey Morgan Allen Colton Williams Aniq Shahid

Main Focus

Research project investigating relationships between **media** concerning a company and that company's **stock price**.

Theoretically, the stock price of a company is based in part on public perception of their market. This perception can easily be affected by media sources. The project team hypothesizes that **social media sources and news sources may be used to predict daily stock price fluctuations** within the technology industry**.**

This team was able to construct a model that predicts the closing stock price with around 3% error.

Motivation

Considering the most basic supply-demand model in economics, stock prices are functions of the supply and demand for ownership in a company. Because the "consumers" in this supply and demand model are viewing a stock as an investment, demand represents confidence in the performance of the company. Furthermore, **an increasing stock price represents increased confidence in the potential of the company**. Thus the stock price may be affected by a change in the public's understanding of a **company's *perceived* growth and risk**. Forms of media, such as news articles, may have a huge impact in these perceptions.

This project will examine the extent to which media sources affect stock prices, and how well we can use media data to predict stock price changes.

Data and Methods

Numerical Data from Bloomberg recording daily stock price
Google Trends measuring how often companies are searched
Text Data from news sources scraped off the web

Companies representing the technology industry
- HP
- IBM
- Intel
- Seagate
- Western Digital

Regression to establish a relationship between stock price and Google trends
Logistic Regression to model impact of words on an up-down change in the stock price
Topic Modeling to determine different themes occurring within articles
Recurrent Neural Network (RNN) as our main learning model which takes numerical, textual and google trends data and outputs final price prediction

Results

Figure 1 and Table 1 show how **Google Trends** were a very **effective predictor** of **stock price** in a regression model

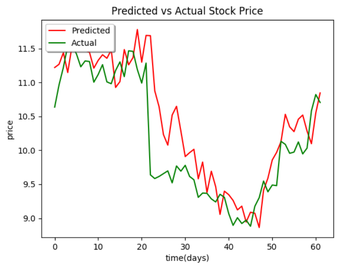

Figure 1. Regression Results
Regression results using Google trends data for company HP

Company	MSE (%)
HP	0.9
IBM	4.0
Intel	0.7
Seagate	1.0
West Dig	0.6

Table 1. Mean Squared Error
Error from the regression model by company

Figure 2 and Table 2 show how **article text data** was an effective predictor of **stock price** in a logistic regression model

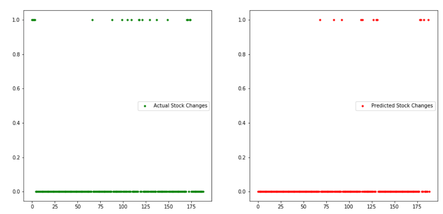

Figure 2. Logistic Regression Results
Stock increase/decrease predictions based on article data for Seagate.
0 represents decrease from previous day; 1 increase

Company	Accuracy (%)
HP	85
IBM	76
Intel*	---
Seagate	84
West Dig*	---

Table 2. Testing Accuracy
Error from the logistic regression text model by company

Topic modeling is commonly used in Natural Language Processing (NLP) to reduce a large number of sparse features into dense and meaningful topic

Topic	Words			
Topic 1	Devices	Share Buyback	Strategic	Win
Topic 2	Price	Provided	Buy right	Revenues
Topic 3	Company	Stock	Market	NYSE
Topic 4	Packard	Buying	Company	IBM
Topic 5	Revenue	Year	New	Development
Topic 6	IBM	Intel	Data	Year

Table 3. Topic Modeling Classes
Articles divided into topics, trying to find some predictive pattern buried in the data.

Table 3 gives insight into the main "breakdown" of our text data. These topic classes became features in our main RNN model.

Results

The RNN model, a NN model that adapts to and "remembers" previous data, used current features, such as raw text data, 25 classes generated by topic modeling, and Google trends to ultimately predict the closing stock price.

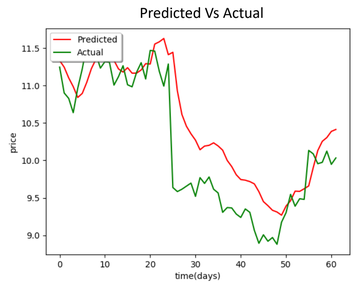

Figure 3a. RNN with Topic Modeling Features
Complete RNN model predictions for HP

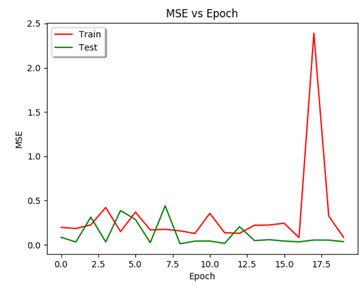

Figure 3b. MSE vs Epochs
Relationship between MSE and number of epochs in RNN model for HP

Figures 3a and 3b show the predictions and model accuracy for the total RNN model. The testing accuracy is quite high, on average 3% error.

It's worthwhile noticing from Figure 3a that the RNN model almost always overestimated the stock price. In addition, the predicted values also had significantly less variance than the actual stock values. This can lead the model to be better predicting long-term trends as opposed to short term valuations.

Conclusions and Discussion

The team was able to flesh out a solid model, with low error rates. This model can easily be built upon to include more data sources, such as twitter and annual legal data.

One of the biggest challenges was getting enough clean data from a variety of sources. Despite our efforts, there were often significant gaps in dates between published articles.

RNN's are potentially a great tool for stock predictions, but they require a large amount of data to achieve maximum potential. In a low data situation, a generative model is one of the best choices. However, as more data is collected and the gaps are filled, the RNN will continue to improve and eventually outperform a generative model.

Resources

Seeking Alpha α Bloomberg WSJ TensorFlow python™ jupyter

Graphs generated from TensorFlow and Juptyer Notebook models created by this team

* Logistic regression model data not currently available