IDS 594 – COGNITIVE COMPUTING ANALYTICS

CODE DOCUMENTATION

Fall 2019

Ajay Srivats

Anchit Jhingan

Kavya Ravi

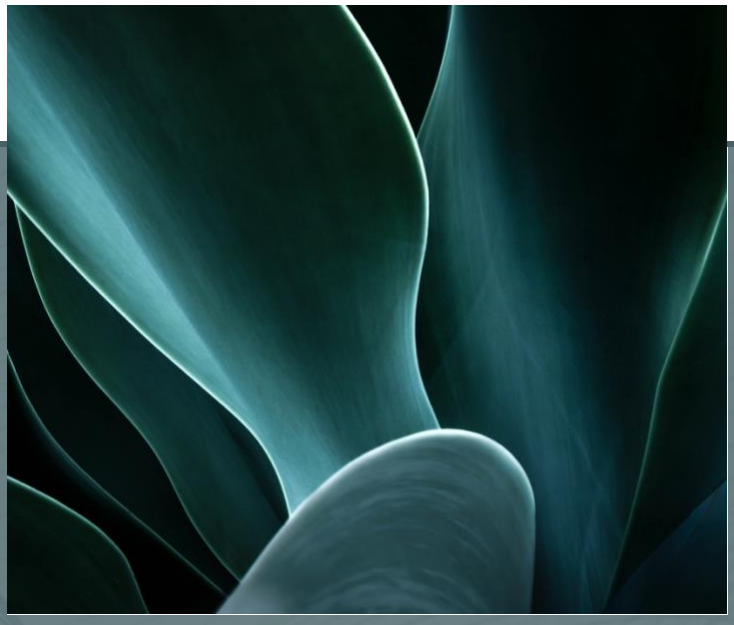
Pavlo Diatkovych

Shubham Puri

ABOUT WHAT WE'VE DONE

Our business gathers sentiment data and incorporates it with historical price data to accurately detect and summarize the effects that sentiment has on stock price movements.

The objective of this project was to deploy different machine learning algorithms to predict sentiment based on news headlines and find correlation with stock prices.



THE PROCESS

WALK THROUGH

During the course of this project, we've been successfully able to successfully scrape posts/tweets from popular social media platforms, scrape news headlines from credible business sources like businesstimes, finviz, stocktwits etc. We have also collected yahoo finance data for several different stocks.

This raw data had to undergo extensive feature engineering before it could be used for predictive modeling & sentiment analysis.

The next several pages dive deep to describe the strategies implemented and the results we were able to derive.

Our analysis is two part, performed both on historical data as well as sentiment data.

STEP 1 – IMPORTING THE DATA

Sentiment Data:

Included in our archive package are html files which need to be placed in a folder inside the Jupyter environment ('dataset').

We then proceed to extract the headlines data from these tables into one dataframe.

Historical Data:

We look at stock prices over the past years, we set our start date at January 1, 2015.

We then use the web.DataReader function to scrape data for one stock at a time. The first argument is the series we want, the second is the source ("yahoo" for Yahoo Finance), third is the start date and fourth is the end date.

STEP 2 – DATA PRE-PROCESSING

Historical Data:

Our pre-processing steps for historical data included:

1. Set the index as date
2. Create a dataframe with date and the target variable
3. Split into train and validation (did not use random splitting since that will destroy the time component of our data)
4. We set last year's data as our validation and the 4 years prior to that as our training set.

Sentiment Data:

Our pre-processing steps for sentiment data are as follows:

1. Iterate through all the 'tr' tags in news_table
2. Read the text from tr tag into text
3. Split the text into a list
4. Check the length of 'date_scrape',
   If = 1 then we load time as our only component
   Else, we load date as our first element and time as second
5. Extract the ticker from the file name, get the string up to the 1st '_'
6. Append ticker, date, time and headline as a list to the 'parsed_news' list

STEP 3 – OUR ANALYSIS

Historical Data:

We train our models on 4 years of historical data and create predictions for our validation set. These predictions are checked using actual values and the results are plotted within python.

The following models were deployed during our analysis:

1. Linear Regression
2. K-Nearest Neighbors
3. Time Series & Seasonality analysis using fb Prophet
4. LSTM

Sentiment Data:

We used SentimentIntensityAnalyzer within the NLTK Vader package for sentiment analysis. We Instantiate the sentiment intensity analyzer with the existing lexicon. Before we could proceed, we wanted to update our sentiment lexicon with a few words and values particular to our use-case.

We then proceed to apply the NLTK using the Vader model, which is majorly used in financial industries.

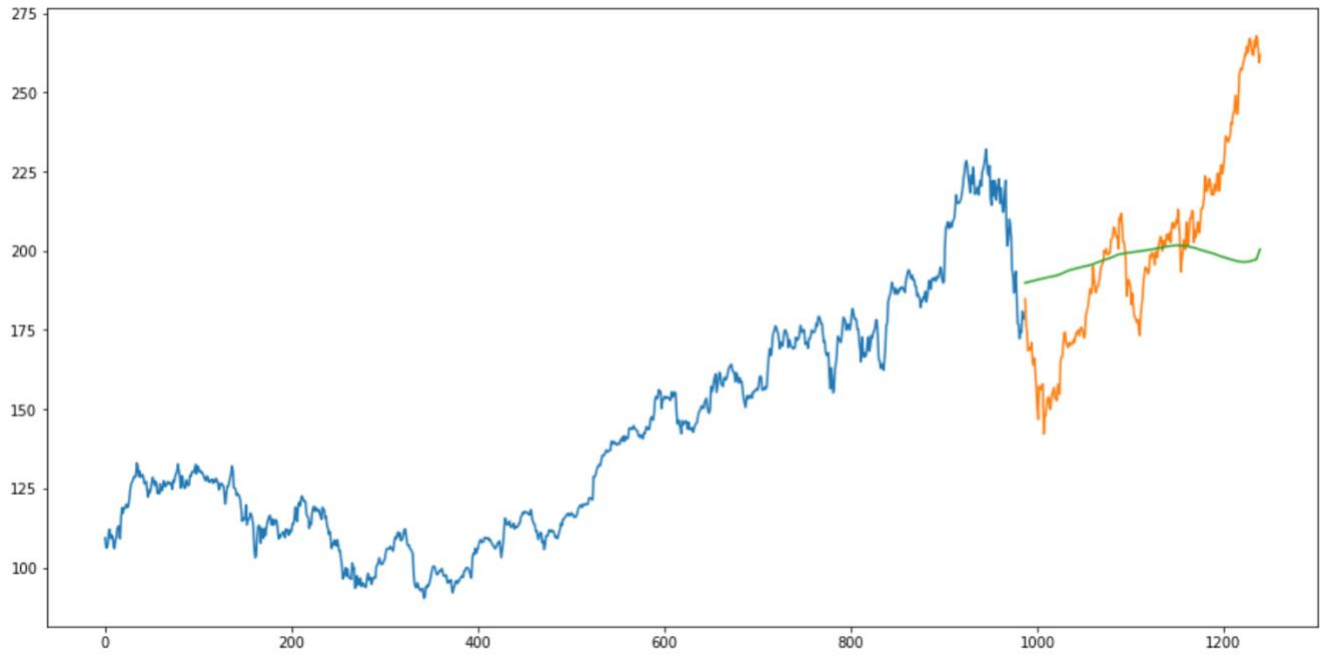
1. Set column names
2. Convert our list of lists (headlines) into a single dataframe
3. Iterate through these headlines and generate Vader polarity scores
4. Convert the resulting list of dicts into a new dataframe
5. Join the dataframe and convert the date column from string to datetime.

STEP 4 – VISUALIZING OUR RESULTS

We use the matplotlib library inside python to plot our results for both sets of data.

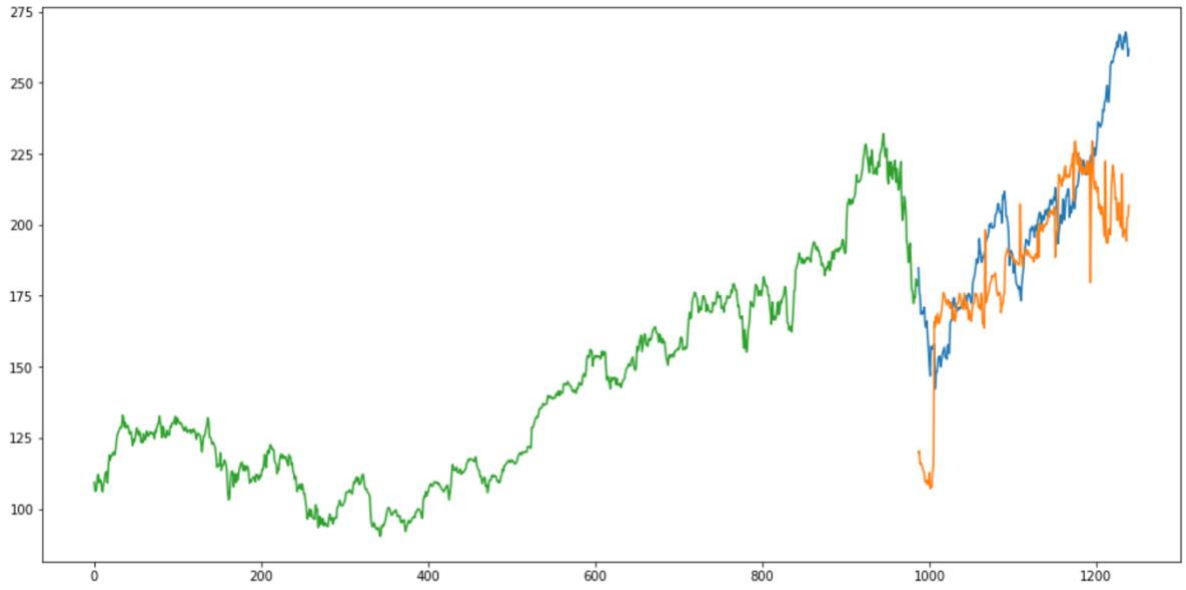
Some of our most prominent plots are shown below, rest of which can be found within the notebook file.

Linear Regression (Historical Data)

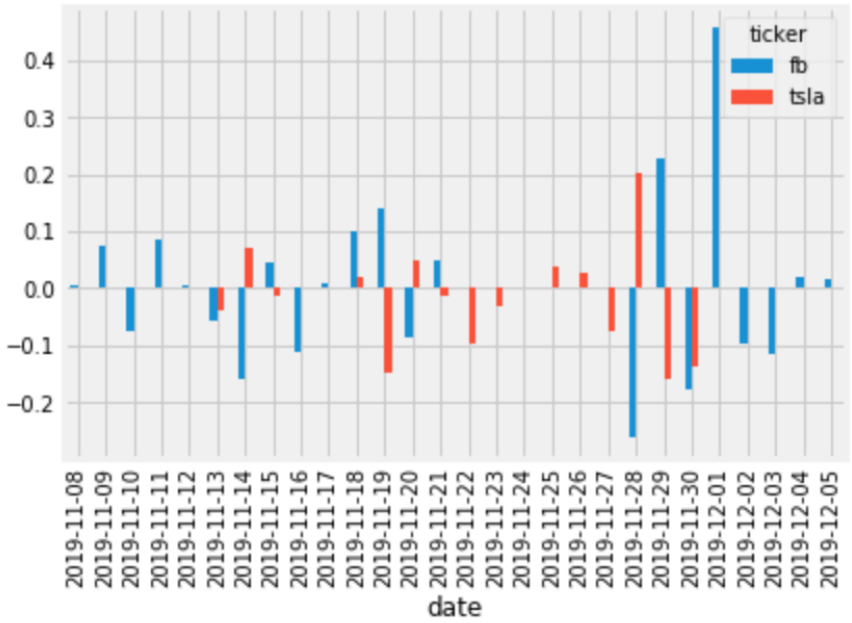


STEP 4 – VISUALIZING OUR RESULTS

K – Nearest Neighbors (Historical Data)

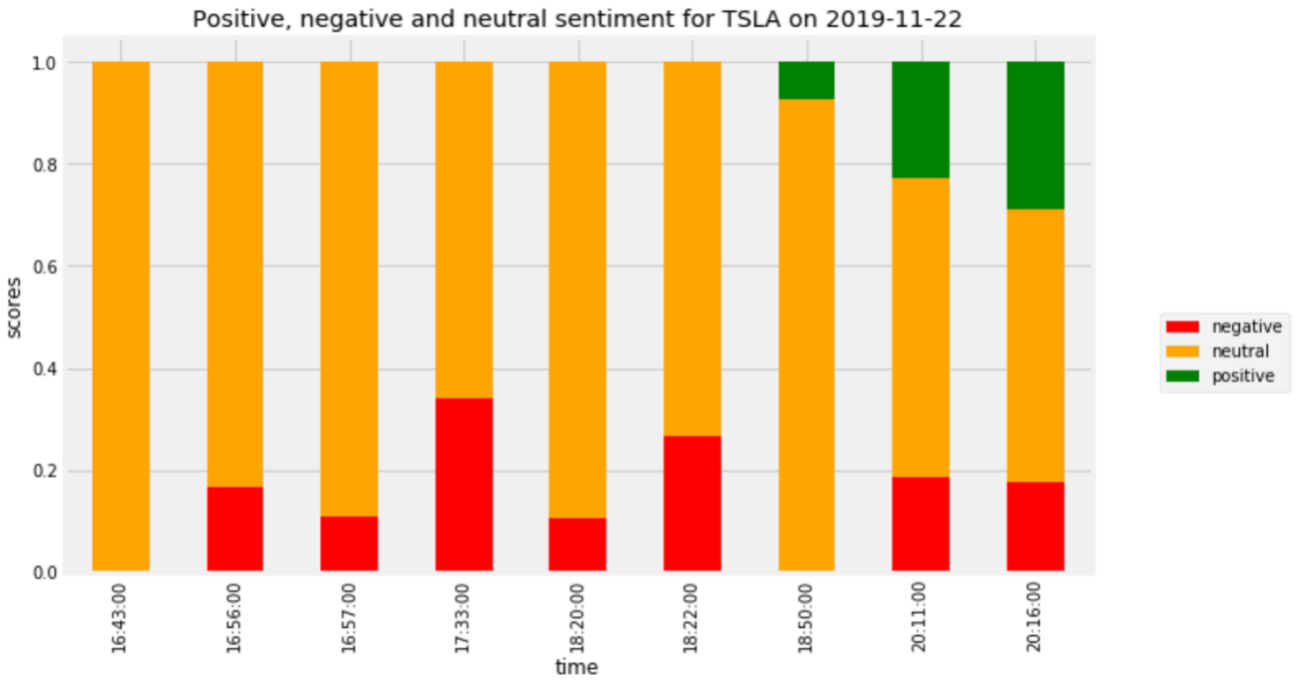


Plotting the mean of scored_news



STEP 4 – VISUALIZING OUR RESULTS

Sentiment scores for one stock & one day



SET UP

EXECUTING OUR NOTEBOOK

Download both notebooks attached and open using Jupyter. To run both these files successfully, a few steps need to be followed.

1. Both our included notebooks should come with all necessary imports
2. Make sure the path to the ('dataset') folder is set correctly and the files inside the dataset folder are copied correctly
3. Run all the cells in order and do not uncomment any line which has been placed inside comments.
4. Now run all the remaining cells. Everything should run without any errors.