Machine Learning Approach to Stock Prediction and Analysis

Dr. Satish Kumar T.[1]

[1] BMS Institute of Technology, Bengaluru, Karnataka, India

Krishna Prasad R [2]. Bhal Chandra Ram Tripathi.[3]

Global Academy of Technology, Bengaluru, Karnataka, Germany
rkp_rgp@yahoo.co.in

Visheshwar Pratap Singh [4]

Aegis School of Business, Data Science and Telecommunication, Mumbai, Maharashtra, India

Abstract The Stock Market is becoming a highly anticipated field of analysis. The data emerging every moment in stock market globally is in petabyte size. The data analyzers are working continuously on the data generated in the stock market to make capital-based predictions. The effort is to predict the future of stocks by using the data and make the best use of the financial environment. The crucial perspective of all the analysis is generation of the relevant data and through reliable resources. The experiment is dependent on Twitter for generation of data through its API. The database of one-million data used to make the accurate prediction of 75%-79%. The use of Sentimental Analysis, Natural Language Processing and Convolutional Neural Network makes the backbone of the overall research. The benchmark algorithm named STOCKP is an attempt to touch the expected accuracy of prediction of stock market and make the best monetary stability.

Keywords: Stock Market, Twitter API, Sentimental Analysis, Natural Language Processing, Convolutional Neural Network, Data

1 Introduction

There is continuous flow of data all around from digital to analog resources. There are certain platforms and benches that used by resource personalities to produce their views [8]. The views are hypothetically a good ground of analysis and enjoyed most by the data analyzers to make certain assumptions. The researchers are applying a continuous attempts to make prediction against the most important aspect of economy i.e. Stock Market [11].

The Business Organizations and global leaders use the platforms such as Twitter, Facebook, and LinkedIn to produce their views through multiple resources. The views are when analyzed against the trends and sudden changes in stock market are found to be key role playing agent. The researchers and market analysts are closely using this generated data to perform certain level of future predictions. The researchers are continuously attempting to make the predictive algorithms to be 70% accurate but it is 55%-60% [9].

The use of iterative resources and traditional data analysis is lacking to solve the real time cases of prediction. The iterative approach consisting of extraction of data continuously and making the algorithm to run upon does not satisfy the increased demands of market. The attempts used are mostly requiring the human intervention, making the next model to be available for analysis late and the result affected by 1% -2%.

The proposed research work by us is employing the automated approach and use of machine learning. The domain of machine learning is a great advantage to the research as it is allowing the CPU and GPU to work in a balanced way and produce the best-optimized result. The source of data for the analysis is from Twitter using the Twitter API and social media platforms. The data generated when well balanced with the automated approach and algorithm being developed will be capable to produce the benchmark prediction[11,2]. The use of parallel computing to make the best use of GPU and CPU makes the computing of results in microseconds. The modules of Machine learning and parallel computing provides the freedom to make the vision to reality; the modules such as sentimental analysis, natural language processing, hash cloud, and etc. helps inn making the research possible.

In this research, we are capable of producing the benchmark results and capable of comparing with the results generated from existing platforms. The initial iterations and improvements are capable to match a benchmark of 55.7% but with the rigorous improvements in data and architecture, we are currently at a benchmark of 62%.

In this paper we are proposing a new vision to handle the feelings of individual and determining its significance in the current trends of regional stock markets. Using the architecture to be trained against the historical data and perform the training for few weeks. The initial data set consists of one million records from different individual's views over social platforms and the reaction to those views. The architecture is then made automated after rigorous iterated process.

The paper is divided into multiple sections including this section which provides the motivation of performing the research. The other sections include 2 Experimental Setup to perform the research 3 Motivation to perform the research 4 Architecture of the research 5

2 Experimental Setup

The tools, benchmarks, environment used in this research is mentioned in this section.

2.1 Processing Unit

There is immense amount of inflow and outflow of data used for the research through various platforms; the use of GCC compiler with better optimization resulted in the approximate usage of resources. We have considered to work with Open64 [3], LLVM [4], ROSE [5], Phoenix [6], GCC [7] but we moved forward with GCC. The GCC is an open source compiler with more language supports and better optimization result handling in current open source environment. It is capable of core data processing with the availability of results in microseconds.

2.2 Optimization

The open source GCC provides all around support to optimize the processing but the use of proper flags is an important user input. The initial training requires the appropriate flags to input with the data while processing. There are many flags available in GCC, which are unique in their perspective that makes it necessary to use global levels (-Os,-O1,-O2,-O3). After the use of global levels, we can make it more automated results in handling flag. The optimization of data is an important task due to the modules needed to deploy on the optimized data will lead a more accurate result.

2.3 Platform

The research is generic in nature and the platform used is common to most consumers. We have surveyed and found that the developers and predictors work more on Linux environment than other Generic environment. The use of cluster of Intel Core processor with Octa Core Processing, which is supported by 8 GB RAM and GPU of 2 GB and having Linux support, allows making the results be approximated.

2.4 Benchmark of Evaluation

We have set certain benchmarks for the inputs to be processed. The benchmarks are important, as it will make the quality of processing always improving rather than varying against examples. There are certain negotiable situations, which needs special attention, while evaluating. The views generated during the tension like needed to be taken utmost care, as the views are available few weeks before the actual tension. The benchmarks of predicting the stock market is very dependent upon the preprocessing of the data that is the views and if the data is lagging in certain preconfigured benchmarks then it is not recommendable to process further. If there are chances to repair the data generated through views then we can run appropriate module to make the cycle run. The stock market core algorithm is an engine, which requires preprocessed elements rather than the processing done there.

3 Motivation

The researchers in the field of financial ecosystem are facing a large amount of data for analysis. The data generating though social platforms, live television debate broadcasting, views and others are heterogeneous in nature and are complete indifferent in benchmark evaluation. There are about 20 stock exchanges around the world, which is self-capable of generating data of size petabytes. The stock markets are also interdependent in business; this makes it tough to analyze and make any prediction regarding next moment of stock. There are multiple approach to solve the business problem of stock prediction; mostly are acquiring the iterative and manual intervention roles[1]. There are cases that due to improper handling of views of dignities and the response to those views has made the opposite of expected results and made to lose the faith of individuals in such predictions.

We are motivated with the challenge and created an all-around research in collecting the important data generating to certain organizations views on social platforms. We chose Tesla and Amazon to analyze their stocks by the statement of their respective officials in public domains. There have been continuous monitoring of stock markets, which trade most socks of these organizations such as of countries India, United States, China, Switzerland and we found that there are certain criteria that these organizations make before releasing some statements in the platforms. We examined statements by Elon Musk on twitter and their relation to different country stock exchanges and we found that the most affected market is of United Stated stock exchange and the least is from China stock exchange. We made attempt to collect all such previous statement data on Twitter and response against them in stock prices; which made us realize that there are much similarity in history of events after analyzing one million records. Figure 1 shows the graphical representation of views presented by Executive of tesla on Twitter and their effect on stock market. It clearly shows that the people welfare views are affecting most the stocks and making the stock of Tesla getting most profit.

The research is solely dependent on the development of a Robust Algorithm and with the continuous attempts; it is near to touch the accuracy as expected. The algorithm employs the basic functionalities of Sentimental Analysis, Key Extraction, Key Communication, Neural Establishment, and Neural Optimization.

4 Architecture

The architecture of the research is simple but it requires a high processing background. In this section we will consider the Data Identification, Data Extraction, Data Cleaning, Core Algorithm-I, Core Algorithm-II phases of the architecture.

4.1 Data Identification

It is the crucial phase in the research. The need to extract the data is an important milestone but identification of appropriate data is much important than extraction. There is availability of surplus of data; it is important that what responses against a statement by an Executive would accepted for use. The views by an important official over twitter can lead to several million reactions but we need to extract the profiles of reactionary's.

This will help to save the processing time and to extract only the responses of people whose responses are analyzed as a de-facto.

The data identification phase not only saves the unnecessary data to be processes but also the resources, which used wisely, can produce more optimized results.

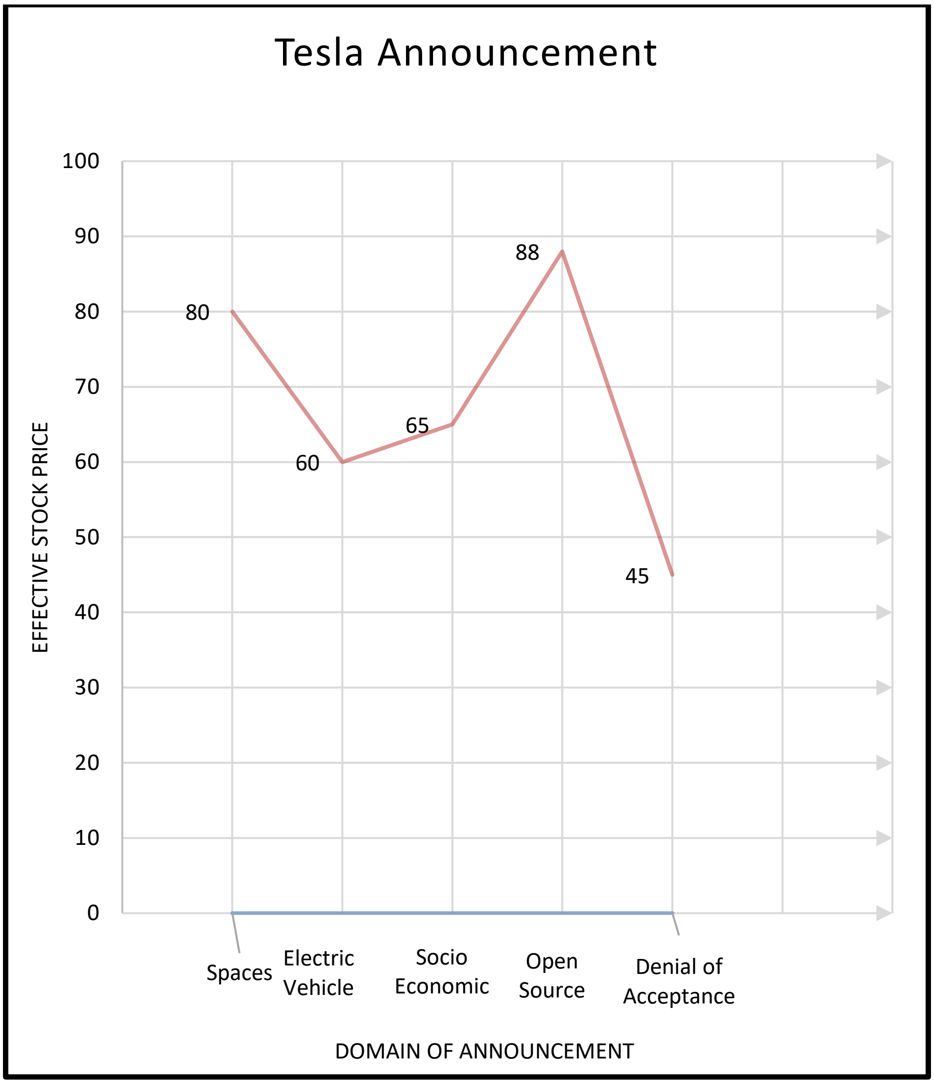

Figure 1 – The analysis of announcement by Tesla Executives on Twitter and its effect on stock market

4.2 Data Extraction and Data Cleaning

The scrappers is feed into with the results of data identification and it makes the extraction to be particular. The research consider in case having the twitter as a platform to be considered for extraction of data then the syntax to make it possible is-

```
for tweet in tweepy.Cursor(api.search,q="#%s"%u,tweet_mode='extended',
                        count=n,lang="en", since="2019-01-01").items():
      df = df.append({'TimeStamp':tweet.created_at,'Tweet':tweet.full_text},
                          ignore_index=True)
```

There are multiple ways to extract with the help of identification keywords.
The data is then cleaned to remove the symbols and unnecessary characters such that it is in pure sentence form without any extras. The below snipper is an one level cleaning and is part multilevel filtering.

```
def tidy(x):
   x = re.sub("@[\w]*", " ", x)
   x = x.lower()
   x = re.sub("(https?://[\w./]*)", " ", x)
   x = re.sub("[^a-z#]", " ", x)
   x = re.sub("#[\w]*", " ", x)
   x = ' '.join([w for w in x.split()])
   return x
```

4.3 Core Algorithm

The algorithm is only the heart and soul of our research. We are developing two phases of algorithm phase-I and phase-II, which will be making the expected results to be touched. GCC being a great optimizing compiler have supported the algorithm to work in full flourished manner[7]. The phase-I is having the modules such as sentimental analyzer, natural language processor, shift analyzer, redundancy settle. The algorithm is still in test phase with giving the expected result of 69% accuracy. The base of both the algorithms is to create a real world on what our imaginations[10,6].

The phase-II gets the inputs from phase-I and it is the only phase helping to touch the expected accuracy of 75%. It has modules, which are taken from dynamic programming, and word based analysis. The Figure 2 is able to showcase the overall flow of data and appropriate analysis. The technical aspects are fixed with the above mentioned processing capacity.

5 Conclusion

The aim of the research is to make the use of machine learning and neural network to make a reliable platform, which is accurate in financial stock predictions. We are very much confident with its infusions and beginning results; the algorithm based on basic to advance libraries has a perfect blend of imagination and reality. We are hopefully looking for better-optimized results in future enhancements by using cross-domain technology. We hope the future advancements will robust out vision.

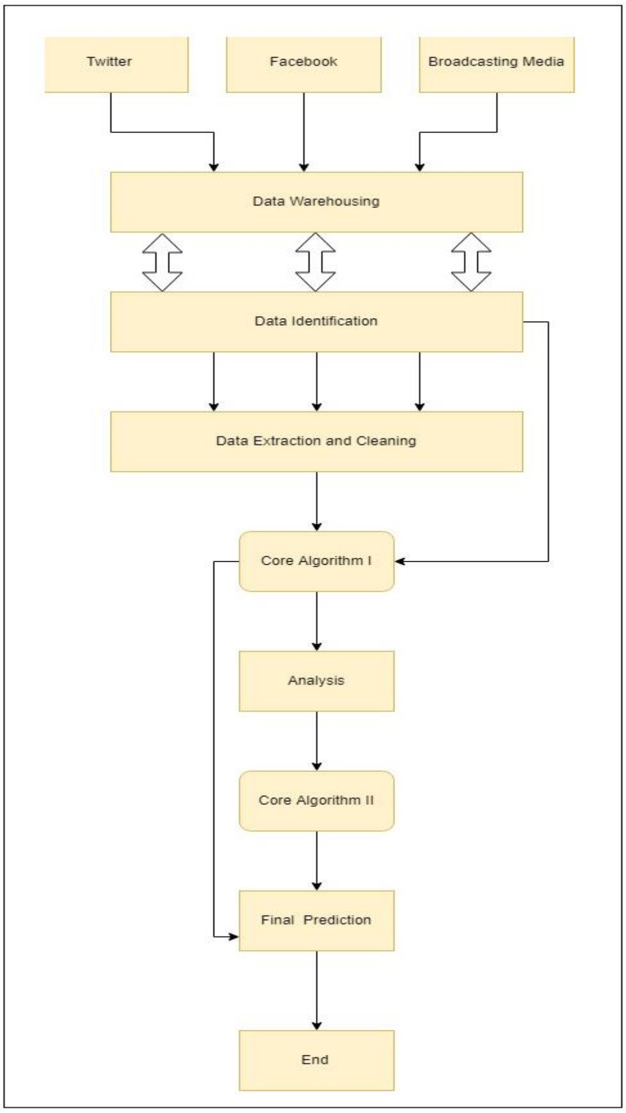

Figure 2. The Architectural Flow Diagram

6 Reference

[1] Asadifar, Somayyeh, and Mohsen Kahani. “Semantic Association Rule Mining: A New Approach for Stock Market Prediction.” *2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC),* 2017, doi:10.1109/csiec.2017.7940158.

[2] Youning jhang,“Sentiment Analysis.” *Mastering the Stock Market*, 2015, pp. 59–77., doi:10.1002/9781119204916.ch3. IEEE, 2016

[3] Open64: an open source optimizing compiler suite. http://www.open64.net

[4] LLVM: the low level virtual machine compiler infrastructure. http://llvm.org
[5] ROSE: an open source compiler infrastructure to build source-to-source program transformation and
analysis tools. http://www.rosecompiler.org/
[6] Phoenix: software optimization and analysis framework for microsoft compiler technologies. https:// connect.microsoft.com/Phoenix
[7] GCC: the GNU Compiler Collection. http://gcc.gnu.org
[8] Yauheniya Shynkevich-T. Mcginnity-Sonya Coleman-Yuhua Li-Ammar Belatreche - 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) – 2014
[9] (2018) Market Reactions at the Equity Offerings Announcement: A Short Window Event Study. Journal of Accounting and Finance. doi: 10.33423/jaf.v18i8.116
10 Jadhav, R. and M. S., W. (2017). Survey: Sentiment Analysis of Twitter Data for Stock Market Prediction. *IJARCCE*, 6(3), pp.558-562
[11] Donald A Bradley Stock Market Prediction Paperback – Import, Dec 1984