MAKERERE UNIVERSITY
P.O. Box 7062 Kampala Uganda
E-mail: info@cis.mak.ac.ug
URL: http://www.cis.ac.ug
Tel: +256 - 41 - 540628 / 534560 / 1-97
URL: +256 - 41 - 540620
COLLEGE OF COMPUTING & INFORMATION SCIENCES

FOREIGN EXCHANGE FORECASTING SYSTEM (FFS)

by

CSC19-37

Department of Computer Science
School of Computing & Informatics Technology

A Project Report Submitted to the
School of Computing and Informatics Technology for the Study Leading to
a Project Report in Partial Fulfilment of the requirements for the
Award of the Degree of Bachelor of Science in
Computer Science of Makerere University

Supervisor:
Dr. John Ngubile

Department of Computer Science
School of Computing & Informatics Technology

sepember, 2018

Declaration

We Group CSC19-37 do hereby declare that this Project Report is original and has not been published and/or submitted for any other degree award to any other University before.

#	**Names**	**Registration Number**	**Signature**
1	KAYIIRA TREVOR DUANE	16/U/5793/EVE	
2	KUIIMAKWE MARTIN	16/U/6385/EVE	
3	KYOBE JEREMIAH	16/U/6485/EVE	
4	NAKABIRI VERON	16/U/8434/EVE	

Date:

Approval

This Project Report has been submitted for examination with the approval of the following supervisor.

Signed:

_____

Date:

_____

Dr. John Ngubile
BSc. CS, MSc. CS,
School of Computing & IT
College of Computing & IT
Makerere University
johngubile@cis.mak.ac.ug

Dedication

We dedicate this report to the Almighty God without whom we can do nothing. We further dedicate it to our parents and guardians for their unceasing and selfless support throughout our stay in this university.plus all the lectures who provided us with neccessary information like professor Engneer Binomugisha,MR.Joseph Lomwa

Acknowledgement

We are deeply indebted to our project supervisor Dr. John Ngubile whose unlimited steadfast support and inspirations have made this project a great success. In a very special way, we thank him for every support he has rendered unto us to see that we succeed in this challenging study.

Special thanks go to our friends and families who have contained the hectic moments and stress we have been through during the course of the research project.

We thank the school for giving us the grand opportunity to work as a team which has indeed promoted our team work spirit and communication skills. We also thank the individual group members for the good team spirit and solidarity.And not forgetting the project coordinator how was there ever to keep us in line of our goal

November 19, 2018

Contents

Chapter 1

Introduction

Money is the most important and deciding factor in our day to day life and we should know how to use it efficiently. In order to use the money efficiently, countries and their people have started to invest in several businesses. While investing, they have also started to invest in other countries and shares which mainly depend on the currency exchange rate.

The well-planned business will always win. Similarly, a person who plan and invest by understanding the circumstances of the financial market (i.e. the ups and downs of the financial market) also wins. One of the most important deciding factors of the financial market is countrys currency exchange rate which is also known as the Foreign exchange rate.

Foreign Exchange rate market is one of the most difficult markets in the world that is effectively forecast because of its volatile and unpredictable nature. Hence, predicting of this Foreign Currency Exchange Rate (FOREX) has become important in the financial sector. By predicting the exchange rate effectively, the vulnerability of the trade investment can be decreased which provide smooth international trade and most importantly bringing maximum profit to the investors.

The main idea of this project is to predict the FOREX rate efficiently and also in a reliable manner. To achieve this, we will selected the best data mining models based on the past research and the ensemble model created out of it for the better accuracy. This entire model is developed on the big data using Sparkling water (Spark +H2o) in order to predict the FOREX rate quickly and efficiency.

Foreign Exchange Rate is the price of one countrys currency in terms of another countrys currency. Foreign Exchange Rates are one of the most important determinants of a countries relative level of economic health but tracking and understanding of these rates by traders and investors proves to be a difficult activity, they always have to keep an eye on reports and

news headlines to understand the current situation of the Foreign Exchange Market.

1.1 backgroud

Today financial markets are the heart of economies. Forex market is one of the largest financial markets where trillions of transactions are happening each day apart from weekends. It is important to forecast these market predictions for the betterment of the society. Financials markets run by means of information available.

Traders or investors always keep an eye on latest news or reports to understand the current situation of the FOREX market. The proposed system uses news headlines and predicts the directional movements of the market. Mainly, there are two types of market predictions; these are long term and short term predictions. FOREX Intraday trend prediction is short term prediction of FOREX market based on the news headlines released in the past couple of hours. Based on the news headlines, investors or traders will get an idea about the present situation and this information is used to match with the past similar situations and assumption is that the market behaves in the same way as it behaved in the similar situations in the past.
The main objective of the proposed work is to improve the accuracy of prediction of FOREX intraday market based on the related financial news. As the number of information sources increases due to internet, there is a huge availability of high dimensional data available in the internet. This can decrease the accuracy of predictions.

The main motivation behind this work is the use of multilayer dimension reduction algorithm to deal with the semantic redundancy as well as the sentiment integration. The multilayer dimension reduction algorithm consists of 3 layers. In each layer, it takes care of dimensionality reduction in an incremental fashion. The selection and processing of financial news for decision making is a challenging task.
Another motivation towards this work is the use of news headlines instead of news article bodies. News headlines give the summary of the news article and are straight to the point especially in short term predication.

In a typical short term trading environment Forex traders are mainly interested in three mutually exclusive events or outcomes. These three events are to find out whether the change of bid rate in the future between a particular currency and the US dollar will be up, steady or down. Our system predicts which of these three mutually exclusive events will come true

1.2 problem statment

Foreign exchange for the past has been a hidden market in the country where the rate are only publicized in papers and over the televisions to few who can afford time to look through the media. This has been publicized with a high increasing rate without clear reason and so the public left out to ask why but with not answers.

The bank of Uganda has been trying to convise the public that this is as result of micro economic factor but still the magnitude of rais still was over that of the economic change thus ther had yo be a system that can easy and accurately predict these changes before they happen and if there is any remedy that can be taken to keep the increase minimum.

The proposed system predicts the intraday movements of a currency pair from the information available in the financial news-headlines. To predict the foreign exchange rate accurately, we proposed a data-mining model which uses Artificial with machine learning, Regression based Random Forest, or Generalized Linear Model using Sparkling water.

The FOREX is one of the challenging and difficult markets to predict, several authors Yu et al. (2005), suggested that single model is not sufficient to predict the FOREX accurately hence in this research an ensemble of three different model is used. Due to the large volume of data in the real-time the data from 1995 to 2016 (In this research demonstration the dataset from 2000-2016 is only considered because all the research is done in same time frame and the modern era is only considered) which can affect the performance of our ensemble model.

The data in this research also includes the data set of factors (Interest rate, Inflation rate (Consumer price index), Gross domestic value, Shares etc.) which mainly influences the forex. By taking this performance issue under consideration, in this research, big data technology of sparkling water is used.

This project involves several processes of implementation from requirements gathering, Data collection, development of the architecture, pre-processing of data, data analysis and the evaluation are explained in chapter 3 in the methodologies. This proposed model of implementation as per our expectation produced the best results of accuracy but requires careful formulation and refers to what has been identified and needs a solution in the practical world, theoretical world or both.

1.3 Objectives

1.3.1 Main objective

The main objective is to come up with foreign exchange forecasting computer software that predicts FOREX market dynamics in the country

1.4 Specific Objectives

i. To carry out a literature study of existing computer system used in bank of Uganda to calculate the foreign exchange rates .

ii. To carry out a preliminary study on the existing system used in the exchange market currently.

iii. . To design the system for the foreign exchange forecasting t.

iv. To implement the designed system..

1.5 Scope

Foreign exchange Market Analysis of currency is the way of buying and selling goods internationally and mainly with the exchange of currencies.This system focused on predicting the expected rate dynamic using data mining. This will be useful for new investors to invest in any economic market forex market must be well studied based on the various factors considered by the software.

Forex market includes daily activities like forex rate calculation, exchange of currency. The exchange provides an efficient and transparent market for trading in forex. Our software will be analyzing currency rates based on online news headlines and online database. The exchange rates of currency depend on some of the following factors.
News articles are an important part of market information and are widely read by investors and market participants. The behavior, opinions, and mood of some investors, often referred to as investor sentiment in the domain of behavioral nance, is said to be linked to the timing and publication of news. In recent times, studies in the domain of nance have turned to using text analysis to examine the content of news and to create a proxy for investor sentiment .

1.6 Significance

The currency exchange rate data in the Foriegn exchange market was said to be chaotic, random, and noisy in nature. In earlier days, the Random Walk Model and the Efficient Market Hypothesis were the two most widely used models based on fundamental analysis.The proposed system will provide the following advantages:

- Use of linear regression algorithm which is a good predicator algorithm in big data compared to the existing algorithm.
- The fact that countries like Uganda are till developing and so need such a system that can support the international trade.
- eases the work of the bank of uganda to control the rates of the local currency as a ay of forecasting it will be easy to tell the changes expected

Chapter 2

Literature Review

In order to have a clear understanding of work. The proposed system predicts the intraday movements of a currency pair from the information available in the financial news-headlines), the best part is starting with literature review. In our literature review, the research papers from 2000 and above are only considered because our ideal goal of bringing accuracy can be achieved using recent technologies. The more valued and cited papers are only considered for our literature review to get strong foundation for implementing the project.

2.1 A review of Foreign exchange currency rate prediction research projects in the past

There are several traditional FOREX rates theory models such as Interest Rate Parity, Purchase Power Parity and Balance of International payment for predicting the FOREX. But these models find it extremely difficult to predict after the implementation of floating exchange rate. After that, the prediction of FOREX has always been a challenging task and this led to the discovery of several other models for predicting the FOREX .

2.2 Machine learning model review

2.2.1 Auto regressive (AR)

Lin et al. (2013). According to Lin et al. (2013), Auto regressive moving average (ARMA) is best suited for stationary exchange rate and the

models like Auto Regressive Conditional Heteroskedastic (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity(GARCH) are suited for the dynamic exchange rate. After these models, there are several Non-linear models that were developed which produced better results than those ARMA, ARCH and GARCH. The author Lin et al. (2013) also had the same idea like ours in handling the large amount of real-time dynamic data using big data Map reduce concepts of cloud computing. The author used Linear regression prediction model on cloud computing to predict the foreign exchange rate.

2.2.2 Artificial Neural Network (ANN)

The several Non-Linear models are Artificial Neural Network (ANN), Genetic Algorithm (GA) and many others are providing better accuracy compared to other models.

The mode we proposed produced has an average accuracy of 97 percent which indicates that the implementation of projects can also produce better accuracy. There is a great difficulty in predicting the FOREX because of its high volatility and noise. The author Yu et al. (2005) recognized that Artificial Neural Network (ANN) is powerful in FOREX prediction than the traditional models. The literature review also shows that several types of ANN are used and compared for predicting the FOREX. According to author Yu et al. (2005) , the several types of ANN were used in the past such as

The author Yao and Tan (2000), Zhang (2003) provided a great indication that Artificial neural network is best suited for forecasting the foreign exchange rate. The authors demonstrated their model by predicting the foreign currency rate of United States Dollars(USD) against Deutsch Mark(DEM), British Pound(GBD), Japanese Yen(JPY), Australian Dollar(AUD) and Swiss Franc(CHF).

The author Kimata et al. (2015) also used Artificial Neural Network for predicting the forex of Solomon Islands dollar (SBD) against the United States Dollar (USD), Great Britain pound (GBP), Australian dollar (AUD), New Zealand dollar (NZD), and Japanese yen (JPY).

2.2.3 Multilayer feed-forward network(MLFN)

Multilayer feed-forward network(MLFN), Recurrent network, Clustering Neural Network Model (CNN), General Regression Neural Network (GRNN) for predicting the forex similar to our model

The author also used ensemble-based method for predicting the forex, as the author felt that single model is not sufficient for predicting the forex accurately.
The authors in their implementation, they have used United states dollars(USD) against Deutsche Mark(DEM), Great Britain Pounds(GBP) and Japanese Yen(JPY) for demonstrating their model.
The author used the specialized weighted method, in which USD and AUD have weighted value of 80 percent and the rest of the currencies shares the remaining portions.
The author Yu et al. (2005) used ensemble model of GLAR and ANN for predicting the forex and their results were outstanding. This idea of an ensemble can help our project also in providing better accuracy. The author also provided valuable information of splitting the time series data, as he collected data from January 1971 - December 2000 as the training sets and January 2001 - December 2003 as the testing sets.
This process of splitting the data can be followed in our project. As the model proposed by the authors has the lowest Normalized Mean Square Error(NMSE) and return rate (R) is high shows that the model implemented by the author is highly efficient.

2.2.4 Random forest with Regration

According to author Liaw and Wiener (2002), the random forest can be used for both classification and regression and also stated that it is one of the best classifier models which is very robust against over-fitting.

The author also demonstrated the regression based example using Boston House data and their output produced better accuracy results. Segal (2004) used random forest regression in their implementation of machine learning bench-marking. Among all the regression techniques, random forest was performing well compared to others for their implementation.

2.2.5 mathsmatical model

The authors Urrutia et al. (2015) implementation of mathematical model for predicting the exchange rate of Philippines is very supportive and encouraging to our project as the author also had the same idea of ours, in considering the main factors that influence the exchange rate like interest rate, inflation rate, import and export of trade using Auto-regressive integrated Moving Average (ARIMA).

The author implemented several mathematical models to justify the results and have done a detailed study of each factor using the Multiple Linear regression(MLR). The author concluded that Interest rate and Labour rate has the main contribution. As the author Urrutia et al. (2015) suggested that considering the factor influencing, the exchange rate in its prediction plays an important role. I wanted to do further research with Patel et al. (2014) to understand the main factors really contributing to the exchange rate prediction.
The authors theoretically provided promising evidence that the main factors influencing the exchange rate are Inflation rate, Interest rate, Capital Account balance, Role of speculators, Cost of Manufacture, Debt of the Country, Gross Domestic Product, and political stability Carbureanu (2011). These experimental results give a clear idea that the model proposed by the authors accomplished better prediction accuracy and performance.

2.2.6 comparision of the above allgorithim

Kamruzzaman and Sarker (2003) compared the Artificial Neural Network (ANN) with Auto-regressive Integrated Moving Average (ARIMA) for predicting the Forex. The author also compared their implemented model with the traditional Autoregressive Integrated Moving Average(ARIMA) which is used traditionally for predicting the Time series data. In this comparison, ANN produced a better accuracy than ARIMA with ARIMA and ANN of 73. Similarly,
The author also proposed that ARIMA is a traditional model for time series prediction which almost used for two decades and the ANN is the most capable model for handling the time series data. Like Yao and Tan (2000) also demonstrated the prediction for forex using the United States America Dollars (USD), Singapore Dollar (SGD), New Zealand Dollar(NZD), Great Britain Pound (GBP), Japanese Yen(JPY), and Swiss Franc(CHF).

The author used three models of ANN, Bayesian regression, scaled conjugate gradient and back propagation for prediction and they are compared. In their model implementation, the scaled conjugate gradient of ANN performed better than others.
The author Leung et al. (2000), observed and compared the General Regression Neural Network (GRNN) with several other models including Artificial Neural Networks type of multilayered feedforward network (MLFN) with several layers of hidden layers, several random walk models, and Multivariate transfer function for forecasting of foreign exchange rate.

The authors general regression neural network (GRNN) is used for predicting the FOREX of three different countries currencies i.e Canadian Dol-

lars (CAD), Great Britain Pounds(GBP)and Japanese Yen(JPY).
The authors models of GRNN produced better accuracy than the existing ANN.The model General Regression Neural Network (GRNN) implemented by the Chen and Leung (2004) is same as that of Leung et al. (2000), for the forecasting of foreign exchange rate. As the author felt that GRNN is strong structure and consumes less time for the training and prediction he also used GRNN for rectification of the error.

The author developed two stage of error correction for bringing better accuracy in the artificial neural network.

2.3 Related systems implemented

2.3.1 High frequency financial time series prediction

The purpose of this research is to apply machine learning techniques for predicting high frequency nancial time series. Experiments are conducted using several regressors which are evaluated with respect to prediction quality and computation cost. The obtained results are analysed in order to reveal parameter combination for particular regressor that yields the best results according to chosen performance criteria.

Motivation
Machine learning is a rapidly evolving subeld of computer science. It has enormous amount of applications. One of the application domains is nancial data analysis. Machine learning was usually applied for analysis and forecasting of daily nancial time series. Availability of high frequency nancial data became another challenge with its own specics and diculties. Regressors, being a signicant part of machine learning eld, have been selected as study subjects for this project.

Methods
An extensive quantied literature search is conducted in order to gain insight on nancial data analysis and prediction techniques with focus on machine learning approaches. High frequency nancial data from Oslo Stock Exchange is the main source of information to work with. Open-source machine learning library Scikit-learn, designed especially for Python programming language, is used to perform all experiments on the given data set. A list of regressor implementations is adopted in order to perform regression analysis and make predictions for particular values of the data set.

Results
The possible solution for prediction of price uctuations based on the sliding window approach is proposed in this paper. The approach is tested in a

series of experiments on four dierent regressors. The combination of parameters that yields the best results in terms of predened performance criteria are chosen as optimal for each regressor. A comparative analysis of the regressors performance is conducted. The test results suggest that short-term prediction (approximately 1 minute ahead) is more favourable for the given high frequency nancial data.

Conclusion

All the tested regressors have demonstrated the best prediction quality on short periods of time. The multilayer perceptron regressor has demonstrated the best results in terms of both error values and time expenses. Possible improvements to prediction technique have been suggested

2.3.2 StreamliningSmartMeterDataAnalytics

The system offers an information integration pipeline to ingest smart meter data; scalable data processing and analytic platformfor pre-processing and miningbig smartmeterdatasets and aweb-based portal fo rvisualizing data analytics results. The system incorporates hybrid technologies, including big data technologies Spark and Hive, the high performance RDBMS PostgreSQL with the in-database machine learning toolkit, MADlib, which are able to satisfy a variety of requirements in smart meter data analytics.

The data processing layer

.The core component of this layer consist of the job scheduler implemented using Quartz library , and the data processing algorithms running in thi splatform.The algorithms can be run on Spark,Hive,LinuxShell,SQLEngine or Python environment.

The data processing algorithms are the functional modules implemented for a certain purpose, such as data cleansing, data transformation, data anonymization, streaming processing, anormal data detection, etc.

The sources and the sinks represent the places where a job reads the datafrom and writes the nal results to,respectively.Since this platform is an open data platform, it supports various data sources and targets in the underlying, which is simply to implement the corresponding read and write interfaces.

2.4 Identified the gaps

The authors Leung et al. (2000), Kamruzzaman and Sarker (2003), Yu et al. (2005), Kimata et al. (2015), Lin et al. (2013) have done a great amount of work by developing several prediction models for predicting the FOREX in the field of finance.

As most of the authors except Urrutiaet al. (2015) and Yao and Tan (2000) are really focused on predicting the forex with several models and finding out which one better suits for forex prediction. They should also consider other attributes which also mainly contributes to the prediction of forex.

2.5 Conclusion

As these literature review is backbone of our project that really guided us to implement our model for prediction of forex. The literature review from Yao and Tan (2000), Chen and Leung (2004), Kamruzzaman and Sarker (2003), Czekalski et al. (2015) provided us that Artificial Neural Network(ANN) model has performed better than other model for predicting the Forex,since its the resent most used but also compaered it with the autoregrasion thereby we have considered deep learning algorithm as one of our prediction model which act as the ANN for better accuracy.
The author Yu et al. (2005), gave rigid foundation to our project that ensemble model that is using of two or more models can produce better accuracy than single model. Similarly, Lin et al. (2013) also supported the idea of using big data for prediction with his cloud computing model produced better accuracy. Especially, Urrutia et al. (2015) was the base idea to implement our project by considering the influencing factors for prediction. Thereby taking all these positive and supportive ideas we were able to predict to have project successfully with great efficiency, accuracy and better performance.

Chapter 3

Methodology

3.1 introduction

Our goal is to predict the foreign currency exchange rate(Forex) more accurately by implementing the proposed prediction model using sparkling water. In this process of implementation several phases are involved, which are requirement gathering, development of proposed architecture and data mining models, data analysis,proccessing and presentation of results, and comparison of the independent data set with the output for successful testing. The complete information about each phase is explained in detail below.

3.2 Development approach

3.3 System Analysis methods

To make this system more about solving the unexpected dynamics in the rates , we will need to collect data from the population we are designing the system for. Details like what percentage of the population own smart phones, what type of smart phone, how long they spend on their daily commutes etc. which information will be relevant as we develop the system. A questionnaire that will be used in data collection is attached in the appendix.

To collect this information, we will use random sampling as we need an unbiased technique in order to get a proper image of the population. Also we will use ODK (Open Data Kit) as a data collection tool to collect this information from the population.

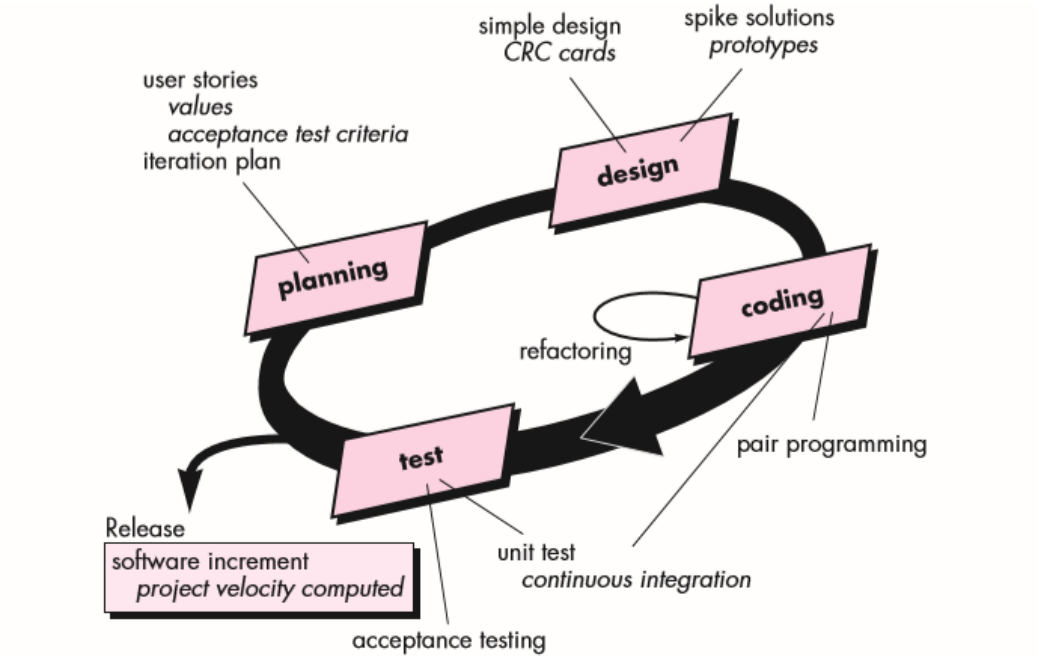

Figure 3.1:

3.3.1 system planing

Every implementation has to follow system development life cycle, that are Analysis, Design, Development and Evaluate. In the process of Analysis, we first analyzed our main goal of predicting the FOREX accurately and efficiently.

In analysis, we did several literature review which is discussed above, will help us to find out the best model for prediction, identifying factors that influences the FOREX rates, handling data for splitting test and train, and evaluating the results.

In the data analysis,The requirements gathering are done to find the type of data, the availability of data, storage system, tools for processing, analyzing the data and visualizing the results.

In the process of design and development, the design of the architecture is developed, implementation of proposed architecture and model will be conducted. Especially,

in the development phase data exploration, pre-processing, and data analysis using proposed models will be achieved. The results are presented for the evaluation and visualization.

Figure 3.2:

In order to make sure that the process flow implementation will be conveyed properly, every process of the process flow will be provided with detailed explanation in the figure below

time graph as show below

3.3.2 preminarly review

3.3.3 data collection

The first process of the requirement specification is the collection of the dataset, which is the main and prioritized requirement. It is unimaginable to complete this project successfully without finding the proper dataset. Since that project is about predicting the FOREX we first need to collect FOREX datasets. In order to evaluate the model perfectly, the currencies exchange rate of Uganda shilling- ugx against United State Dollars - USD, Canadian Dollars - CAD and Great Britain Pounds - GBP are used and the data collection are explained in detail below.

The historical data of foreign currencies exchange rate of Uganda shilling UGX against United State Dollars - USD are collected from http://www.investing.com , the Canadian Dollars - CAD and Great Britain Pounds - GBP are from https://www.oanda.com . As per the literature review suggestions, the factor that influenced the Forex rate are considered and their respective dataset are collected separately.

The factors that mainly influencing the Forex rate are, Interest rate, Gross

Domestic Product, Inflation rate / Consumer price index, Debt of the country, Producer price index, Share Prize, and Gold prize. The interest rate is one of the major factor that influences the Forex rate, it is the rate of interest charge by the country for any financial transaction. According to Compare remit (2016) and Patel et al. (2014), the interest rate is highly correlated with Forex rate.
The country with high interest rate will cause severe impact on the local business of the country. The country with higher interest rate decreases the purchase power of consumers and also people who plan to take loans has to pay high interest which will ultimately reduce the investors.

The data for the interest rate of United states, Canada and United Kingdom are collected from http://www.federalreserve.gov ,
http://www.bankofcanada.ca/rates/ and http://www.bankofengland.co.uk , http://www.bankofugand.co.ug respectively. Gross Domestic Product is the measure of goods and services of ones countries, and it also signifies the economic health of the country.
It is mainly based on complete expenses made by the country for their businesses, private activities and exports of goods. The country with higher GDP indicates that it has good economic growth which brings the focus of the investors Patel et al. (2014), Kayal (2010). The country with higher GDP has the lower foreign currency exchange rate. The dataset for GDP of all three countries (USA, Canada and U.K) is taken from https://data.oecd.org/gdp/.

The inflation rate determines the price of the product in a country and it mainly depends on the export strength of the country. The country with low inflation rate will attract the investors to invest more in these countries therefore, inflation rate has the important role in the valuation of the foreign currency exchange rate Investopedia (2013). In our project, the Consumer price index highly correlate with the inflation rate and was finding hard to collect the data set of the inflation for all the three countries of the same duration. Hence, the dataset for consumer price index is considered and collected from http://stats.oecd.org/ for all the three countries.
The country with more expenditure and also with low-income will cause debt to the country, this also reflects their economy. No countries in the world will get more attracted towards the country with high debt this makes no investors to invest in countries with high debts. So the country with high debts will have high exchange rate Uddin et al. (2013). The datasets for the debts of all the three countries are collected from http://stats.oecd.org/.

The producer price index is a weighted index to represent the countries

producing power from small scale to large scale. It shows the producing power of the wholesale, commodities market and manufacturing industries, this PPI helps in understanding the CPI of the country. Since CPI is highly correlated with exchanges the PPI is also considered for the prediction of Forex. The dataset for the PPI is collected from https://data.oecd.org/price/ for all three countries Investopedia (2013).
The financial economy of the countries also determined by the financial shares by the countries. The dataset for the shares for all the three countries are collected from https://data.oecd.org/ The Gold price is also another factor considered for the prediction of the Forex, the dataset for the Gold for all the three countries are collected from http://www.gold.org. All these datasets consist of the historical values of their respective fields and the data within the range of 2000 to 2016 is only considered for our analysis.

3.4 System Design methods

inthis section we will see the designe tools that we iused to descride the system to the users so that its easly understood by every one both developer and novic users.this tool include the architecture diagram over view structure etc.

3.4.1 system over view

This phase of our project is a very important phase because the raw data collected from several websites are treated with several techniques to make them useful for the model to predict the foreign exchange rate effectively and accurately. This phase involves the complete data analysis which involves the integration of data where all different datasets are integrated to each other to make a single complete data set fit into the model. The integrated data is then cleansed so that data is free from noise and NAs, the cleansed data has to undergo some transformation like converting the date to several useful information like which day of the week, is the day is weekday or weekend, etc. , the transformed data is now completely fit to the model for prediction, then these data are transferred to data analysis using data mining techniques where the training and the testing are done and the last part of the analysis are presenting and evaluating the results. as seenin figure 3.1

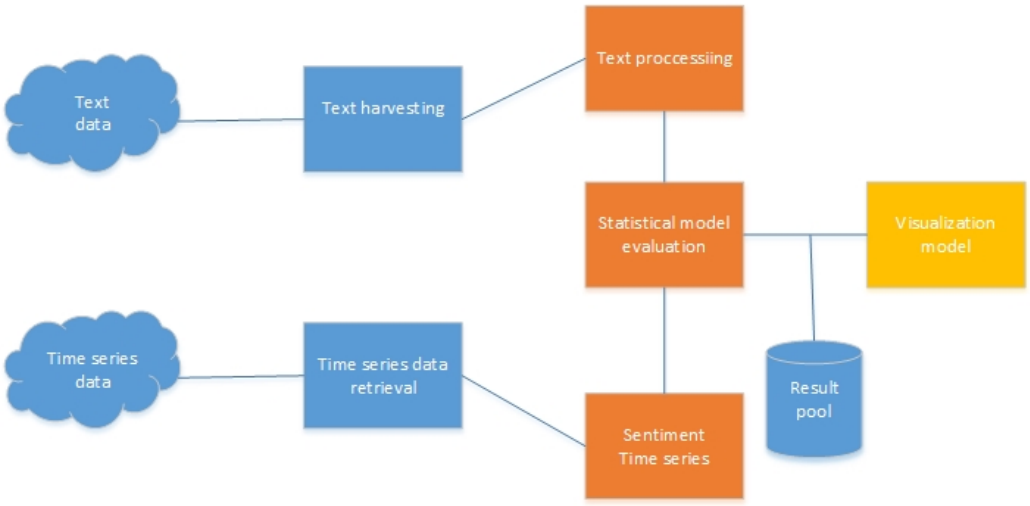

Figure 3.3:

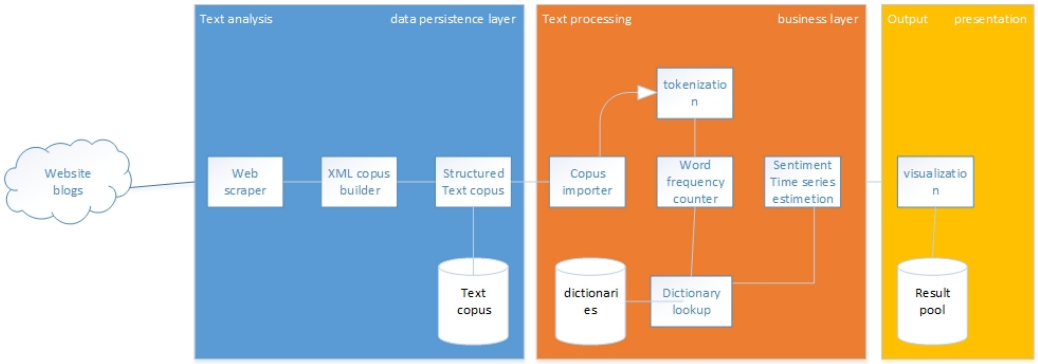

Figure 3.4:

3.4.2 system architecture design

In order to develop a strong and robust model for predicting the FOREX, a better architecture is developed that delivers the best accuracy, efficiency and performance. Architecture diagram is a great support for making this project a success, it has a dramatic representation of our strategy followed to make this project successful.

In this architectural design, it discusses what kind of tools and techniques are used in design, how they are used, in what order they are used and the interaction between them in order to attain our goal of Forex prediction accuracy. Any architecture in the Information Technology like Software de-

velopment, Data analytics.
Data modelling system follows three tier -architecture, i.e. Data persistence Layer, Business Layer, and Presentation Layer and the same type of architecture is followed in our system.
The Data Persistence Layer deals with the type of data to use, how it is used, where is stored and how other layers can interact with it are explained in this layer, the business Layer deals with all computation and the logic of our models are implemented in this layer and in the presentation layer, how the results are provided to . The users are implement in this layer. As discussed above, the data persistence layer has information about what storage system is used, why it is used, what type of data, are discussed here.
In this project, the data storage system used is Hadoop or spark sql or nosql and the reason choosing it is our model uses a large amount of historical data. It also supports the parallel processing which provides high computation facility by which we can produce high accuracy Wu et al. (2014) Deng et al. (2014). That is the reason, that the storage system used in this project is Hadoop distributed file system (HDFS). This entire HDFS system is available in the form of virtualized application using Hortonwork sandbox or juypter
The data collected for all the four countries including their respective factors datasets are integrating and stored in this Hdfs system and the type of data used is comma separated value file (.CSV) is used for our computation. There are three CSV files in the HDFS called USD.csv, CAD.csv,UGX.csv and GBP.csv which are integrated FOREX data of United States of America, Canada uganda and the United Kingdom respectively.
The business layer, where all the computations are taking place, the computations includes processing of data, training and testing of the model, prediction of the data are taking place. In order to make use of the big data to its maximum, the sparkling water (Spark + H2O) will be used for the spark cluster computing system with the deep learning of H2o. This helps in easier and faster data processing.
This is the layer where our ensemble model of deep learning, Distributed Random forest, and deep generalized linear regression is used for predicting the FOREX.

The presentation layer, helps the users to interact with the business layer where the model resides and the output of the model are evaluated and presented to the users. The H2o has a specialized GUI, where the users can interact with it, where the flow is created and the performance of the model is shown. The system is also provided with Spark terminal to scale program for interacting with data and the model. In this project, the visualization tool tableau is also used to compare the performance of the model by com-

paring the with the actual FOREX value with the predicted FOREX values.

3.5 System implementation methods

Statistical Tool or Econometrics Views, which is a new version of a set of tools for manipulating time series data, was used by the researchers in conducting SARIMA Modeling and Multiple Linear Regression. EViews provide regression and forecasting tools on Windows computers. With EViews a statistical relation can be develop from the data and then use the relation to forecast future values of the data. Areas where EViews can be useful include: sales forecasting, cost analysis and forecasting, financial analysis, macroeconomic forecasting, simulation and scientific data analysis and evaluation.

3.6 Linear Regression algorithim

Linear models attempt to quantify the relationship between present values of a random variable with the present value given all available information. In nancial time series analysis this random variable can include the returns or price of an asset. Considering the return of an asset rt, a simple model will attempt to explain or capture the linear relationship of rt using information available before time .

This information typically begins with historical values of rt (linear regression) and can include additional random variables with useful information (multiple linear regression). Additional variables might contain information about the economic environment and market structure in which the assets value is determined.

How these variables relate to each other plays an important role in model specication. Understanding how correlation and autocorrelation occurs is a basic tool for studying time series and as such is the building block of linear estimation for stationary series. Often a marginally statistically signicant autocorrelation is observed in nancial returns This suggests that a previous return value rt-1 may be useful in providing information about rt. This can be summarised in a typical autoregressive equation:

rt = A + 1rt-1 + b

where
A = the intercept value
1 = the regression coecient of the previous return value
b = the residual term assumed to have constant mean and
varianc

3.6.1 Multiple Linear Regression algorithim

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. A line in a two dimensional or two-variable space is defined by the equation:

Y=(a)X +(b)W +(c)Z

Where the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant is also referred to as the intercept, and the slope as the regression coefficient or B coefficient. In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. In general then, multiple regression procedures will estimate a linear equation of the form:

3.6.2 Stepwise Multiple Linear Regression algorithim

Stepwise regression is a modification of the forward selection technique in that variables already in the model do not necessarily stay there. [33] Stepwise regression is a semi-automated process of building a model by successively adding or removing variables based solely on the t-statistics of their estimated coefficients.
It is especially useful for sifting through large numbers of potential independent variables and/or finetuning a model by poking variables in or out.

3.6.3 Autoregressive Integrated Moving Average

An Autoregressive Integrated Moving Average (ARIMA) model is a statistical analysis model that uses time series data to predict future trends.
It is a form of regression analysis that seeks to predict future movements along the seemingly random walk taken by stocks and the financial market

by examining the differences between values in the series instead of using the actual data values. Lags of the differenced series are referred to as autoregressive and lags within forecasted data are referred to as moving average. For a response series * +, the general form of the ARIMA model is:

3.6.4 scraping and data set fomulation

Traditional beliefs held in nance of an ecient market with rational participants are not as denitive as rst theoris. The strict idea of an ecient market, that all information is incorporated into price at all times, would lead investors to believe that news and exogenous information would have no impact on the markets, having already absorbed any and all possible innovations.

The existence of speculative bubbles however, suggests that market participants are acting in a way that is devoid of information about the underlying value, exhibiting irrationality and making decisions based on emotional bias or assumptions and possibly overreacting to new information Financial analysts, particularly those engaged in fundamental analysis employ a number of data sources to make judgements on the true value of an asset.

Data sources can include raw nancial data such as price, exchange rates, or company earnings, market and industry reports or announcements, emergency and disaster events, geopolitical events, and government regulation. This information can be unstructured, qualitative in nature, and dicult to aggregate.

Much of this information is reported in news and contained in text. As such, quantitative methods to measure and aggregate this qualitative information are appealing. Information contained in news may inuence market participants. This information may then provide some predictive insight into the reaction of investors and therefore market movements.

The idea that there exists market participants trading on information, and those who trade on misinformation or stale information, described as noise, was discussed in the formative work by Fischer Black in Noise [11]. Black also highlights that noise is necessary for a market.

Although noise may contribute to making a nancial market imperfect, it supplies liquidity, of which those trading on information can take advantage of. As the number of noise traders trading increases, those trading on information must increase their position, taking on more risk. For liquidity to exist in a market, prices will become less ecient and as stated markets are ecient almost all of the time

machine laerning evolved a way the news can be used to determin the change in economy by use of data mining methods like scraping the fianancial new and time series in scrapy modules and libraries like beautiful soup4 in

spark environment to form csv files the can be treated as an input to any algorithm .

Python also comes up with a way of aggregate the information scraped from the web by by using its laerning abilities to fit the relevant data set and store as a csv file on the topic of your intreset in courps which can be imported at any time

3.6.5 data Aggregation and analysis

A regression model is built with dummy variables to account for business cycles and control variables for time series anomalies. The main findings reiterate that of previous studies that news text helps predict market while during recessionary periods the impact of the negative sentiment is stronger but less signicant than in expansionary periods.

Other findings were that by using the negative sentiment series extracted from the data columns, a one standard deviation change in negative words results in a basis point impact on the recessionary periods and basis points during expansionary periods. In the study by Tetlock the time varying impact of sentiment is also noted but not as thoroughly investigated. A partial reversal is also noted which reinforces the idea of the impact being a non-informational event but that of investor sentiment and a reaction to news.

A wealth of data is becoming available with databases of financial data being released for free and an even larger volume of unstructured data being published on the internet. Although these data are available, issues can arise with the collection and aggregation of such information.

Problems such as conflicting, non-unique, and ambiguous variables surround data collection and defination. Time zones, price quotes, and non-synchronous data can mean dierent time series of price data being retrieved from exchanges and brokers. It may not always be possible to accurately dene and sample data but a good knowledge and description of the dataset is beneficial and contributes to competent model estimation. These are some of the considerations in data collection and aggregation.

3.6.6 Data proccessing

The functions and models in this component were implemented in python. The main functions are divided into a data processing component and a modelling component.
The data processing component performs the retrieval of financial data and aligns this with the imported results from the text analysis component.
The modelling component runs dierent models to assess the impact that the

sentiment variable will have on changes in the price of financial assets.
These data and output are passed to a web based visualisation, the full output from the models is also saved to a file
All necessary transformations are performed using functions built with the python base library. These transformations include calculating returns for financial price data, z-score values for sentiment, aggregating data to other (lower) time frequencies if required, or other transformations that may be necessary to create a stationary time series.The first function of the transformations component calculates financial returns from the price series of the downloaded financial data.
The financial data downloaded will include several columns of time series data such as trading volume, or high and low price values for a time period. For a single observation the price quoted from an exchange is often the final price or transaction price for that time period.
For a daily price series, the price quoted for a particular day will be the price at the end of that trading day. This close price is what is extracted from the financial data and aggregated with the sentiment data.
The last component of the data pipeline can work in conjunction with the transformations component, this component tests whether the time series are linear and stationary.
Nonstationary data will often have trends or cycles occurring in the data which results in changes over time in central moment behaviour, changing mean, variance, and covariances.
Due the presence of these patterns, it is often unpredictable and dicult to model these data with linear models. An awareness of the anomalies of the series will help in determining what transformations may be necessary to obtain a stationary series.
For instance, by de-trending a series, of a deterministic trend can be removed by simple subtraction that does not affect the observations of the series but instead may transform the process from non-stationary to a stationary one. This allows the transformations component to be informed by the stationarity tests.
After data processing the time series data are passed to the model estimation component . In this component a vector autoregression, rolling regressions are computed.

The results and output of the models are stored and can be passed to an interactive GUI front end where some of the results are displayed. This is the final component of the system. The processed time series are passed to a vector autoregression (VAR) model that estimates the model and regression coecients for each of the included variables. The details of this model have

been described previously in literature .

The regression equation to be estimated can be specied prior to running the system. The system can update the data series and re-estimate and update the models periodically. In the VAR model estimation, a prior equation is supplied that is used in the evaluation of the system and has been drawn from the literature.

3.6.7 data visualisation

The presentation layer, helps the users to interact with the business layer where the model resides and the output of the model are evaluated and presented to the users.This is done by a specialized GUI, where the users can interact with it, where the flow is created and the performance of the model is shown.
The system is also provided with Spark terminal to python program for interacting with data and the model .This project will be using visualization tool such as tableau and matplotlab. this also is used to compare the performance of the model by comparing the with the actual FOREX value with the predicted FOREX values.

The visualisation component is the final part of the system and consists of a GUI interface built in python using the matplotlab and tabulea package and allows a clean interface to be wrapped around the statistical analysis component of the prototype system.

The GUI component is based on a JavaScript and HTML framework and allows the creation of a user interface with server side capabilities. All applications built using this framework use web based client side and server side design principle. Using the mentioned packages a user interface script must be created that controls the layout and appearance of the application (client side).

A server script must also be created to act as the controller of the application (server side). A nal helpers script is also used with the server script to provide additional logic, data retrieval, and functions that are necessary for data and results preparation for the GUI. Much of the framework acts as a wrapper for R functions and scripts, where more detailed functionality can be specied and then mapped to the UI created.

The goal of this framework is to allow users to showcase data analytic work with a minimum overhead of GUI design and deployment of a web application.
The web based framework makes deploying the system to a web server with a front-end possible and allowing the prototype to be more portable and work independently from a user or individual computer.

3.7 System Testing methods

3.7.1 introduction

The Evaluation of Prediction Model to be able to understand the performance of the models as it is the goal that we have to predict the Forex accurately. In the Evaluation process several techniques and tools are used to evaluate the model

In order to test how well the model t and able to predict the data, the test followed is called Goodness of Fit, which provided the accuracy of the model. This technique is used to test the accuracy of each and every model used in this project.

The models under the sparking water for our prediction of Forex uses two important fitness function namely, R-Square and Mean Squared Error Willmott and Matsuura (2005). R-Square, R-Square, shows the strength of the model and how well the model can predict the data and also explains the correlation between the actual value and the predicted values of the model. R-Square is also defined as the correlation between independent and dependent column.

3.7.2 Granger Causality Test

Granger causality is a statistical concept of causality that is based on prediction. Its mathematical formulation is based on linear regression modelling of stochastic processes. A variable x is said to Granger cause another variable y if past values of x help predict the current level of y given all other appropriate information. The simplest test of Granger causality requires estimating the following two regression equations:

3.8 overall testing and evalution

As a result, all the deep learning model was predicting the Forex accurately, among all the three model the Distributed Random forest was predicting the Forex accurately.

3.9 overall Conclusion

Our main goal is predicting the forex accurately, which we can successfully achieved by developing the prediction model using Deep learning algorithm, Distributed Random forest, Generalised Linear model in sparkling water. We were able to predict the model accurately at a fast pace which is the industry really aiming for. Our research, also provided the contribution of factors inuencing the forex and also made us clear that any exploratory analysis is not required for Deep learning.

As a future analysis, several other factors like political stability is unidentied and nding how the political stability inuences the forex rate. Also other factors like day today news, Fuel price, countrys president speech can also be considered of how it inuencing the forex rates.

3.10 Apendex

	items	**quatiity**	**unit cost**	**total cost**
	laptop	1	1500000	1500000
	internet package	3 month	150000	450000
	printing	4 copies	15000	60000
	transport	10 meet ups	16000	160000
	anonymous costs	unkwon	unkwon	100000
	air time	four month	20000@ month	80000
	total			2350000

3.11 Reference

Carbureanu, M. (2011). The analysis of currency exchange rate evolution using a data mining technique., Petroleum-Gas University of Ploiesti Bulletin, Economic Sciences Series 63(3).

Chandar, S. K., Sumathi, M. and Sivanandam, S. (2014). Neural network based forecasting of foreign currency exchange rates, International Journal on Computer Science and Engineering 6(6): 202.

Chen, A.-S. and Leung, M. T. (2004). Regression neural network for error correction in foreign exchange forecasting and trading, Computers Operations Research 31(7): 10491068.

compareremit (2016). money-transfer-guide. URL: http://www.compareremit.com/money-transferguide/key-factors-aecting-currency- exchange-rates/

Czekalski, P., Niezabitowski, M. and Styblinski, R. (2015). Ann for forex forecasting and trading, Control Systems and Computer Science (CSCS), 2015 20th International Conference on, IEEE, pp. 322328.

Deng, J., Qu, Z., Zhu, Y., Muntean, G. M. and Wang, X. (2014). Towards ecient and scalable data mining using spark, Information and Communications Technologies (ICT 2014), 2014 International Conference on, pp. 16.

Eng, M. H., Li, Y., Wang, Q.-G. and Lee, T. H. (2008). Forecast forex with ann using funda- mental data, Information Management, Innovation Management and Industrial Engineering, 2008. ICIII08. International Conference on, Vol. 1, IEEE, pp. 279282.

Gan, W.-S. and Ng, K.-H. (1995). Multivariate forex forecasting using articial neural networks, Neural Networks, 1995. Proceedings., IEEE International Conference on, Vol. 2, pp. 1018 1022 vol.2.

investopedia (2013). Forex tutorial: Economic theories, models, feeds data available. URL: http://www.investopedia.com/university/forexmarket/forex5.asp
Kamruzzaman, J. and Sarker, R. (2003). Comparing ann based models with arima for prediction of forex rates, Asor Bulletin 22(2): 211.

Kimata, J. D., Khan, M. and Paul, M. T. (2015). Forecasting exchange rate of solomon islands dollar against euro using articial neural network, 2015 2nd Asia-Pacic World Congress on Computer Science and Engineering (APWC on CSE), IEEE, pp. 112.

Leung, M. T., Chen, A.-S. and Daouk, H. (2000). Forecasting exchange rates using general regression neural networks, Computers Operations Research 27(11): 10931110.

Lin, S.-Y., Chen, C.-H. and Lo, C.-C. (2013). Currency exchange rates prediction based on linear regression analysis using cloud computing, system 6(2).

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models, Encyclopedia of statistical sciences .

Patel, P. J., Patel, N. J. and Patel, A. R. (2014). Factors aecting currency exchange rate, economical formulas and prediction models, International Journal of Application or In- novation in Engineering