Financial Data Infrastructure

Feb-April 2023

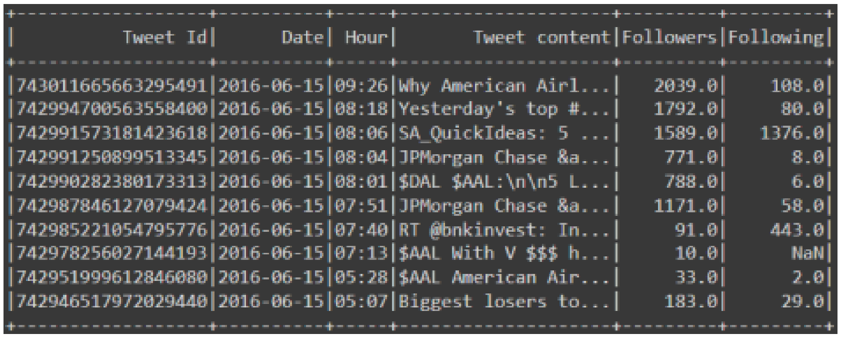
Predicting the American Airlines stock price through tweet text mining using Spark

Introduction

Throughout this course, text mining techniques have been studied with the purpose of extracting valuable information that can give its user useful insights to aid their decision-making process. Text mining in finance is important because there is a significant amount of soft information embedded in texts that is usually overlooked (as finance professionals tend to look more at numbers) but is incredibly valuable. Our team was motivated to use text mining in practice by using this technique with tweets that mentioned American Airlines and aiming to predict their stock prices using the information this technique would yield. Therefore, the team created a series of models that would attempt to accurately predict the stock prices, first by using tokenized words information and then by using sentiment analysis. We made use of two different machine learning models, linear regressions, and random forests. To obtain more accurate results, we took an unorthodox path where we did not split the data into train and test sets. Instead, we used all the observations to train the models. Ultimately this yielded better results that were neither overfitting nor underfitting, but it is important to keep in mind that this is not one of the best practices when one makes these types of machine learning models. In the end, the model that yielded the best results (a root mean squared error of 1.0953) was a random forest model using sentiment analysis and the whole dataset to train the model.

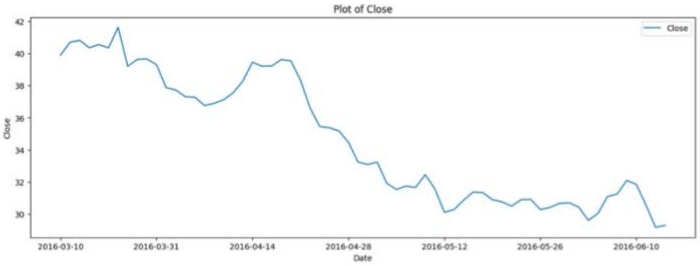
Data

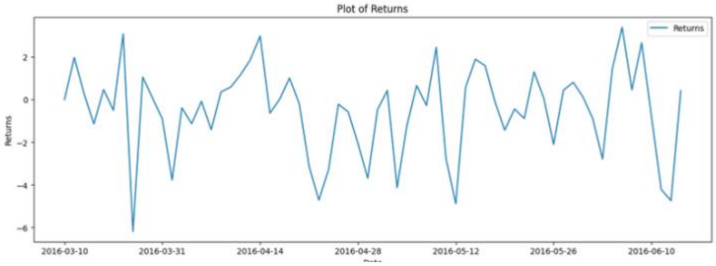
We obtained closing prices for American Airlines stock and obtained a database of tweets that mentioned American Airlines. In the tweets database we had their id, date and hour of publication, the text that was contained in the tweets, how many people followed the person who published it, as well as how many people the person who published it was following. The first observations in the dataset are shown in the table below.



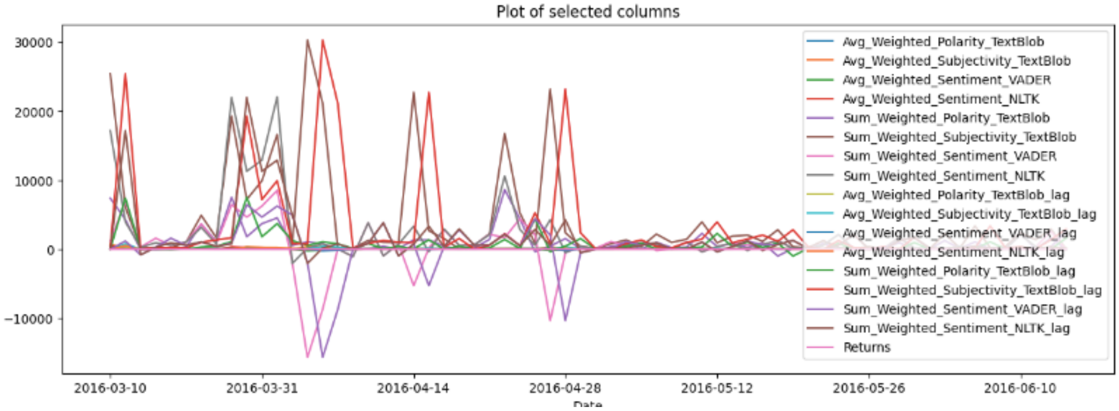
```
+------------------+----------+-----+--------------------+---------+---------+
|          Tweet Id|      Date| Hour|       Tweet content|Followers|Following|
+------------------+----------+-----+--------------------+---------+---------+
|743011665663295491|2016-06-15|09:26|Why American Airl...|   2039.0|    108.0|
|742994700563558400|2016-06-15|08:18|Yesterday's top #...|   1792.0|     80.0|
|742991573181423618|2016-06-15|08:06|SA_QuickIdeas: 5 ...|   1589.0|   1376.0|
|742991250899513345|2016-06-15|08:04|JPMorgan Chase &a...|    771.0|      8.0|
|742990282380173313|2016-06-15|08:01|$DAL $AAL:\n\n5 L...|    788.0|      6.0|
|742987846127079424|2016-06-15|07:51|JPMorgan Chase &a...|   1171.0|     58.0|
|742985221054795776|2016-06-15|07:40|RT @bnkinvest: In...|     91.0|    443.0|
|742978256027144193|2016-06-15|07:13|$AAL With V $$$ h...|     10.0|      NaN|
|742951999612846080|2016-06-15|05:28|$AAL American Air...|     33.0|      2.0|
|742946517972029440|2016-06-15|05:07|Biggest losers to...|    183.0|     29.0|
+------------------+----------+-----+--------------------+---------+---------+
```

To make the time series data of stock price stationary, we calculated the returns and worked with this variable throughout the rest of the project. This is illustrated in the graphs below, the first one is the stock price (where we can see a clear downward trend) and the second one is the returns.





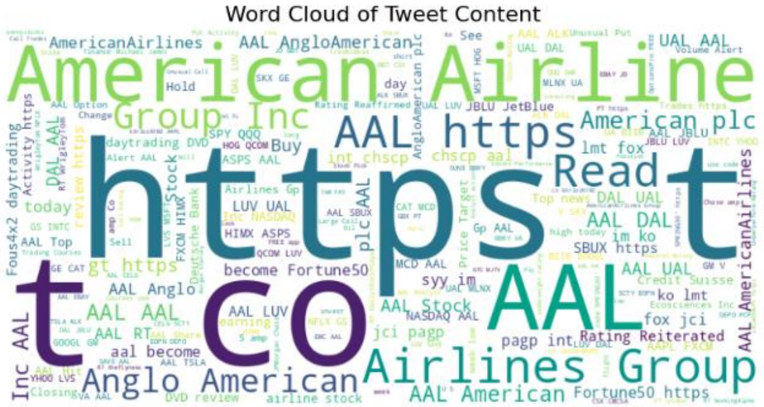
With text mining techniques first we excluded stop words (words that are used very frequently that do not yield useful information such as “the”, “and, “with”, etc.) and proceeded to create a dictionary with the words that could serve of indication of how the company was performing. We tokenized these words with an id that would be used in the model based on word tokenization. Furthermore, a dictionary of sentiment analysis was created. Sentiment analysis focuses on whether the tweet content is positive or negative. It would be expected that circulating negative tweets would push stock price and returns downwards while positive tweets would push them upwards. The project relies mostly on sentiment analysis to make the predictive models since, compared with word tokenization by itself, sentiment analysis was proving to give more accurate results. Below are the created variables of sentiment analysis that were used to create the models.

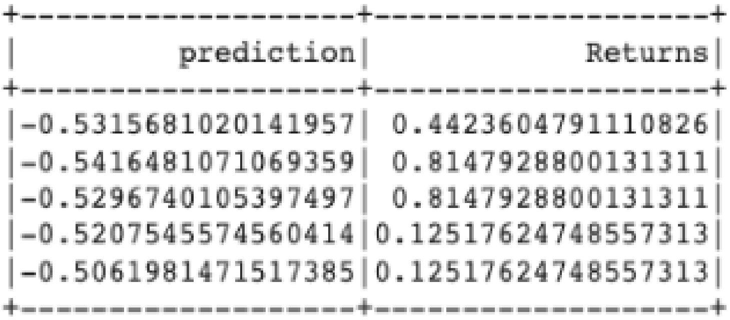


The Models in Spark

Model 1: Linear Regression with Text Mined Tokenized Words

The first model we explored was a linear regression using the tokenized words from the tweets dataset. However, by creating a word cloud the algorithm was taking into account (shown below), we noticed it considered some words that did not have a very valuable meaning for our purposes such as “https”, “group” or “read”. Some of the predictions this model made are also shown below.



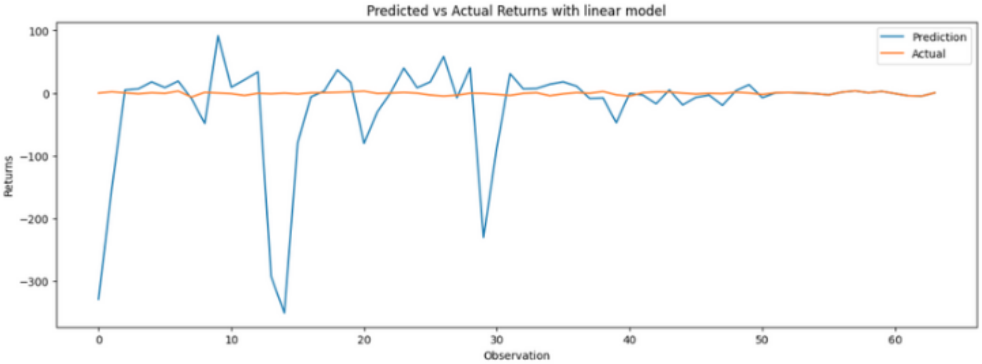


```
+-------------------+-------------------+
|         prediction|            Returns|
+-------------------+-------------------+
|-0.5315681020141957| 0.4423604791110826|
|-0.5416481071069359| 0.8147928800131311|
|-0.5296740105397497| 0.8147928800131311|
|-0.5207545574560414|0.12517624748557313|
|-0.5061981471517385|0.12517624748557313|
+-------------------+-------------------+
```

This model had a root mean squared error of 2.77134. We believed we could obtain better results and decided to switch the approach using the sentiment analysis.

Model 2: Linear Regression with Sentiment Analysis (Train and Test Split)

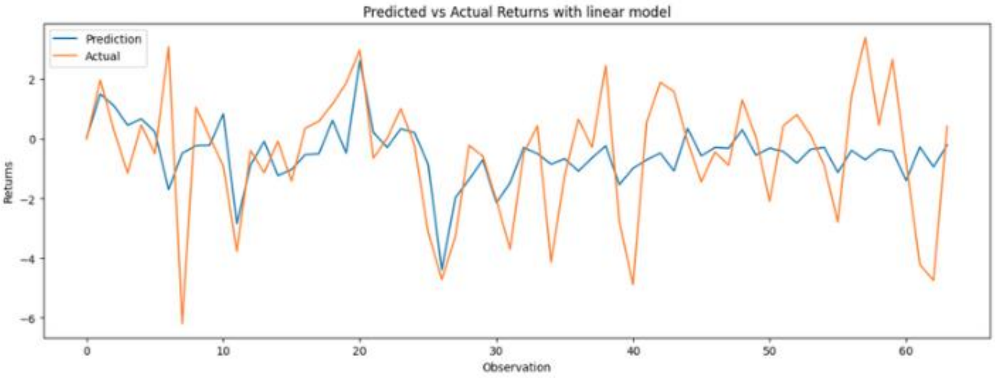
For this linear regression model, we used the sentiment analysis variables shown in the Data section following the train and test split procedure. A graph of the predictions given by this model against the actual data is provided below. The blue line are the predictions and the orange is the actual values (same for the other models).



From the graph, it is possible to see that this model severely overreacts to the data given. It does not provide accurate stock returns predictions and has a very elevated RMSE of 83.1507. This is when we were motivated to try to train the model with the whole dataset.

Model 3: Linear Regression with Sentiment Analysis (Whole Dataset to Train)

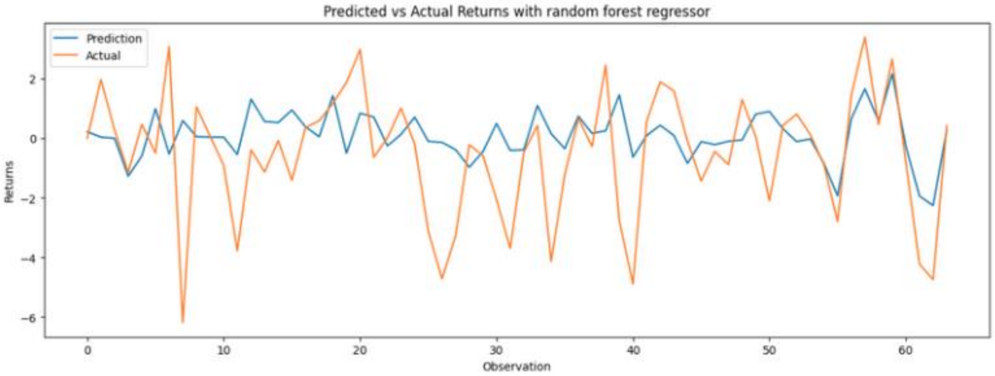
This model is the same as the previous one, except that it was trained with the whole data available. This is not following common procedure of this machine learning model, but, as it is possible to see in the next graph, it provided better results.



This third model has a very good fit to the data. With the first observations, one might say the model is somewhat overfitting and for the most recent observations it is underfitting. However, so far this model had the lowest RMSE (2.64012), meaning it was more accurate than the others. We were now encouraged to try a different machine learning model. We opted to try out a random forest model because we thought it could work well with the kind of data we possessed.

Model 4: Random Forest Model with Sentiment Analysis (Train and Test Split)

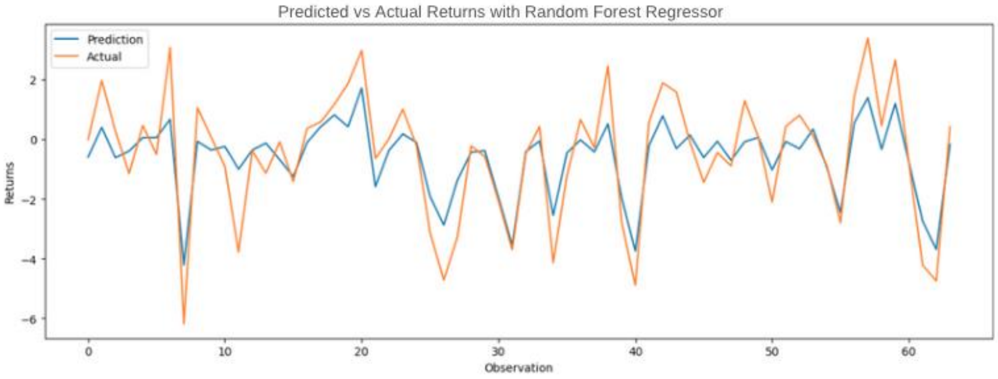
We started respecting the train and test split for this random forest model. Below is the graph of the predictions of this model against the actual data.



This model does not seem to have an overfitting issue. One could argue it seems to be underfitting. Nevertheless, this random forest model had a lower RMSE than the linear regression. Its RMSE was 1.93085. The final model we would try would be this same random forest with sentiment analysis but taking in all the dataset to train it (similarly to Model 3).

Model 5: Random Forest Model with Sentiment Analysis (Whole Dataset to Train)

In this last model, once again, we went against procedure seeking to obtain more accurate predictions and took the whole dataset as training material for our model.



This model yielded the most accurate predictions. It had the lowest root mean squared error of 1.09533. We can see from the graph that it is neither overfitting nor underfitting which is an indication that this model could work well with new data given to it. This is the model we would recommend that investors use to predict the stock returns of American Airlines based on soft information from tweets that mention this company.

Result

Summary of the results

A comparison of RMSEs across the five different models is offered below:

1. Linear Regression (Tokenized Words) …………………………………… 2.77134
2. Linear Regression (Sentiment / Train and Test) ………………………… 83.1507
3. Linear Regression (Sentiment / Whole Dataset) ………………………… 2.64012
4. Random Forest (Sentiment / Train and Test) …………………………….. 1.93085
5. **Random Forest (Sentiment / Whole Dataset) ………………………… 1.09533**

Conclusion

Concluding points, limitations of the project, and further works

The model with the most accurate predictions and with no overfitting nor underfitting was Model 5, the random forest model with sentiment analysis and taking the whole dataset to train it. However, here comes the first limitation of the project. Our best model did not follow the established procedure of train and test split. We would therefore advise users of this model to be cautious when considering it for their decision-making process. Another limitation of our project is that companies use bots or pay people with influence in social media to shift the public perception to their favour. If American Airlines did this, this could be creating some noise in the dataset and making our model not as accurate as it could be. Furthermore, the airline industry is extremely sensitive to shocks in the economy. If our dataset includes periods of economic inflection, this could also be creating noise for our model. For further research, it would be useful to aggregate other variables to obtain more accurate predictions such as financial ratios, analysts' opinions, macroeconomic indicators, and industry-specific data. Future researchers are also encouraged to try other kinds of models to explore is they provide more accurate results.