VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

"Stockport | Predictive Sentiment Analysis"

Submitted in partial fulfilment for the award of degree(18CSI85)

BACHELOR OF ENGINEERING IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

ABHINAV PRAKASH

1TJ19CS001

Conducted by
VARCONS TECHNOLOGIES PVT LTD

Department of Computer Science and Engineering
T. John Institute of Technology

Gottigere, Bengaluru-560083

2020-2021

T. John Institute of Technology

Department of Computer Science and Engineering

Gottigere, Bengaluru-5600832020-2021

CERTIFICATE

This is to certify that the Internship titled **“Stockport | Predictive Sentiment Analysis”** carried out by **ABHINAV PRAKASH,** a bonafide student of T JOHN INSTITUTE OF TECHNOLOGY, in partial fulfillment for the award of **Bachelor of Engineering**, in **COMPUTER SCIENCE AND ENGINEERING** under Visvesvaraya Technological University,Belagavi, during the year 2022-2023. It is certified that all corrections/suggestions indicated have been incorporated in the report.

The project report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the course Internship / Professional Practice (18CSI85)

Signature of Guide **Signature of HOD** **Signature of Principal**

EXTERNAL VIVA

Name of Examiner

1) ____________________

2) ____________________

Signature & Date of Examiner

1) ____________________

2) ____________________

DECLARATION

I, **ABHINAV PRAKASH**, final year student of Computer Science and Engineering, T JOHN INSTITUTE OF TECHNOLOGY- 560083, declare that the Internship has been successfully completed, in **VARCONS TECHNOLOGIES PVT LTD**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Branch name, during the academic year 2022-2023.

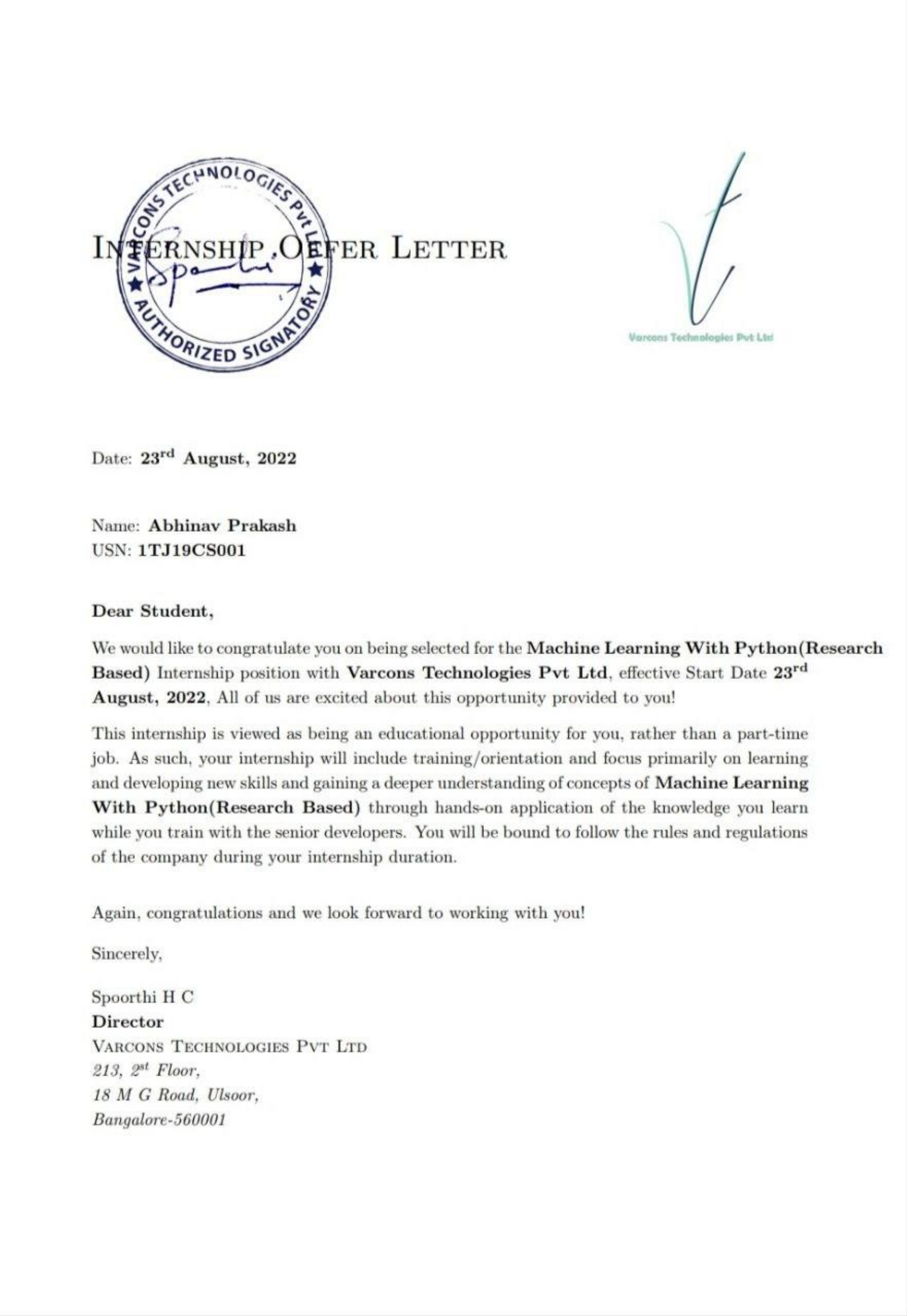
Date :25/09/22 :

Place : Bengaluru

USN : 1TJ19CS001

NAME : ABHINAV PRAKASH

OFFER LETTER



VARCONS TECHNOLOGIES Pvt Ltd ★ AUTHORIZED SIGNATORY ★

Varcons Technologies Pvt Ltd

INTERNSHIP OFFER LETTER

Date: **23rd August, 2022**

Name: **Abhinav Prakash**
USN: **1TJ19CS001**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With Python(Research Based)** Internship position with **Varcons Technologies Pvt Ltd**, effective Start Date **23rd August, 2022**, All of us are excited about this opportunity provided to you!

This internship is viewed as being an educational opportunity for you, rather than a part-time job. As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts of **Machine Learning With Python(Research Based)** through hands-on application of the knowledge you learn while you train with the senior developers. You will be bound to follow the rules and regulations of the company during your internship duration.

Again, congratulations and we look forward to working with you!

Sincerely,

Spoorthi H C
Director
VARCONS TECHNOLOGIES PVT LTD
213, 2st Floor,
18 M G Road, Ulsoor,
Bangalore-560001

ACKNOWLEDGEMENT

We are grateful to our institution **T.JOHN INSTITUTE OF TECHNOLOGY** with its ideals and inspiration for having provided us with the facilities, which has made this report a success.

We would like to express our gratitude to our chairman **Dr. Thomas PJohn** for providing us withthe necessary facilities for the successful completion of the project.

We also thank **Dr. P Suresh Venugopal,** Principal, T.John Institute of Technology, for providing us an educative environment to work.

We also thank **Ms. Suma R** Associate Professor & Head, Dept. of CSE, for his inspiration during the completion of report.

We worked on the project under the guidance of **Ms. SONIA DAS,** Assistant Professor, Dept. of CSE, whose wisdom and experience has enabled us to conduct the seminar work successfully.

We would also like to take this opportunity to thank other faculty members of our department who have helped us in various ways while preparing for this project. We are also very grateful toour **family** members and **friends** for their support and encouragement.

ABHINAV PRAKASH(1TJ19CS001)

ABSTRACT

In this paper, we apply sentiment analysis and machine learning principles to find the correlation between ”public sentiment” and ”market sentiment”. We use twitter data to predict public mood and use the predicted mood and previous days’ DJIA values to predict the stock market movements. In order to test our results, we propose a new cross validation method for financial data and obtain 75.56% accuracy using Self Organizing Fuzzy Neural Networks (SOFNN) on the Twitter feeds and DJIA values from the period June 2009 to December 2009. We also implement a naive protfolio management strategy based on our predicted values. Our work is based on Bollen et al’s famous paper which predicted the same with 87% accuracy.

Table of Contents

CHAPTER 1

COMPANY PROFILE

1. COMPANY PROFILE

A Brief History of Varcons Technologies

Varcons Technologies Private Limited is a Private incorporated on 11 July 2022. It is classified as Non-govt company and is registered at Registrar of Companies, Bangalore. Its authorized share capital is Rs. 1,000,000 and its paid up capital is Rs. 10,000. It is inolved in Other computer related activities [for example maintenance of websites of other firms/ creation of multimedia presentations for other firms etc.]

Varcons Technologies Private Limited's Annual General Meeting (AGM) was last held on N/A and as per records from Ministry of Corporate Affairs (MCA), its balance sheet was last filed on N/A.

Directors of Varcons Technologies Private Limited are Haralahalli Chandraiah Spoorthi and Chikaegowdanadoddi Kariyappa Somalatha.

CHAPTER 2

ABOUT THE COMPANY

2. ABOUT THE COMPANY

Varcons Technologies is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Compsoft Technologies specialize in ERP, Connectivity, SEO Services, Conference Management, effective webpromotion and tailor-made software products, designing solutions best suiting clients requirements. The organization where they have a right mix of professionals as a stakeholders to help us serve our clients with best of our capability and with at par industry standards.They have young, enthusiastic, passionate and creative Professionals to develop technologicalinnovations in the field of Mobile technologies, Web applications as well as Business and Enterprise solution. Motto of our organization is to "Collaborate with our clients to provide them with best Technological solution hence creating Good Present and Better Future for our client which will bring a cascading a positive effect in their business shape as well". Providing a Complete suite of technical solutions is not just our tag line, it is Our Vision for Our Clients and for Us, We strive hard to achieve it.

Products of Varcons Technologies.

Android Apps

It is the process by which new applications are created for devices running the Android operating system. Applications are usually developed in Java (and/or Kotlin; or other such option) programming language using the Android software development kit (SDK), but other development environments are also available, some such as Kotlin support the exact same Android APIs (and bytecode), while others such as Go have restricted API access.

The Android software development kit includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on QEMU, documentation, sample code, and zutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows 7 or later. As of March 2015, the SDK is not available on Android itself, but softwaredevelopment is possible by using specialized Android applications.

Web Application

It is a client–server computer program in which the client (including the user interface and client- side logic) runs in a web browser. Common web applications include web mail, online retail sales, online auctions, wikis, instant messaging services and many other functions. web applications use web documents written in a standard format such as HTML and JavaScript,which are supported by a variety of web browsers. Web applications can be considered as a specifific variant of client–server software where the client software is downloaded to the client machine when visiting the relevant web page, using standard procedures such as HTTP. The Client web software updates may happen each time the web page is visited. During the session, the web browser interprets and displays the pages, and acts as the universal client for any web application. The use of web application frameworks can often reduce the number of errors in a program, both by making the code simpler, and by allowing one team to concentrate on the framework while another focuses on a specifified use case. In applications which are exposed to constant hacking attempts on the Internet, security-related problems can be caused by errors in the program.

Frameworks can also promote the use of best practices such as GET after POST. There are some who view a web application as a two-tier architecture. This can be a “smart” client that performs all the work and queries a “dumb” server, or a “dumb” client that relies on a “smart” server. The client would handle the presentation tier, the server would have the database (storage tier), and the business logic (application tier) would be on one of them or on both. While this increases the scalability of the applications and separates the display and the database, it still doesn"t allow for true specialization of layers, so most applications will outgrow this model. An emerging strategy for application software companies is to provide web access to software previously distributed as local applications. Depending on the type of application, it may require the development of an entirely different browser-based interface, or merely adapting an existing application to use different presentation technology. These programs allow the user to pay a monthly or yearly fee for use of a software application without having to install it on a local hard drive. A company which follows this strategy is known as an application service provider (ASP), and ASPs are currently receiving much attention in the software industry.

Security breaches on these kinds of applications are a major concern because it can involve both enterprise information and private customer data. Protecting these assets is an important part of any web application and there are some key operational areas that must be included in the development process. This includes processes for authentication, authorization, asset handling, input, and logging and auditing. Building security into the applications from the beginning can be more effective and less disruptive in the long run.

Web design

It is encompasses many different skills and disciplines in the production and maintenance of websites. The different areas of web design include web graphic design; interface design; authoring, including standardized code and proprietary software; user experience design; and search engine optimization. The term web design is normally used to describe the design process relating to the front-end (client side) design of a website including writing mark up. Web design partially overlaps web engineering in the broader scope of web development. Web designers are expected to have an awareness of usability and if their role involves creating mark up then they are also expected to be up to date with web accessibility guidelines. Web design partially overlaps web engineering in the broader scope of web development.

Departments and services offered

Varcons Technologies plays an essential role as an institute, the level of education, development of student's skills are based on their trainers. If you do not have a good mentor then you may lag in many things from others and that is why we at Compsoft Technologies gives you the facility of skilled employees so that you do not feel unsecured about the academics. Personality development and academic status are some of those things which lie on mentor's hands. If you are trained well then you can do well in your future and knowing its importance of Compsoft Technologies always tries to give you the best.

They have a great team of skilled mentors who are always ready to direct their trainees in the best possible way they can and to ensure the skills of mentors we held many skill development programs as well so that each and every mentor can develop their own skills with the demands of the companies so that they can prepare a complete packaged trainee.

Services provided by Varcons Technologies.

- Core Java and Advanced Java
- Web services and development
- Dot Net Framework
- Python
- Selenium Testing
- Conference / Event Management Service
- Academic Project Guidance
- On The Job Training
- Software Training

CHAPTER 3

INTRODUCTION

3. INTRODUCTION

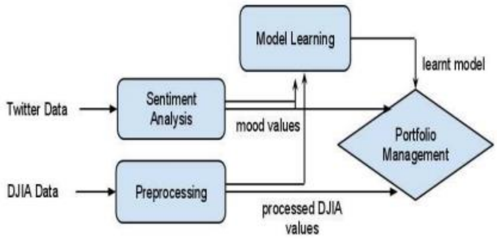
Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli. In this paper, we test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decison making process, thus, leading to a direct correlation between ”public sentiment” and ”market sentiment”. We perform sentiment analysis on publicly available Twitter data to find the public mood and the degree of membership into 4 classes - Calm, Happy, Alert and Kind (somewhat like fuzzy membership). We use these moods and previous days’ Dow Jones Industrial Average (DJIA) values to predict future stock movements and then use the predicted values in our portfolio management strategy. Related work Our work is based on Bollen et al’s strategy [1] which received widespread media coverage recently. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. The authors considered the tweet data of all twitter users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy. They cross validated the resulting mood time series by comparing its ability to detect the public’s response to the presidential elections and Thanks giving day in 2008. They also used causality analysis to investigate the hypothesis that public mood states, as measured by the Opinion Finder and GPOMS mood time series, are predictive of changes in DJIA closing values. The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values. Their results show a remarkable accuracy of nearly 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA). The rest of the paper is organized as follows. The second section briefly discusses our general approach towards solving the problem and the following sections discuss the individual components in greater detail. In Section 3, we briefly discuss the dataset that we have used for this paper and data preprocessing measures adopted. Section 4 discusses the sentiment analysis technique developed by us for the purpose of this paper. Section 5 includes in detail, the different machine learning techniques to predict DJIA values using our sentiment analysis results and presents our findings. In Section 6, we use the predicted values and devise a naive strategy to maintain a profitable portfolio.

CHAPTER 4

SYSTEM ANALYSIS

4. SYSTEM ANALYSIS

Algorithm:The technique used in this paper builds directly on the one used by Bollen et al. [1]. The raw DJIA values are first fed into the preprocessor to obtain the processed values. At the same time, the tweets are fed to the sentiment analysis algorithm which outputs mood values for the four mood classes for each day. These moods and the processed DJIA values are then fed to our model learning framework which uses SOFNN to learn a model to predict future DJIA values using them. The learnt model as well as the previous DJIA and mood values are used by the portfolio management system which runs the model to predict the future value and uses the predicted values to make appropriate buy/sell decisions. Figure 1 shows a brief flow diagram of our technique. The following sections discuss each component of our technique in greater detai DATASET In this project, we used two main datasets- • Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009. The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.



Publicly available Twitter data containing more than 476 million tweets corresponding to more than 17 million users from June 2009 to December 2009. The data includes the timestamp, username and tweet text for every tweet during that period. Since we perform our prediction and analysis on a daily basis, we split the tweets by days using the timestamp information.

Data Preprocessing

The data obtained from the above mentioned sources had to be pre-processed to make it suitable for reliable analysis. We pre-processed the DJIA data in the following manner1. While the Twitter data was available for all days lying in the giving period, the DJIA values obtained using Yahoo! Finance was (understandably) absent for weekends and other holidays when the market is closed. In order to complete this data, we approximated the missing values using a concave function. So, if the DJIA value on a given day is x and the next available data point is y with n days missing in between, we approximate the missing data by estimating the first day after x to be (y+x)/2 and then following the same method recursively till all gaps are filled. This approximation is justified as the stock data usually follows a concave function, unless ofcourse at anomaly points of sudden rise and fall.

2. If we observe the general movement of stock markets, it is associated with a few sudden jumps/falls and a brief period of small fluctuations around the new value. However, such jumps/falls are due to some major aberrations and cannot be predicted. Moreover, as we know the public memory is very short and even though the market may be trading at a much higher level than the previous year, that does not mean that calmness will be much higher than previous year;

public mood is a very local metric. Therefore, we adjusted our stock values by shifting up/down for steep falls/jumps, re daily directional trend (up/down movement of stock prices).

3. Even after shifting the values in step 2, the values contained significant periods of volatile activity which are very difficult to predict. We pruned our dataset by removing these periods for final training and testing. Finally, in order to ensure that values were small and comparable, we computed the z-score of each point in the data series ((x−μ)/σ) and used that in our analysis (The original values were of the order of 104 , so MATLAB was giving a precision error when computing functions like exp(-x2))

CHAPTER 5

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Sentiment analysis was an important part of our solution since the output of this module was used for learning our predictive model. While there has been a lot of research going on in classifying a piece of text as either positive or negative, there has been little work on multi-class classification. In this project, we use four mood classes, namely, Calm, Happy, Alert, and Kind. We tried several standard tools like OpinionFinder, SentiWordnet [5] etc. for our problem but found them inadequate and/or inefficient and therefore decided to develop our own analysis code. The methodology we adopted in finding the public sentiment is as follows1. Word List Generation We develop our own word list based on the well known Profile of Mood States (POMS) questionnaire. POMS is an established psychometric questionnaire which asks a person to rate his/her current mood by answering 65 different questions on a scale of 1 to 5 (For example, rate on a scale of 1 to 5 how tensed you feel today?). These 65 words are then mapped on to 6 standard POMS moods-Tension, Depression, Anger, Vigour, Fatigue and Confusion. In order to do automate this analysis for tweets, the word list needs to be appropriately extended. Bollen et al. [1] used the Google n-grams data for the same. We followed a much simpler approach of extending the list by considering all commonly occuring synonyms of the base 65 words using SentiWordNet and a standard Thesaurus. 2. Tweet Filtering As mentioned earlier, the tweet data is enormous and will take several hours to be processed if used as it is (which makes the task of daily predictions difficult). Therefore, we filtered and considered only those tweets which are more likely to express a feeling, i.e. we consider only those tweets which contain the words ”feel”, ”makes me”, ”I’m” or ”I am” in them. 3. Daily Score Computation We used a simple word counting algorithm to find the score for every POMS word for a given dayscore of a word = #of times the word matches tweets in a day #of total matches of all words The denominator accounts for the fact that the number of tweets could vary from one day to another. This works well for our problem because of the nature of tweets which contain simple sentence structures and only a maximum of 140 characters (in most cases much less). We tried using the Stanford coreNLP software for word tagging and then using a word’s position in the sentence to find its importance. However, similar to our experience working with OpinionFinder, we observed that this process, besides being extremely slow was not too beneficial. 4. Score Mapping We map the score of each word to the six standard POMS states using the mapping techniques specified in the POMS questionnaire. We then map the POMS states to our four mood states using static correlation rules (for example, happy is taken as sum of vigour and negation of depression). It is important to note that, given our formulation, it does not make much sense to compare the value of one mood against another; they should only be used to compare mood trends across days. We cross validated the results of our sentiment analysis technique by comparing the values returned by our algorithm around significant events like Thanksgiving day and Michael Jackson’s death. As shown in Figure 2, the moods show a sharp rise of various mood states on Thanksgiving whereas one day after MJ’s death, there is a sharp decline in happiness

CHAPTER 6

DESIGN ANALYSIS

6. DESIGN & ANALYSIS

Granger causality is based on linear regression, but the correlation between stocks and moods is certainly non linear. Therefore, after finding a causality relation between the past 3 days moods and current day stock prices, we tried 4 different learning algorithms (Linear Regression, Logistic Regression, SVMs, Self Organizing Fuzzy Neural Networks) to learn and study the actual correlation. For SVM we used the LIBSVM [2] library, but we implemented the other three in MATLAB ourselves as we could not find good working libraries for them. The Self Organizing Fuzzy Neural Network (SOFNN) is a five layer fuzzy neural network which uses ellipsoidal basis function (EBF) neurons consisting of a center vector and a width vector. We implemented the online alorithm for creating SOFNNs as introduced in [3] in which neurons are added or pruned from the existing network as new samples arrive. Neural networks have been considered to be a very effective learning algorithm for decoding nonlinear time series data [4], and financial markets often follow nonlinear trends. The authors in [1] showed the 87% correlation using SOFNNs only, and our results also indicate that SOFFNs do the best among all other algorithms, giving nearly 75.56% accuracy. In order to measure accuracy, we developed a novel validation technique called the k-fold sequential cross validation (k-SCV). In this method, we train on all days upto a specific day and test for the next k days. The direct k-fold cross validation method is not applicable in this context as the stock data is actually a time series unlike other scenarios where the data is available as a set. Therefore, it is meaningless to analyze past stock data after training on future values. For the purpose of our analysis, we use k = 5. Our Granger Causality analysis indicates that Calm and Happy are causative of the DJIA values. But to confirm the inverse dependence of other mood dimensions on DJIA we investigated a total of 7 different possibilities. In Table 2, ID denotes the 5-SCV accuracy when only the past 3 days DJIA values are given as features. Similarly ICD, ICHD, ICAD, ICKD, ICHKD, ICHAD denote the accuracy when features are the past 3 days DJIA values (represented by D) along with the past 3 days mood values (C=Calm, H=Happy, A=Alert, K=Kind) in different combinations. MAPE indicates the Mean Absolute Percentage Error between our predicted values and the actual normalized stock values. The Direction accuracy indicates the percentage matchings in the trends (up/down) predicted by our training vs the actual daily trends in the stocks. When using classification algorithms, we used the up/down trends as class inputs and used the algorithm to directly predict trends whereas when using regression algorithms, we fed the normalized stock values as input and used the predicted stock values to obtain direction (up/down) trends. We find that ICHD gives the best results in all the algorithms considered, indicating that Calmness and Happiness are more predictive of the stock values, confirming the Granger causality analysis and unlike the [1] result. They showed 87% correlation when features were Calm and DJIA values of past 3 days, and reported the accuracy only on a specific test set, without reporting any cross validation error average. Our results are also in conjunction with the philosophy that happiness should in general be causative of the stock values. Figure 4 shows a graph of the normalized stock values as predicted by our SOFNN algorithm vs the actual normalized stock values when trained on the Happy+Calm+DJIA feature set. We find a very close correlation and hence the small value of Mean Absolute Percentage Error (MAPE) and good Directional Accuracy. We can draw several important conclusions from the Table Figure 4: Predicted vs Actual Stock Values using SOFNN on Calm+Happy+DJIA for 40 Consecutive Days 2. Firstly SVMs and Logistic Regression perform badly on this dataset, giving the same percentage values for Direction

Accuracy for all mood combinations. This shows that classification (directly predicting trends) is not the ideal methodology for this problem. Linear Regression performs pretty good, which is in conjunction with the Granger Causality results, whereas SOFNN performs the best. If we look at the Direction Accuracy for SOFNN we observe that the best value is for ICHD (75.56%). Also on adding any other mood dimension the Direction Accuracy worsens, ie. the Direction Accuracy is clearly lower for ICHAD, ICHKD. This shows that by adding more features we would essentially be overfitting the data. If we try to remove some feature and find the Direction Accuracy, we observe that the result still worsens. Hence Calm and Happiness are indeed more indicative of the stock values than any other moods. If we consider the MAPE values, we find that Calm and Alert mood dimensions do marginally better than the others, but they are poor in predicting the Direction or the trends of stock movement. To compare the significance of our results on the Directional Accuracy with those of [1], we see that 5 fold sequential cross validation is clearly a better indicator. [1] have used a specific test period which gives a 87.6% Directional Accuracy, and they have proved that statistically the probability of that event happening with random success is pretty low. But we have on the other hand cross validated over the entire period, using our 5-fold SCV technique. As mentioned earlier, this technique makes more sense for stock data because the usual cross validation would essentially be using future stock values to predict the past ones, which is an incorrect technique when financial data is concerned.

PORTFOLIO MANAGEMENT

Having predicted the DJIA closing values one day in advance, we can use these predicted values to make intelligent sell/buy decisions. We develop a naive greedy strategy based on a simple assumption that we can hold at most one stock at any given time (or s stocks if all stocks are always bought and sold together) Following are the steps/features of our strategy- • Pre-computation We maintain a running average and standard deviation of actual adjusted stock values of previous k days • Buy Decision If the predicted stock value for the next day is n standard deviations less than the mean, we buy the stock else we wait. • Sell Decision If the predicted stock value is m standard deviations more than the actual adjusted value at buy time, we sell the stock else we hold. Note that the above strategy has three parameters- k, n and m. Our experiments show that the optimal parametrization is n = m = 1 and k = 7 or 15 (Note that in order to trade effectively, we needed a large enough test set, containing at least 30-40 entries, thereby limiting the scope of our experiments because of limited test data.) The profit obtained using our strategy is as followsk = 7 - P rof it = 527.2 Dow Points k = 15 - P rof it = 543.65 Dow Points The total range of stock movement in the same period is 920.72 Dow Points. Please note that while the above analysis is entirely in terms of Dow points, it is easy to correlate a profit in Dow points with a monetary value. For example, the Dow Diamonds is an exchange-traded fund that holds the 30 stocks that comprise the DJIA and researchers have shown a 98.5% correlation between the movement of DJIA and the Dow Diamonds. Similarly, we can find many other funds which can translate this profit in Down point to a corresponding profit in Dollars. 7. CONCLUSIONS AND FUTURE WORK

CHAPTER 7

IMPLEMENTATION

7. IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively.

The system can be implemented only after thorough testing is done and if it is found to work according to the specification. It involves careful planning, investigation of the current system and it constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning.

Two major tasks of preparing the implementation are education and training of the users and testing of the system. The more complex the system being implemented, the more involved will be the system analysis and design effort required just for implementation.

The implementation phase comprises of several activities. The required hardware and software acquisition is carried out. The system may require some software to be developed. For this, programs are written and tested. The user then changes over to his new fully tested system and the old system is discontinued.

TESTING

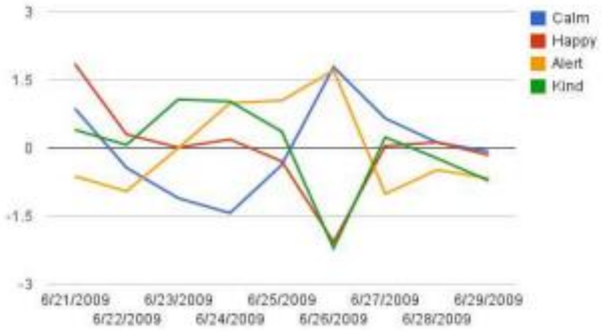
The testing phase is an important part of software development. It is the Information zed system will help in automate process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied. Software testing is carried out in three steps:

1. The first includes unit testing, where in each module is tested to provide its correctness, validity and also determine any missing operations and to verify whether theobjectives have been met. Errors are noted down and corrected immediately.
2. Unit testing is the important and major part of the project. So errors are rectified easily in particular module and program clarity is increased. In this project entire system is divided into several modules and is developed individually. So unit testing is conducted to individual modules.
3. The second step includes Integration testing. It need not be the case, the software whose modules when run individually and showing perfect results, will also show perfect results when run as a whole.

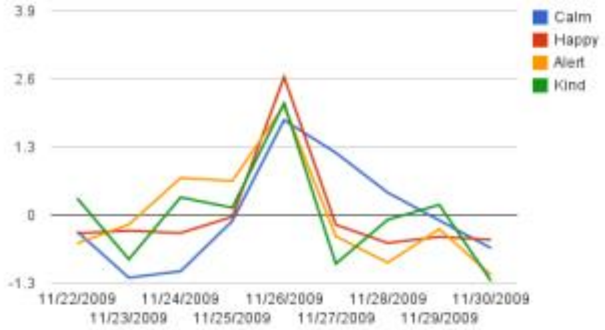
CHAPTER 8

SNAPSHOTS

8. SNAPSHOTS



(a) Various moods after Michael Jackson's death on 25 June 2009



(b) Various moods on Thanksgiving day on 26 November 2009

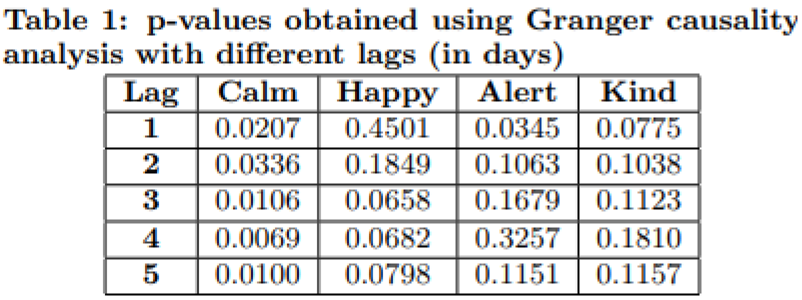
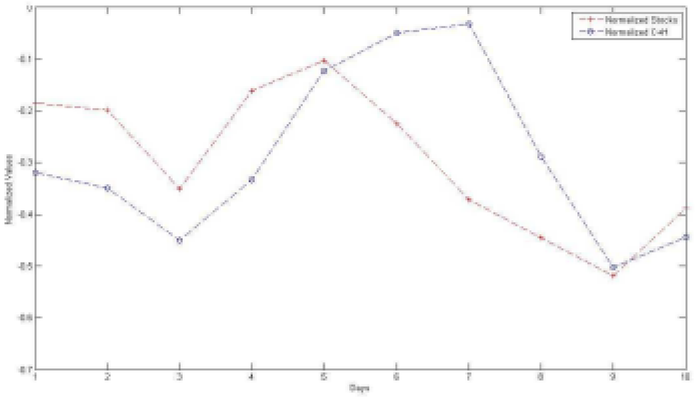


Table 1: p-values obtained using Granger causality analysis with different lags (in days)

Lag	Calm	Happy	Alert	Kind
1	0.0207	0.4501	0.0345	0.0775
2	0.0336	0.1849	0.1063	0.1038
3	0.0106	0.0658	0.1679	0.1123
4	0.0069	0.0682	0.3257	0.1810
5	0.0100	0.0798	0.1151	0.1157



The DJIA, and Calmness + Happiness curves superimposed to show correlation

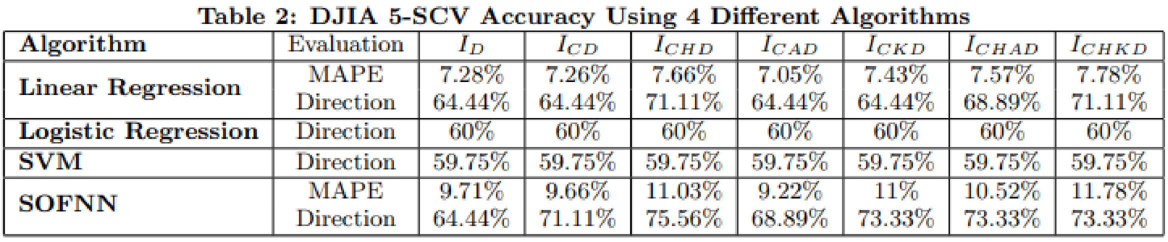
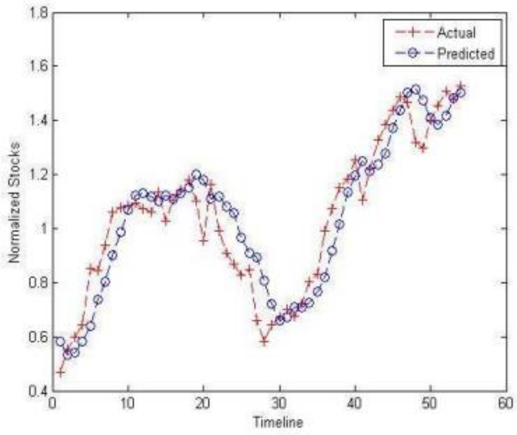


Table 2: DJIA 5-SCV Accuracy Using 4 Different Algorithms

Algorithm	Evaluation	I_D	I_{CD}	I_{CHD}	I_{CAD}	I_{CKD}	I_{CHAD}	I_{CHKD}
Linear Regression	MAPE Direction	7.28% 64.44%	7.26% 64.44%	7.66% 71.11%	7.05% 64.44%	7.43% 64.44%	7.57% 68.89%	7.78% 71.11%
Logistic Regression	Direction	60%	60%	60%	60%	60%	60%	60%
SVM	Direction	59.75%	59.75%	59.75%	59.75%	59.75%	59.75%	59.75%
SOFNN	MAPE Direction	9.71% 64.44%	9.66% 71.11%	11.03% 75.56%	9.22% 68.89%	11% 73.33%	10.52% 73.33%	11.78% 73.33%



Predicted vs Actual Stock Values using SOFNN on Calm+Happy+DJIA for 40 Consecutive days

CHAPTER 9

CONCLUSION

9. CONCLUSION

We have investigated the causative relation between public mood as measured from a large scale collection of tweets from twitter.com and the DJIA values. Our results show that firstly public mood can indeed be captured from the large-scale Twitter feeds by means of simple natural language processing techniques, as indicated by the responses towards a variety of socio-cultural events during the year 2009. Secondly, among the observed dimensions of moods, only calmness and happiness are Granger causative of the DJIA by 3-4 days. Thirdly, a Self Organizing Fuzzy Neural Network performs very good in predicting the actual DJIA values when trained on the feature set consisting of the DJIA values, Calm mood values and Happiness dimension over the past 3 days. The performance measure we have used is kfold sequential cross validation, which is more indicative of the market movements for financial data. Finally a naive implementation of portfolio management using our strategy indicates a decent profit over a range of 40 days. Our results are in some conjunction with [1], but there are some major differences as well. Firstly our results show a better correlation between the calm and happy mood dimensions with the DJIA values, unlike their result, which showed high correlation with only calm mood dimension. Secondly, we haven’t been able to obtain high percentage result of 87%, but our 75.56% result using k-fold sequential cross validation gives stronger evidence that the correlation is over the entire range of data. The profits in out naive implementation of portfolio management shows that our MAPE estimates are pretty much accurate. Finally, its worth mentioning that our analysis doesn’t take into account many factors. Firstly, our dataset doesn’t really map the real public sentiment, it only considers the twitter using, english speaking people. It’s possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people’s mood indeed affect their investment decisions, hence the correlation. But in that case, there’s no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research.

10. REFERENCE

[1] J. Bollen and H. Mao. Twitter mood as a stock market predictor. IEEE Computer, 44(10):91–94.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[3] G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. Neural Networks, 17(10):1477–1493.

[4] A. Lapedes and R. Farber. Nonlinear signal processing using neural network: Prediction and system modeling. In Los Alamos National Lab Technical Report.

[5] A. E. Stefano Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC