Title

* Introduction and Objectives
  + Better understand the forces which drive the Apple stock price
  + Machine learning and sentiment analysis is able to quantifiably
* Context
  + Get some facts and figures about people doing this at banks
  + Quants?
* Methods
  + Data Gathering

Initially, this project looked to follow the work of Coyne, Madiraju, and Coelho by using text data from the trading social website Stocktwits (Coyne, et al., 2017). This site consists of essentially a trading-focused community where members may post their opinion into the public area and discuss. However, this paper found that since then, obtaining permission and access to Stocktwits’ database has become untenable through current company administrative limitations. Therefore, this paper reasoned that Twitter is a highly suitable alternative dataset. While it differs from the dataset used by Coyne, Madiraju, and Coelho, it is also likely to have a far larger base of information as Twitter’s monthly active userbase is approximately 310 million, compared to 1.5 million from Stocktwits (Roof, 2016; Statista, 2021). Furthermore, while Stocktwits is more likely to have an informed audience relative to the context of this project’s problem, this paper argues that more people invest in large blue-chip companies like Apple than is contained within Stocktwits, while also overall sentiment of the population and market is more likely to be better captured within a larger population like that of Twitter. Therefore, this paper concludes that Twitter is a highly appropriate source of sentiment data for the purposes of this analysis.

The primary dataset of text was sourced using the Twitter API to gather all tweets within the specified timeframe of 11/05/2020 to 11/05/2021 containing the hashtags “#apple”, or “#AAPL” (relating to Apple’s US stock market tag). Gathering this data was a significant activity and less straightforward than initially projected. The opening approach selected to obtain the tweets was the Postman tool. This has a user-friendly interface and an uncomplicated query builder. However, the researcher quickly noted that a core part of retrieving tweets from Twitter is the limitation on academic projects of a maximum of 500 tweets per request. While a key known as the “next token” provides a link to the subsequent part of the query result, it was clear the Postman tool was not easily optimised to automate gathering the full set of data. Therefore, this project reviewed alternative options to gather the necessary data. One option reviewed was Tweepy.

The alternative which was implemented was to adapt manual Python code which directly interacts with the Twitter API without a ready-made interface. This process was reasonably lengthy as there was a great deal of customisation required such as how to structure the loops, data formatting, and storage. Furthermore, the Twitter API has a number of restrictions on the number of tweets which may be requested over a variety of timeframes, and therefore a series of IF statements needed to be adapted to the code to pause it when necessary, ensuring the process would run uninterrupted and without any automatic throttling from the Twitter API. However, this process was positive through relative comprehensibility and flexibility as the dataset was constructed on the familiar and flexible Python Jupyter notebook interface.

The Apple stock price dataset was sourced directly from the NASDAQ website which contained a variety of information regarding the stock price, including its daily open and close prices, the trading volume, and the high and low for each day. For the purposes of this project, the daily price change was extrapolated and merged alongside the relevant dates for each text entry retrieved from Twitter, allowing the project’s models to view each tweet in context of that day’s price change.

Data Pre-processing

One significant pre-processing step was to ensure the full text of each tweet was delivered. While standard tweets were satisfactory, retweets had their text automatically truncated, minimising the available text data for these tweets. A retweet is a …

This paper considered and concluded that while a retweet is not the same as someone writing and expressing their opinion, it can be one’s way of expressing one’s opinion through the medium of another’s words. Therefore, this paper can assume that if a tweet is retweeted, the user has read the initial tweet and feels a similar sentiment which they wish to echo amongst their online social group. Therefore, this paper has concluded that these retweets are as important as original tweets and will be treated as such. Consequently, the truncation of retweets was rectified by identifying which tweet the retweet referred to, and copying the original untruncated text from the original tweet onto the retweet, ensuring there would be no truncation and compromise with the sentiment analysis.

* + Data Analysis
  + Requirements Analysis
  + Design
  + Implementation
  + Testing/Evaluation
* Experiential Results
* Discussion and Critical Evaluation of Findings
* Evaluation, Reflections, and Conclusions
  + Lessons Learnt
  + Future Research
* References