Market Movement Predictor

Tensorflow Machine Learning Algorithm Utilizing Analyzed Tweets and OHLC Market Data

University of Regina

ENSE 489

Maksim Sharoika and Muhammad Zaman

July 2023

Executive Summary

In today's day and age, around 54% of people in America are involved in financial markets, allowing them to set themselves up for a possibly early retirement or a higher standard of living in retirement. Out of all the investors, around 90% of day traders (swing traders) end up losing money instead of exiting the market with profits. This is because of a few reasons, some of them being, lack of knowledge, unrealistic expectations, and psychological reasons such as fear of missing out (FOMO). This is why it is usually recommended that investments are done for the "long-term" instead of the "short-term"; i.e. investing for retirement instead of for quick gains. Therefore we want to focus on the long-term; attempting to increase long-term performance instead of chasing short-term unrealistic gains. The majority of the factors influencing social markets are human factors that we believe can be mitigated with Machine Learning, but human factors in many cases can be unpredictable and must be treated as such. We built a machine learning algorithm that uses stock data alongside Twitter sentiment, to give us an educated prediction on the stock's performance - this is done by comparing today's close to the predicted close tomorrow; and giving a prediction on if it will increase or decrease.

Installation instructions; and further technical usage details are available in the README.md within our project repository.

GitHub Repository: https://github.com/sharoika/MarketPredictor

Introduction

The average salary in the United States of America is $54,000.00 per year, with a salary like this, the likelihood of retiring with a significant savings portfolio (enough to retire) is rather bleak - this almost states that investing is mandatory for retirement; and not enough people - especially young people are doing it. Investing is powerful due to the idea of compound interest, which means that interest is earned on both the initial principal amount and the accumulated interest from previous periods.

The younger you begin investing the more "compound growth" you experience and thus the larger your portfolio can be by retirement. We aim to increase compound growth by a few percentage points (which when compounded is significant) so that the quality of life in retirement is higher; or for some people, it becomes a possibility. This would also have the added benefit of allowing people to start investing in the later years of life but still have the possibility of retiring with a fully replaced income; many people believe that they are "too old" to start investing and just don't until they are forced to work into their 60s and 70s.

For scope, we will be focusing on a simple prototype that realistically should not be used in the real-world as markets are an unpredictable environment and we do not hold any liability for losses incurred.

Analysis of Social and Economic Issues

As of 2016, the percentage of families that invest in America is 52%. This accounts for families that have some level of investment in the market (this includes individual stock, mutual fund, 401k, IRA, etc.) (Parker, 2020), and in the long-term in many cases is the start of creating generational wealth. The average return of the SP500 index fund (average return) over the last 30 years is 7.5%; if we assume an individual has been investing $1000.00 monthly for the last 30 years they would currently have roughly $1,350,000.00 which is significant, but realistically not enough to currently retire in Canada; sources state you need about $2,500,000.00 in investment to retire with an "average Canadian income" (Srivindhya Kolluru, 2023).

Let's examine the famous investing rule that Warren Buffet mentioned: loses beat gains; it means if you lose 10% in the markets, you need to earn about 11% to get back to where you are. This leads us to believe that to invest "above" the average we would need to have a method of avoiding losses instead of securing large gains - this is where our AI comes in with next-day

close price predictions. We could attempt to guide investors to drop assets at a time when they are worth more if they are predicted to decrease. This, in essence, would factor towards "loss-avoidance" with an added-on benefit of knowing if a stock is favored to move upwards - so assuming we can bring the average investor's rate of return from 7.5% to 10% which is reasonable in respect to above average performance of investing giants such as Berkshire Hathaway ~20%. Their final investment portfolio amount would be roughly $2,260,000.00. This is significant; this is the difference in tens of thousands of retirement income and quite frankly the difference between enjoying your "golden years" or working till the day you die.

As a reminder, this should not be used as the primary decider for whether you should invest or not, as this is not 100% accurate and it will be inaccurate at times, it should be used as a confirmation or a second opinion after doing your research. Relying on an AI will hinder your ability to research on your own, and also not give you any meaningful skill when researching/trading.

Ethically there is no negative on the individual basis; it increases the quality of life - socially.

Design of Prototype System

Originally we thought of building just a simple market predictor, which would provide predictions based on closing price and other OHLC metrics. That was decided to be too simplistic and would ignore social sentiment; which is a huge part in markets therefore, not serve many benefits, so we backed our predictions with research done via Twitter sentiment scraping.

Therefore the requirements for our solution that we set for ourselves were the capability of reading and analyzing some form of social media sentiment (Twitter), the capability of receiving and using stock metric data (Yahoo) and finally, the ability to combine all of those inputs with a machine learning algorithm trained in tensorflow to give accurate predictions.

The Twitter scraping and analysis were designed from scratch using a web library instead of the Twitter API due to their API currently being extremely unreliable and more importantly not being open for public use - we were worried we would lose access for some sort of infraction.

The tools we used are as follows: TextBlob, yFinanace, PlayWright, Tensorflow, and various generic python mathematical libraries.

It is important to note the changes from our proposal; we decided the capability of predicting the actual closing price would be too difficult with the smaller data set and simpler inputs we are using. This was supported by extremely incorrect prediction during testing - therefore we transitioned to a binary “increase” or “decrease” model.

Implementation of Prototype System

The architecture of our solution is rather robust; 4 major parts are all crucial for the flow of data from collection to our market prediction. They can be broken down into 4 major functions, Twitter data collection, Twitter data analysis, stock data collection, and finally all of that data is fed into the predictor algorithm that we trained. The training was done with data from stock and tweet datasets from September 30th, 2021 to September 30th, 2022; and the process of training was the same except for us having the historic increase and decrease from day to day to train off of. The Tweet dataset was retrieved from Kaggle (Yukhymenko, 2022) and the stock dataset was requested from yFinance via API calls.

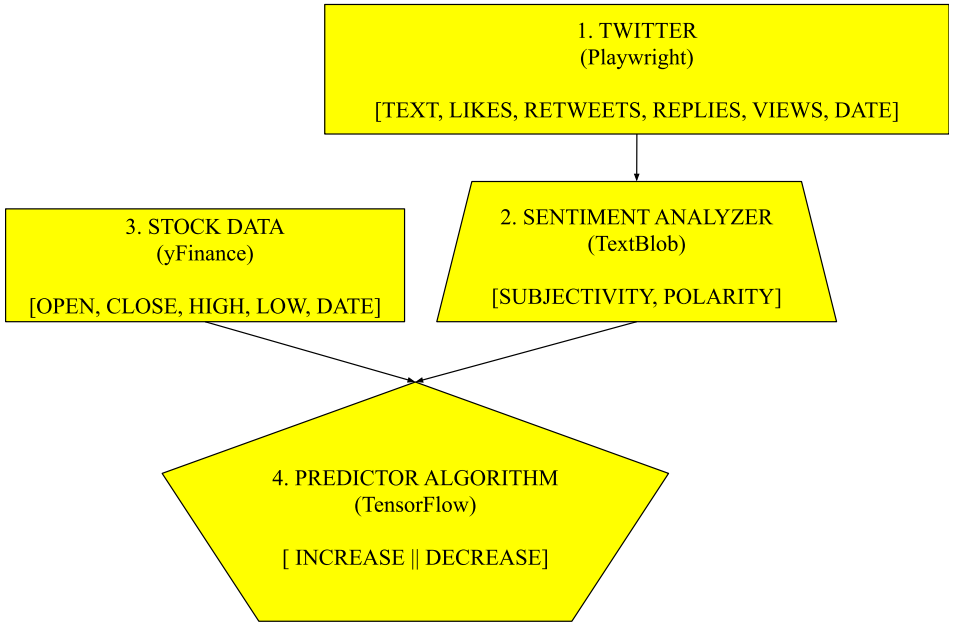

Figure 1: Market Predictor Software Architecture

The information within the “( )” brackets, is the tool that is used for that portion; the “[ ]” brackets represent the data which is passed down from that portion; the final step passes the data back to the user. The only input that the user needs to provide is a stock ticker; the ticker is a small character string used in the financial market to designate an investment.

1. Twitter data collection: This portion is broken down into two main components; Twitter authentication, and Twitter collection. The authentication is built with 2 pathways; the first is using the email followed by the password, and the second is using email, then the username, and then the password. The path chosen upon login is random. We combat this by attempting path one; if it is taking too long we timeout and attempt path two afterwards. This leads to the login being successful at all times. Next, we use the input stock ticker along with a hashtag in the form of "#[ticker]" to search for relevant tweets; using the DOM we then scrape them and remove duplicates uploading the data to a JSON file within the project.
2. Twitter data analysis: This portion of our algorithm is rather simple. First we read the entire tweet JSON from our project and save it within an array of variables. Then, using the sentiment analysis tool within the TextBlob package we analyze each tweet and get two important float values from it; one being subjectivity, how sure we are that this is discussing the markets; and polarity, which is the actual sentiment that the tweet is discussing (positive or negative). This data is added and saved into the project structure for later use.
3. Stock data collection: For our software, we want to get the OHLC (open, high, low, close) for the stock we are analyzing. This is easily accomplished by using another Python package called yFinance (Yahoo finance) that has a simple API call that we can provide a ticker to and receive the information we require within mili-seconds. This is done; the data is then saved in a JSON within the file structure for later use.
4. Predictor algorithm: The predictor algorithm will finally read into memory the average subjectivity, the average polarity of the stock, and the OHLC data. The OHLC data will be normalized by dividing all the metrics by the open price; this is done to get a ratio usually between 0.9 and 1.1 of the stocks metrics; this is to account for the price disparity of stocks. Once that is done, the machine learning algorithm is fed the average polarity, average subjectivity, close-ratio, low-ratio, and high-ratio, and generates a prediction of tomorrow's close price in respect to today's close price it will either an increase or a decrease.

Discussion

All in all, the project accomplishes its goal of loss-prevention insights - Throughout the term of this project, we completed 4 major components that when they come together accomplish our objective with a reasonable accuracy of 72.56%. This was done by creating

predictions for 5 stocks per day, for 2 weeks, and then the results were tabulated and analyzed against the closing prices per day. A red cell means the prediction was incorrect while a green cell means the prediction was correct; this high accuracy is unexpected due to the simplicity of our design and I believe that the Twitter data we scraped at the time of analysis had above-normal significance.

STOCK	July 10 2023			July 11 2023			July 12 2023			July 13 2023			July 14 2023		
	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION
AAPL		188.61	INCREASE	189.16	188.08	INCREASE	189.68	189.77	INCREASE	190.5	190.54	DECREASE	190.23	190.69	INCREASE
AMZN		127.13	DECREASE	127.75	128.78	INCREASE	130.31	130.8	DECREASE	134.04	134.3	INCREASE	134.06	134.68	DECREASE
MSFT		331.83	INCREASE	331.06	332.47	INCREASE	336.6	337.2	INCREASE	339.56	342.66	INCREASE	347.59	345.24	INCREASE
NVDA		421.8	INCREASE	424.81	424.05	INCREASE	430.33	439.02	INCREASE	445.18	459.77	INCREASE	465.83	454.69	INCREASE
TSLA		269.61	INCREASE	268.65	269.79	INCREASE	276.32	271.99	INCREASE	274.59	277.9	INCREASE	277.01	281.38	INCREASE
STOCK	July 17 2023			July 18 2023			July 19 2023			July 20 2023			July 21 2023		
	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION	OPEN	CLOSE	PREDICTION
AAPL	191.9	193.99	INCREASE	193.35	193.73	INCREASE	193.1	195.1	DECREASE	195.09	193.13	DECREASE	194.1	191.94	
AMZN	134.56	133.56	DECREASE	132.71	132.83	INCREASE	133.39	135.36	DECREASE	134.07	129.96	INCREASE	131.34	130	
MSFT	345.68	345.73	INCREASE	345.83	359.49	INCREASE	361.75	355.08	DECREASE	353.57	346.87	DECREASE	349.15	343.77	
NVDA	462.89	464.61	INCREASE	467.01	474.94	INCREASE	474.64	470.77	DECREASE	465.07	455.2	INCREASE	457.88	443.09	
TSLA	286.63	290.38	INCREASE	290.15	293.34	INCREASE	296.04	291.26	DECREASE	279.56	262.9	INCREASE	268	260.02	
All the above predictions were done in the evening of the night before the next day; a green cell means the prediction from todays close to tommorows close was correct; while a red cell means it was incorrect.															
TOTAL	45	CORRECT	34	INCORRECT	11	ACCURACY	75.56								

Figure 2: Market Predictor Results Spreadsheet

Socially, the pros of our approach are simplicity and understanding. We know the situation that many Canadians and Americans are currently in and want to do our very best to support them. Unlike many other tools out in the market currently, we take into account the “social” aspect which is extremely powerful when you consider that the stock market is completely human-driven which means social-driven. The cons of our approach is that it is heavily reliant on the data we can retrieve from Twitter, if there is incorrect data due to people lying, or just trying to pull a prank on one another our machine learning algorithm will take their statements and still analyze them for the prediction.

Economically, the pros of our system are in the long-term we attempt to reduce losses; this can lead to almost a doubling of someone's net worth over a 30-year horizon; The issue with this is that the world changes extremely fast - and the reliability of Twitter as a social source can easily become outdated; it's important for us to integrate other sources such as Facebook, and Reddit and so on. Additionally, another con is the idea that we are playing with people’s life savings; mistakes could cost thousands and thousands of dollars and cause users to have a lower standard of living if those losses are realized; the complete opposite of our goal.

Conclusion

In conclusion, we believe that our tool can be very powerful in the long-term perspective, increasing a retirement portfolio by millions of dollars if working properly. We believe our

strength comes from leveraging reliable Python packages to do a lot of the “legwork” which allows us to focus on more theory of our predictor algorithm such as implementing the normalization of data features discussed earlier. The weakness of this program lies in the fact that it's not entirely accurate. Due to many factors, the stock market is extremely volatile and difficult to predict due to a lot of noise, information imbalance, and common global market shocks that are unpredictable with machine learning. The markets are unpredictable because quite frankly human beings are unpredictable - that is why we focus on long-term loss avoidance as the goal.

Future Work

There are a few improvements that we would have implemented if the project was longer-term; these include more automation; the ability for users to “save” the stock tickers they want monitored daily with email notifications. Alongside this greater insight if a stock ticker enters the “decrease” zone; we would also want to provide more than just a binary output of “increase” or “decrease” , perhaps adding a “significant decrease” and “significant increase” to better support users and improve economic insight. Socially, we looked into adding an ESG option (economic social and governance) input while training and analyzing to see if there is a market correlation for social “well-being.”

From a more technical perspective; a longer data set would have been preferable; one which had experience in a bear and bull market; Alongside this increase in social input locations since Twitter tends to favor stocks such as TSLA in the current environments (sharing ownership) and finally we would have wanted the input domains for the predictor algorithm to be a range, not a single day of OHLC insight; for example a 7 or 30 day range. Lastly, using more metrics than OHLC, there are few data sets with reliable long-term in-depth insights so we would most likely have to create our own using API calls from yFinance. We would have to go day by day and for a long period of time and compile the JSONS ourselves before training.

Division of tasks

We believe our group (Muhammad Zaman and Maksim Sharoika) put equal amounts of effort, care, and attentiveness into this project. Also, ChatGPT was used for the making of this project, although it was not used in writing any code, mainly to confirm our sources of information, and used as a source of getting information which we confirmed later on.

Maksim Sharoika: 50%
Muhammad Zaman: 50%

References

Gillham, D. (2010, August 17). *Trading the stock market – why most traders fail*. Wealth Within. https://www.wealthwithin.com.au/learning-centre/share-trading-tips/trading-the-stock-market#:~:text=Anyone%20who%20starts%20down%20the,per%20cent%20make%20money%20consistently.

Jones, J. M. (2023, May 31). *What percentage of Americans own stock?*. Gallup.com. https://news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx

Parker, K. (2020, March 25). *More than half of U.S. households have some investment in the stock market*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/03/25/more-than-half-of-u-s-households-have-some-investment-in-the-stock-market/

Russo, F. (2022, December 7). *How much does the average American make in 2022?*. First Republic . https://www.firstrepublic.com/insights-education/how-much-does-the-average-american-make

Srivindhya Kolluru. (2023, April 17). How much money do you really need to retire? here are the factors you should consider. Toronto Star. https://www.thestar.com/business/personal-finance/how-much-money-do-you-really-need-to-retire-here-are-the-factors-you-should/article_d26774cd-901e-5c4a-80b8-321af4eafdb2.html

Yukhymenko, H. (2022, December 5). Stock tweets for sentiment analysis and prediction. Kaggle. https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction