# Financial Market Prediction using Google Trends

Farrukh Ahmed
Research Student, Department of Computer Science & Software Engineering
NED University of Engineering & Technology, Karachi, Pakistan, 75270

Dr. Saman Hina
Assistant Professor, Department of Computer Science & Software Engineering
NED University of Engineering & Technology, Karachi, Pakistan, 75270

Dr. Raheela Asif
Assistant Professor, Department of Computer Science & Software Engineering
NED University of Engineering & Technology, Karachi, Pakistan, 75270

Muhammad Muzammil
Research Student, Department of Computer Science & Software Engineering
NED University of Engineering & Technology, Karachi, Pakistan, 75270

***Abstract*—Financial decisions are among the most significant life-changing decisions that individuals make. There is a strong correlation between financial decision making and human behavior. In this research the relationship between what people think and how stock market moves is investigated. The data from 2010 to 2015 of some of business, political and financial events which directly impact the local stock market in Pakistan is analyzed. The data was collected from search engine Google via Google trends. The association between internet searches regarding the political or business events and how the subsequent stock market moves is established. It was found that increase in search of these topics may lead to stock market fall or rise. The overall objective of this research is to predict Karachi Stock Exchange (now known as Pakistan stock exchange) 100 index by quantifying the semantics of international market. In addition to that, the relation between what an individual thinks while searching on Google which affects the local market is also investigated. The collected data has been mined by Multiclass Neural Network and Multiclass Decision Trees. The result shows that Multiclass Decision Trees performed best with an accuracy of 94%.**

***Keywords—Google trends; financial market; stock market; Karachi stock market; multiclass neural network; multiclass decision trees***

## I. INTRODUCTION

We are living in a world where data is generated from all domains of life. For example, from every social media interaction do, from every computer, every mobile, every sensor and now even from watches and other wearable gadgets. The real question is how we can convert this data into meaningful information for decision making such as to predict stock market behavior. Stock market prediction is a domain of challenging factors which is based on many important aspects and collective thinking of the financial experts. Stock Market data can be acquired from different sources. Its impact has generated considerable scientific attention due to its complexity and size. Despite its huge size, such data sets capture only the final action taken at the end of a decision-making process. No insight is provided into earlier stages of this process, where traders may use this information to determine what consequences of various actions and factors may be.

The aim of this study is to predict the behavior of local stock market in Pakistan based on available historical data and International market factors such as International Gold Rates, US dollar Rates, International stock markets and foreign exchange reserves etc. For today's world, data gathering frequently comprises of seeking on the web sources. Few years back Google has given access to cumulative information on the volume of queries for different search terms and how these volumes change over time, via publicly available service named Google Trends [1]. The gathered data was pre-processed using data cleaning and data filtering. The preprocessed data was than analyzed using Machine learning algorithms.

## II. LITERATURE REVIEW

The advent of Internet has seen people have used it as a main source of Information gathering and search engines like Google have become a gateway to this information. This fine grained data available on internet has opened up new options for researchers. Studies have found that large volumes of search queries for a specific word can linked up to real life events, such as forecasting the housing prices and sales [2]; popularity of films, games, and music on their release [3]; unemployment rate [4]; This openly available digital data also help researchers to find how the stock market moves, a recent study has found that increase in search volumes of some topics tends to precede stock market falls [5]. Some researchers have successfully found the relationship between behavior of people through social media (like twitter) and prediction of the stock market [6].

Karachi Stock Market (KSM) is one of the top 10 markets in the world. There are dozens of factors which impacts stock exchange directly or indirectly. That's why this research was intended to work on unique factors which impacts stock exchange. The objective of this research was to predict complex behavior of Karachi stock market (KSM) using historical data of KSM in combination with International economic factors such as US Dollar rates, gold prices, Oil

price, NYE 100 index data; Shanghai 100 index data are few factors which influence KSM.

## III. Data Gathering and Preparation

### A. Data Gathering

The datasets used in this study was gathered from different sources. These sources are described below:

#### 1) Data Generation from Google Trends Analytics

Research was primarily concentrated on gaining insight knowledge of what is going through people's mind and the most right source for it is what they search on Google. As we are aware of the fact that different events put a huge impact on financial markets. Therefore, the focus is on finding out the impact of these events via what people searched about it on Google.

The first task was to collect data about major events in Pakistan from 2010 to 2014. It includes finding major events from Wikipedia and other sources. Then these events are searched in Google trends and their *csv* files were downloaded. The major events from 2010-2015 are listed in **Error! Reference source not found.**.

TABLE. I. Year Wise Major Events in Pakistan

| Year | Major Events |
|---|---|
| 2010 | Air blue airline crash, Imran Farooq murder, Aman ki asha. |
| 2011 | Karachi bombing, Karachi target killing, Lahore bomb blast, Salman Taseer assassination, Osama bin laden killing, Raymond Davis scandal, Resignation of Hussain haqani. |
| 2012 | Karachi Factory fire case, Malala Yousuf Zai attempted killing, Memo gate scandal. |
| 2013 | Finance bill 2013, General Elections 2013, IMF Loan 2013, Meer Hazar Khoso become takecarer prime minister, Nawaz Shareef becomes prime minister of Pakistan. |
| 2014 | Ban on Geo TV, Attack on Hamid Meer, Karachi Airport Attack, Model Town Incident, Peshawar School Attack, Imran Khan long march, Operation Zarb-e-Azab. |

#### 2) KSM Dataset

The past 15 years of KSM data was collected. This includes Date, Index points, Volume, High and Low. Data was filtered from 2010 to March 2015 for research purposes.

#### 3) Dollar Rate Dataset

According to financial analysts, the International Gold Rate impacts KSM. Therefore, it was decided to put dollar rate prices, up and down measure in dataset for predicting Stock Market. The Dollar Rate data was obtained from Quandl [7]. It includes Date, Rate on that date, highest and lowest price and percent change in Rate. This data was obtained from year 2010 onwards.

#### 4) Gold Rate Dataset

Gold rate is another factor which impacts stock markets. Therefore, it was decided to put international market gold rate in this research. It was obtained from Quandl from 2010 and onwards. Parameters includes Date, Gold rate in USD and change in percentage.

#### 5) Newyork Stock Exchange Dataset

According to financial analysts, International stock markets impact KSM greatly. Therefore it was decided to use Newyork stock market index data set for building prediction model. Data was obtained from Quandal which includes date, volume, opening and closing index parameters. In addition to that, US International Index, composite Index and 100 Index were also used.

#### 6) Shanghai Stock Exchange Dataset

Another international market which impact KSM exchange according to financial analysts is the Shanghai Stock Exchange. It was decided to also use shanghai stock exchange parameters in the dataset. The data of Shanghai stock exchange was obtained from year 2010 onwards. Parameters includes data, opening, closing, volume and change in percentage.

### B. Data Pre-Processing/Cleaning

This section is focused on cleaning the data which was obtained by the process described above. Data was cleaned by going through series of steps each of which are illustrated below. Trends data obtained from Google Trends was not in proper format. Data was in three formats i.e weekly, daily and monthly. But the desired final csv was intended to be in day format from year 2010 to 2015. The data transformation was done from weekly to daily basis and is described in the following section.

#### 1) Weekly/Monthly to Daily Transformation

Weekly/monthly data was binded by date index column. Hence the string is aplitted in two parts that are start and end date. Thus a range was developed and then this range is converted into days by using *DateTime library* in python pandas package. The repeating values for the whole week/month generates a data file for each event .

#### 2) Merging of Data

This includes the following steps:

- Gather all event files year wise.
- Calculate mean of each file and saving the results in their respective year file. In this way, the five years of trend data in five different files can be acquired.

##### a) Merging of Data other than Trend data.

This step includes merging the data of Karachi stock exchange dataset, Newyork stock exchange dataset, Gold rates data set, Dollar rate data set, and Shanghai stock exchange data set.

*b) Developing the complete dataset*

The final dataset was formed by merging Newyork stock exchange dataset, Gold rates data set, Dollar rate data set, Shanghai stock exchange data set, Karachi stock exchange dataset and Trends dataset. The attributes of these datasets are shown in Table 2.

TABLE. II. ATTRIBUTES AND THEIR DATASET

| Dataset | Attributes |
|---|---|
| Newyork stock exchange dataset | NYSEInternational_100Index(Open), NYSEInternational_100Index(High), NYSEInternational_100Index(Low), NYSEInternational_100Index(Close), NYSEUS_100Index(Open), NYSEUS_100Index(High), NYSEUS_100Index(Low), NYSEUS_100Index(Close), NYSEComposite(Open), NYSEComposite(High), NYSEComposite(Low), NYSEComposite(close) |
| Shanghai stock exchange dataset | ShanghaiStockExchange(Open), ShanghaiStockExchange(High), ShanghaiStockExchange(Low), ShanghaiStockExchange(Close) |
| Dollar Rates dataset | Dollar Rates, Dollar (High), Dollar ( Low ) |
| Trends dataset | Trend |
| Karachi stock exchange dataset | Index(KSE) |

## IV. EXPERIMENTS AND RESULTS

A new column is added in the final dataset named Index-Difference-Class which is the label class. This label is obtained by the difference of subsequent observation of Index (KSE) column. If the difference is less than 100 observations than the value for label class becomes -1 and if the difference is greater than 100 observations than the value of label turns out to be 1, and 0 if the difference is between

-100 and 100.

Once the label class is obtained, the final dataset is uploaded in Azure Machine learning studio. The cleaning missing data module is used in order to remove observations which have missing data. After that Meta editor module is used to define the label class, which is index-difference-class in this case. Then split data module is used to divide data into training and evaluating data. The default setting that is used is70:30 in Azure machine learning studio[1]. The training data percentage was 70% of the real data and the testing data is 30% of the actual data.

The following modules of Azure machine learning studio are used to obtain the results.

*1)* Cleaning Missing data module[2].

*2)* Project column module**Error! Bookmark not defined.**.

*3)* Meta data editor**Error! Bookmark not defined.**.

*4)* Split data module**Error! Bookmark not defined.**.

*5)* Training model module**Error! Bookmark not defined.**.

*6)* Score model module**Error! Bookmark not defined.**.

*7)* Evaluate model module**Error! Bookmark not defined.**.

### A. Applying Multiclass Neural Network

Multiclass Neural Network is constructed by using training data and evaluated by using testing data. The trainer mode is single parameter with 100 numbers of hidden rows and learning rate of 0.1. The initial learning weight diameter was set to be 0.1 with momentum 0.

The steps for applying Multiclass Neural Network for training and testing data are as follows:

- *Training Model Module* is used for training data on applied algorithm.
- *Split Data module* was connected with Multiclass Neural Network module and Training Model Module.
- *Score Model Module* is used for testing a trained classification or regression Module.
- Split data module and Training Data Module are connected with Score Model.

The resultant confusion matrix for this experiment is shown in **Error! Reference source not found.**.

### B. Applying Multiclass Decision Forest

Secondly, the Multiclass Decision Forest is applied to the final dataset for evaluating results. The Resampling method is Replicate and trainer mode is single parameter. Number of decision trees are 8 and minimum depth of decision trees are set to be 8. Maximum depths of the decision trees are 32 and number of random splits per node is set to be 128. The minimum number of samples per leaf node is 100.

After applying all these settings, similar steps are executed as performed in the previous experiment. The confusion matrix for this experiment is shown in Fig. 2.

[1] https://studio.azureml.net/
[2] https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx

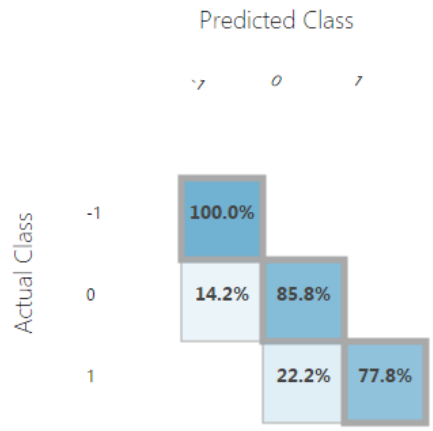


Fig. 1. Confusion matrix of multiclass neural network algorithm.

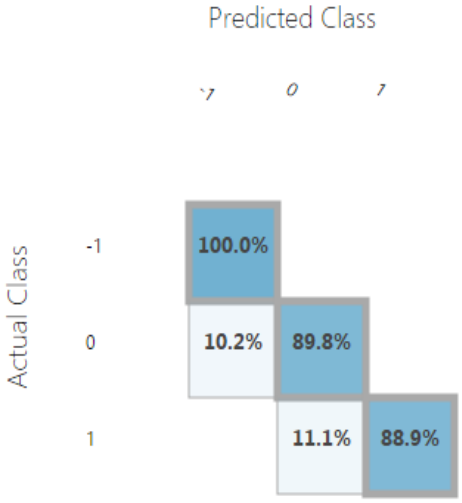


Fig. 2. Confusion matrix of multiclass decision forest.

### C. Evaluating the results.

The Precision, Recall and overall results of Multiclass neural network and Multiclass decision forest are shown in **Error! Reference source not found.**.

As it can see from Table 3, that multiclass decision forest has a higher accuracy than multiclass neural network.

TABLE. III. PREDICTING STOCK MARKET

| *Algorithms* | *Overall Precision (%)* | *Recall (%)* | *Average Accuracy (%)* |
|---|---|---|---|
| Multiclass neural network | 85.3 | 87.8 | 90.7 |
| Multiclass decision forest | 89.3 | 92.8 | 94.1 |

## V. CONCLUSION AND FUTURE WORK

The research was intended to predict KSM behavior by quantifying the semantics of people with the help of Google Trends Analytics. In addition to the datasets from Google Trends, this work also involves international factors which impact financial markets.

The research work can be expanded by introducing additional features in dataset. Other factors like inflation rate, interest rate, tax figures etc. can be used as supporting factors for improving the overall accuracy of algorithms. Once a stable model is established, work on a data product can be done which will be beneficial for investors. The data available on Google Trends or other resource can be utilized as an input to web service and getting the analysis and the prediction about how financial market moves.

The product will directly analyze people behavior and it will help investors in decision making process for buying or selling stocks. Investors will know about the overall geo political situation and will act upon it to get better financial outcomes.

## REFERENCES

[1] T. Preis1 T, H. S. Moat, H. E. Stanley, “Quantifying Trading Behavior in Financial Markets Using Google Trends”, SCIENTIFIC REPORTS 3 Article number: 1684, DOI: 10.1038/srep01684J., 2013.

[2] W. Lynn, E. B. Jolfsson, “The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities”, *ICIS 2009 Proceedings*. 147. http://aisel.aisnet.org/icis2009/147,2009.

[3] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, “Predicting consumer behavior with Web search”. Proc Natl Acad Sci USA 107(41):17486–17490, 2010.

[4] N. Askitas, K. F. Zimmermann, “Google econometrics and unemployment forecasting”, Discussion Paper No. 4201, 2009.

[5] C. Chester, P. Tobias, H. E. Stanleya and H. S. Moatb, “Quantifying the semantics of search behavior before stock market moves”, 2013.

[6] J. Bollen, H. Mao, X. Zeng, “Twitter mood predicts the stock market”, 2011.

[7] https://www.quandl.com/data/CURRFX/USDPKR-Currency-Exchange-Rates-USD-vs-PKR.