**Capstone Project proposal**

**1. Problem**

Algorithmic trading has revolutionized the stock market and its surrounding industry. Over 70% of all trades happening in the US right now are being handled by bots. How to accurately predict an equity’s future price is a vitally important factor which determines an investor’s profits.

Generally, investor use two methods to analyze stock market: Fundamental analysis and Technical analysis. Fundamental analysis involves analyzing the company’s future profitability on the basis of its current business environment and financial performance, while technical analysis includes reading the charts and using statistical figures to identify the trends in the market. What’s more, people’s emotional opinion also plays a significant role on equity’s price variation.

For technical analysis, research use common indicator or variables, such as the historical price series, namely time series data, to build a model and try to predict the price fluctuation. On the other hand, someone prefers to harness online information, take for example social media or news outlets’ stories, to observe the stock’s volatility. They leverage NLP methods to process stories information and acquire sentimental evaluation of those stories.

In this project, I want to involve these two methods, combing Time Series model and sentimental analysis, into one solitary predicting model. According to the analysis of price history and mainstream news sentiment, I hope to generate a result to show whether the equity’s price will increase or decrease.

1. **Data**

1. In order to setup time series model, I will use a dataset from Quandl. Additionally, I will use Yahoo Financials to acquire real time price quotation.

<https://www.quandl.com/data/EOD-End-of-Day-US-Stock-Prices/usage/quickstart/api>

2. The task of mining subjective feelings expressed in the text need a huge amount of news stories to train the sentimental model. I have a lot of choices, either using Twitter information or using mainstream news articles.

1. News stories:

* AG news(<https://registry.opendata.aws/fast-ai-nlp/>)
* <https://nlp.stanford.edu/sentiment/>
* <https://datasetsearch.research.google.com/search?query=News%20feed%20sentimental&docid=6srYA%2Bi3qZssY3ztAAAAAA%3D%3D>

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment .

Content, it contains the following 6 fields:

target: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)

ids: The id of the tweet ( *2087*)

date: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)

flag: The query (*lyx*). If there is no query, then this value is NO\_QUERY.

user: the user that tweeted (*robotickilldozr*)

text: the text of the tweet (*Lyx is cool*)

* <http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection>

**Data Set Information:**

From the original readme file (please consult it for more information):  
  
The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987.  
  
In 1990, the documents were made available by Reuters and CGI for research purposes to the Information Retrieval Laboratory (W. Bruce Croft, Director) of the Computer and Information Science Department at the University of Massachusetts at Amherst. Formatting of the documents and production of associated data files was done in 1990 by David D. Lewis and Stephen Harding at the Information Retrieval Laboratory.  
  
Further formatting and data file production was done in 1991 and 1992 by David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago. This version of the data was made available for anonymous FTP as "Reuters-22173, Distribution 1.0" in January 1993. From 1993 through 1996, Distribution 1.0 was hosted at a succession of FTP sites maintained by the Center for Intelligent Information Retrieval (W. Bruce Croft, Director) of the Computer Science Department at the University of Massachusetts at Amherst.  
  
At the ACM SIGIR '96 conference in August, 1996 a group of text categorization researchers discussed how published results on Reuters-22173 could be made more comparable across studies. It was decided that a new version of collection should be produced with less ambiguous formatting, and including documentation carefully spelling out standard methods of using the collection. The opportunity would also be used to correct a variety of typographical and other errors in the categorization and formatting of the collection.

* <https://www.kaggle.com/therohk/global-news-week>

**Context**

This dataset is a snapshot of most of the new news content published online over one week. It covers the 7 Day-period of August 24 through August 30 for the years 2017 and 2018.

Year 2017: 1,398,431 ; Year 2018: 1,912,872

It includes approximately 3.3 million articles, with 20,000 news sources and 20+ languages.

This dataset has just four fields (as per the [column metadata](https://www.kaggle.com/therohk/global-news-week/data)):

publish\_time - earliest known time of the url appearing online in yyyyMMddHHmm format, IST timezone

feed\_code - unique identifier for the publisher or domain

source\_url - url of the article

headline\_text - Headline of the article (UTF8, Any possible languages)

See the ["Basic Feed-Code Exploration"](https://www.kaggle.com/therohk/basic-feed-code-exploration) notebook for a quick look at the dataset contents.

* <https://www.kaggle.com/aashita/nyt-comments>

**Context**

New York Times has a wide audience and plays a prominent role in shaping people's opinion and outlook on current affairs and also in setting the tone of the public discourse, especially in the USA. The comment section in the articles is very active and it gives a glimpse of readers' take on the matters concerning the articles.

**Content**

The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves. The csv files for comments contain over *2 million comments* in total with *34 features* and those for articles contain *16 features* about more than *9,000 articles*.

1. information from social media

* <http://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms>

**Data Set Information:**

This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn.

The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: economy, microsoft, obama and palestine.

This data set is tailored for evaluative comparisons in predictive analytics tasks, although allowing for tasks in other research areas such as topic detection and tracking, sentiment analysis in short text, first story detection or news recommendation.

Further details on the process of building the data set are provided in the article mentioned in the 'Relevant Papers' section.  
An .R file is provided to provide a simple introduction to handling the data set.

1. **Approach**

I view this project as a kind of classification problem. Basically, I will use supervised learning method, which including NLP and time series skills.

I will try both traditional ML algorithms and deep learning methods, and find the best approach. My target is whether an equity will go up or drop down in the future. By doing so, I will handle historical price to build a price feature, and deal with news articles or twitters to build a real time emotional analysis engine. By tuning the sentimental analyzer, I can use the NLP model to predict a news or twits’ affection toward equity’s price. And finally, combing those features together to predict the up/down of a stock.

1. **Final Delivery**

Typically, I will deploy a web application to deal with real time data and give a result of dedicated stock target.

1. **Computational Requirement**

**To be discussed.**

**CPU:**

**Memory:**

**GPU:**