A Decision Support Approach for Online stock forum Sentiment Analysis

Abstract

Prediction and future forecasting is always been an intangible object. Financial analyst and risk analyst works out to analyse the market risk and suggests decisive steps. Stock market is one of the investment covered under risky investments. The risk involved in stock market cannot be mitigated by making use of any single statistical algorithm. Surveys have found that particularly in domain of stock market opinions from experienced investors contributes much better than solely machine driven approach. To analyse such opinions manually would be impossible task. Thus, approach proposed in this artefact is to make use of sentiment analysis and process all such opinions from experienced investors and map it with a time line and companies. Further, such sentiment scores are feed to the machine driven decision system, which makes used of GARCH – SVM model approach to further process this score and statically produce a predicted value for stocks. Performance is evaluated for this combined approach and artefact is also tested with Base approach for better evaluation. This report concluded that proposed artefact have performance accuracy gain of 20% than the base approach and accuracy decays exponentially when duration of data feed and prediction date increases.

Table of Contents

List of Figures

List of Tables

Introduction

Stock forum is a market considered under risky investment which attracts people from world-wide. As mentioned, it is a risky investment, many analysts work to mitigate this risk in market. Financial analysts makes use of few statistical derivations and come up with some conclusion in predicting the future value of a stock. The statistics and algorithm driven results may not prove to be completely successful as at times, market may become completely unpredictable. Thus, apart from statistical driven algorithm, sentiments plays vital role. Sentiment comes from comments of experienced people world-wide who have good understanding and experience about stock market and investment. The value of experience cannot be driven by any algorithm, as it is an intuition. In many cases it has been found that experienced prediction serves much closer to reality than algorithm driven prediction in domain of stock market. Thus, this project intends to make use of sentiment analysis from online stock forums and aid towards mitigating the risk in stock investment. Thus, it proposes to develop a tool that fetches comments and blogs from online stock forums, analyse the sentiments, maps it with the firm for which sentiment is provided and maps with the timeline during which sentiment was given. Dual mapping of sentiment is required to be done in case of stock market. As it varies from company to company and from time to time. Thus, first mapping must be done for time making use of time vectors. This can be obtained by fetching the time in blogs. Moreover, blog posted timestamp can also prove to be useful in such case. This helps in creating a time series vectors, further can be partitioned in month wise and year wise. Another mapping required would be for company name. It is very important to map sentiments with company, otherwise sentiment score would be useless. Based on specific keywords, sentiments can be mapped with respect to companies and time line. This two dimensional mapping shall serve as input to sentiment analysis. Sentiment analysis is expected to provide appropriate score for sentiment and same will be reflected in mapping.

Based on this sentiments and its score, predictive future based on firms and stocks will be listed and a decision supporting algorithm will be proposed. Thus, on basis of such inputs and processing, a future value for particular stock can be predicted and conclusion will be deriving on percentage of risk reduction in stock market and promote more investors to invest on stocks. This will ultimately increase business by attracting more investor for investing and shall help in financial analysis and decisions for all concerned people. However, there is no such guarantee about the success of the proposed system, and thus a testing approach is planned in such a manner that it evaluates the performance. Basic approach is to estimate future value of portion of shares that were actually in past. Thus, output of system and actual past (pseudo future) value of share can be compared and thus performance of the system will be shown. It should be noted that if the efficiency of the system is at least 40%, the system must be considered as successfully in predicting the future value for stock market. 40% is considered as minimum success percentage because as fore mentioned, stock market is one of the most risky investment scheme, and thus predicting future value correctly even up to 40% can at least conclude that particular investment would be a profit or loss.

Aim

Aim of this project is to make use of millions of online forums where people comments about stocks and predict the future value for particular share in form of sentiment and provide their opinions and to analyse such opinions, sentiment analysis shall be used and appropriate scoring would be provided for each sentiments. Furthermore, a dual mapping as aforementioned would be done for time series and company keywords. Thus, a reference would be created in which for particular time and particular company, average of sentiment analysis scores will be mentioned. All such scores will serve as input to the decision support system and based on the nature of the output appropriate algorithm will be developed that predicts the future value of the stock for particular time and company.

Thus, this project intends to develop a web based tool that fetches the sentiments from various stock forums, identifies historic comments and its correlation with reality. Based on this sentiments and its score, predictive future based on firms and stocks will be listed and a decision supporting algorithm will be proposed. Expected outputs of this research is to reduce the risk involvement in stock investment and promote more investors to invest on stocks and take business related decisions, with statistically supported information from worldwide available sentiments.

Objectives

1. Collecting sentiments / opinions related to stocks from online forums
2. Process sentiment and extract the time duration for each sentiments
3. Process sentiments and extract company name for each sentiments
4. Create mapping between sentiments, company name and time series
5. Evaluate sentiments and associate score to each sentiments
6. Partition time series for each month, for every year and produce averaged sentiment score and map it with time series and company name
7. Analyse sentiment scores, along individual company name and time vector using GARCH – SVH algorithm
8. Correlated results with reality for historical sentiments
9. Quality analysis of predicted time series value using pseudo prediction and actual prediction
10. Comparing results with existing tools
11. Reduction in risk calculation

Report Summary

Chapter 1 contains the introduction, which subjects the problem of risk in stock market investment and proposes for a sentiment analysis and decision supporting system based

solution to mitigate the risk. It further includes the aim of the research and objectives as milestones to fulfil the proposal.

Chapter 2 contains the literature survey. Literature survey is divided in to three sections. First section refers to survey on English and Non – English literature. Second Section refers to survey on sentiment analysis its use and trend, challenges with sentiment analysis, exiting practices in sentiment analysis and approaches in sentiment analysis for known problems. Particular survey is regarding sentiment analysis and it is auto correlated. Third section refers to survey for financial domain and its correlation with sentiment analysis. This survey includes cross correlation of sentiment analysis with financial domain and particularly for stock market. It also includes challenges particularly in financial sentiment analysis. Based on literature survey, challenges with respect to the proposal were identified and exact scope of the research is defined.

Chapter 3 contains the technical and functional design of the artefact. It includes functional flow of the system, process flow of the system and how classification of the sentiments is done. From technical point of view, how exactly the coding part is done is also represented in this section. It includes description of all major classes and packages that are used in the artefact. Moreover, for better understanding, artefact is represented in multiple point of views using UML diagrams.

Chapter 4 contains the testing approach. Testing approach is divided into two sections, one is for UAT and other is functional test approach. Using this test approaches, performance of the developed artefact is evaluated. Same performance is plotted in different charts for better analysis.

Chapter 5 contains the conclusion that derived from the results plotted in chart from chapter 4. It also concluded that proposed approach in developed artefact is 20% better than the base approach.

Chapter 6 contains the reference that are used for developing the artefact and analysing the problems and requirements.

Literature Survey

As the research is integrated with multiple domains, literature survey is required to be done in individual domains and also for their correlation to achieve the ultimate aims and objectives. Major domains that require literature survey is the language and lexicon, sentiment analysis and financial and future value analysis.

Role of Languages and Lexicon

Study and research of languages is not directly correlated with proposed system or research, however it serves as base input to the proposed system. Thus, literature survey in lingual domain is very vital. Survey have shown that worldwide people show interest in investing in stock market. However, surveys have shown that local regional people pertaining to that country show almost 95% of interest in frequent buying and selling, however other 5% belongs to different region (Ivković & Weisbenner, 2005). Moreover, many companies have worldwide listing. As mentioned that listing of a company can be worldwide, similarly comment and opinions of people for these stocks also comes in different languages and assents. According to survey, almost 60% of people worldwide prefer to post in their own language or at least they have few words in sentence that comes from their native language apart from English (Ivković & Weisbenner, 2007). Thus, a further survey was done on how to deal with non-English sentiments and how they affect performance of sentiment analysis, and shall ultimately affect performance of proposed system (Denecke, 2008). Further survey divides for challenges in sentiment analysis in correlation with lingual domain.

Sentiments in English

In a case, it can be a fair assumption that most of the sentiment would be in English language. However, surveys have shown that when people comment, tweet or post a blog on forums, they never know that their opinions may be further used my machine learning process. Thus,

each and every comments are not in same format and pattern (Taboada, Brooke, et. al., 2011). For example, particularly for machine learning process, if comments comes in uniform manner like “XYZ Company, in year 2015 will hit billion.” In such case, machine learning approach needs to first Get Company’s name, until word “Company” starts and fetch year, after word “year” end and rest must be considered for scoring of sentence, i.e., positive, negative, etc. However, this is not the case in actual posts. People keep on posting in their manner, which makes it difficult to debug the sentence and fetch meaningful information using machine learning. As per the survey, approaches used so far includes use of master record for company list, which matches all words in sentence with company list lexicon (Taboada, Brooke, et. al., 2011). Year is identified by check consecutive 4 letter number with starting range of 19 – 20 and ending range of 00 – 99. For rest of the words, SentiWordNet lexicon is used to score the sentence as positive, negative or neutral (Denecke, 2008).

Challenges in English posts

However, English sentiments itself have multi challenges in sentiment analysis. Human comments are prone to spelling mistakes. Ultimately sentiments are just a post from people, and not a serious document that people review language before posting. Thus, even for English opinions, spelling mistakes lead to converse perception in machine learning approach. However, in such case word prediction algorithm can be used which is widely used in office tools (Even though its efficiency is only 90% as per the survey). It should be noted that sentiment score prediction errors for English language would be included in discrete literature survey (Hu, X., Tang, et. al., 2013). This, section of literature survey is limited to lingual challenges on its own, and not in complete correlation to sentiment Sentiments analysis.

Challenges in Non- English posts

As aforementioned, even non-English opinions cover almost 20% of total online stock forums, such scenarios must also be considered. Most common approach that is used, it to make use of language converter tools (Wan, 2009). First step includes detection of language, next step includes conversion of language from detected language to English language, and once opinion is in English language, as per previous section, same is applicable (Aiken & Balan, 2011). But, it should be noted that, if English opinions itself have 10% of errors due to wrong prediction and grammatical errors, then for non – English opinions, along with 10% of error, 15% of conversation error will also be applicable. However, there is no direct sum of errors in percentage, but both of this errors have a weightage ratio and cumulatively add up for resultant errors for non – English opinions particularly for lingual domain (Balk, Chung, et. al., 2012). It can be represented in a manner that Xa + Yb = Z, where X = percentage of known error that exists in English opinions due to spelling and grammar mistakes, Y =percentage of known conversion error for conversing non – English onions in to English, a and b are the weightage coefficient for each that shows cross correlation and Z = percentage of total error due to non – English opinions (Balahur & Turchi, 2014).

Trends in Sentiment Analysis

As per the survey, it was found that trends of sentiment analysis started showing popularity in recent decades in multiple domain of business analysis whether related to information technology or not. Most popular use of sentiment analysis is to find the customer satisfaction ratio (Gamon, 2004). Most of the companies makes use of sentiment analysis to fetch the comments related to their products and based on that, they evaluate the customer satisfaction ratio. Not only that, making use of such sentiments, it helps them to solve the issues and problems faced by customers for their product and mitigate them in next release or patch. Thus, sentiment analysis have become major part of business regardless of the domain of

Business. Moreover, such sentiment analysis can also help a company to analyse needs and problems of customers and can serve as roots to idea for developing a new product based on requirements or problems of people (Gamon, 2004). Sentiment analysis is not only limited to service or product providers, but it also helps customers to choose correct product based on opinions of people worldwide. Sentiment analysis makes use of SentiWordNet (Denecke, 2008), which is a lexicon and uses a machine learning method to analyse such opinions. However, a little work is found in domain of sentiment analysis for Non-English sentiments or opinions. Moreover, sentiment analysis have found to be having poor performance over sarcastic opinions. This problems conveyed that a detailed survey is required particularly for sentiment analysis and its trends, challenges and approaches before moving ahead with financial domain. Moreover, it concludes that for preliminary evaluation of the proposed system performance, it is necessary for a human who can actually understand the comments and provide scores to the words, so that errors due to sentiment analysis can be found. It should be noted that this error shall again have correlation with error mentioned in previous section.

Challenges to Sentiment Analysis

A survey was particularly taken to address the challenges in sentiment analysis. It is quite important because, while calculating the performance of the proposed system, the problems of sentiment analysis itself must also be considered. Moreover, it problems with sentiment analysis is addressed, for proposed system it can help to avoid such situations (Pang & Lee, 2008). Major challenge in the sentiment analysis itself as found in survey is the sarcasm and implicit opinions.

Sarcasm in Sentiment Analysis

Sarcasm means use of irony, where speaker conveys meaning of sentence in completely opposite way (Agarwal, Xie, et. al., 2011). Sometime, it becomes difficult even for human to

understand that. Thus, it becomes major hurdle for machine learning approach based tools like sentiment analysis. Just for an example with respect to sarcasm opinions, consider "How can anyone watch this movie?" Here the state of mind of speaker is very important. Sentiment analysis cannot consider it as negative sentence, as there is no straight forwards negative key work like Bad, worst, etc. But the sentence was intended to be said in negative manner. Another example related to stock could be "How can someone invest in this stock?" Here same situation arises, as there is no keyword which matches with lexicon of SentiWordNet. Problem becomes worse when negativity is conveyed in positive manner. For example, "This movie is too good that I had to sleep in theatre hall. Everybody who is facing insomnia must watch this movie." Here at first instance, sentence seems to be having positive meaning and same would be the result of sentiment analysis tool. But, while thorough revision, it can be seen that sentence is ironically mentioned by speaker which conveys that movie is boring. However, there is no solution (Agarwal, Xie, et. al., 2011) to sarcastic sentiments for sentiment analysis and use of such opinions can be considered as threat to the proposed system.

Domain dependency in Sentiment Analysis

Another major challenge that occur in sentiment analysis is the domain dependency. In most of the social blogs, it can be found that the opinions and comments in most of the cases are with respect to a parent problem or question (Ahmad, 2011). For instance, a person starts up a blog about asking a review of some product or itself commenting for a product. Later on, other people who have experienced same or know about it, shows interest in it and add their opinions below that. Every blog sites have a page limit with maximum number of blogs that can be accommodated in single page. And newer blogs keeps on appending them and old blogs goes to older pages. Thus, many times, a single opinion might be in context to an opinion by some other person's opinion in previous pages. However, considering this point,

blogs have introduced to carry forward the old blog chain for which opinion is given, but that is optional and not necessary that every bloggers follow it. Thus, in sentiment analysis it is very important to define the scope on how many old blogs must be taken to evaluate particular opinion, because an opinion could be like, “Yes that book is good too”. Here, it is completely unknown that which book speaker is talking about. Moreover, even if speaker have addressed previous blog and with respect to that, speaker must have told that, book is good. But for sentiment analysis, it becomes very difficult to identify and link such chained messaged (Mukherjee & Bhattacharyya, 2013). Situation becomes worst when speaker says something like, “Peter, your book is good and John, your book is waste of time”. Here, sentiment analysis, cannot judge that Peter and John are part of this conversation or an external people being addressed. Even if they are part of particular conversation, it makes requirement to identify first, which book was proposed by Peter and John. To solve this problem, most recent approach used by social networking sites and blogsites is to allow user to link user to whom writer is addressing. For example, Peter and John have already posted before, thus they are existing users of the blogsite. While new writer mentioning Peter in blog, it provides option to refer multiple Peters from available users. This helps in letting Peter notified that he’s blog is replied and also helps in domain dependency by relating people within the blogsite. One more example, “Go, read this book to know.” Here, it cannot be judge that by what domain the speaker said this sentence, may be in positive manner or negative manner. However, this again is the limitation of sentiment analysis (Mukherjee & Bhattacharyya, 2013). Solution to domain dependency in sentiment analysis is to provide a limited chained opinions, and manually filtered by data entry operator. Without such manual filtering, it would be impossible to define the exact scope of historical dependency. Moreover, in many cases, it has been found that same blog remains live for blogging even after years. For example, opinions for a movie may have started by people posting on it since

2000 and people still post their opinions for same in recent times. Thus, even if a chained message is created for entire domain dependency, all opinions from year 2000 till date would be impossible to map with each other and analyse.

Thwarted Expectations in Sentiment Analysis

Sometimes, author or speaker makes use of thwarted expectations, which shows positive orientation of the sentence and in the end, author itself refutes in the end sentence (Mukherjee & Bhattacharyya, 2013). For example, "This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up." Here from all linked sentences, it can be conveyed that film is good. However, by the last sentence of the author, it conveys completely opposite meaning that movie is ultimately not good. In traditional sentiment analysis, term frequency was considered. That means, for such linked sentence, total count of positive key words and total count of negative keywords were calculated and average of both counts would be the winning score of the sentence. Here, positive keywords are more in number, than the negative keyword, but still in real sense, it conveys that movie is bad. Thus, recent approach includes term presence along with term frequency.

Pragmatics in Sentiment Analysis

Pragmatics refers to hidden meaning of the sentence. In other words, entire situation may not be aware. In context to domain dependency, even situation knowledge is also required (Mukherjee & Bhattacharyya, 2013). Thus it becomes difficult sometimes to understand the perception of the sentence even by human. For example, "You have green light". Here, meaning of sentence cannot be judged as positive or negative, as sentence perception itself is mystery. Here, it could mean that "You have actually green coloured light", or "You are good to go i.e. proceed further with something" and many more. Thus, sentiment analysis cannot provide any score for such sentence, or if it could score, it would be a false score. As solution

to such situation, a manual separation of sentiments is required. Moreover, if sentiments are like mystery, even domain dependent opinions may not help every time. Thus, a provision must be given, where a human who can actually interpret the opinions, can alter them in a manner that machine learning approach would leave correct score for same. For example, in case as mentioned above, same sentiment can be changed to "It is good" if it is not completely mystery or it can be removed if it is difficult for human to interpret is accurately.

Worlds Knowledge in Sentiment Analysis

In most of the sentence, speakers refers directly to x and y objects, and shows comparison between them. First step for machine learning would be to identify that objects and subject present in it. If either one of subject or object is found to be other than year or company's name, it becomes vital to link it with known object or learn it as new object (Mukherjee & Bhattacharyya, 2013). In many cases, it becomes impossible to track such objects in linked conversations. More often, it may relate with a universally known object rather than one present in conversations. For example, "He is a Frankenstein. Just finished Doctor Zhivago for the first time and all I can say is Russia sucks." Here, sentiment analysis tool must know what Frankenstein is and what Doctor Zhivago is. Without such world knowledge, scoring of sentiment analysis becomes inaccurate (Mukherjee & Bhattacharyya, 2013). The only solution to such problem is to make use of search engine and fetch the meanings for unknown objects and subjects. Moreover, this criteria will be applicable, it is found as necessary in domain of stock market. Reason behind it is, in proposed system, there would be lexicon denoting list of all companies and all possible keywords. Thus, it may be possible that sentiments that does not match object or subject with specified list of keywords, can be filtered out directly without any processing further.

Other challenges in Sentiment Analysis

Many such challenges that were found from the survey are, Subjective detection, Entity identification and Negation. Thus, before moving with sentiment analysis, it is efficient to assume that errors due to above mention problem are inevitable (Mukherjee & Bhattacharyya, 2013). To control such kind of errors are out of scope of this project. Thus, a rough estimation of error percentage can be considered from Rentoumi, Vouros, et. al., (2009) and that can be directly subtracted with obtained native system performance. Thus, in corollary to this assumption, it is understood that only straight forward type of comments and feeds will be considered as input to the sentiment analysis for proposed system that many be filtered manually by human or can even be altered manually by a person.

Practices in Sentiment Analysis

Above section, refers to the challenges in sentiment analysis. Thus, it becomes very important to know about existing practices in sentiment analysis and how they overcome the challenges. As per the survey in general practice in sentiment analysis, OSGOOD semantic differentiation with WordNet is a known practice. Here a lexicon is imported that have keywords for almost all possible words as found in common English dictionary. Finding parts of speech is very important in sentiment analysis. It serves as "crude form of word sense disambiguation" to the sentiment analysis. First step of analysis that is required to be done is finding object and subject. Further verb, adjective and etc. are listed. Object is used for measuring the relative reliability of the sentiment and subject is used to map the sentiment said for, i.e. name of movie or brand or product, etc. In case of stock market, it would represent the name of the firm or company for which the sentiment analysis is to be done. Further to this, subject and object as mapped and based on extracted adjectives (for example, "very good"), score is given. The most used part of speech is the adjective in domain of sentiment analysis. It is the most frequent technique used in practices for sentiment analysis.

Literature shows that 80% of accuracy in detecting the sentiments correctly is found by scoring sentiments based on adjectives. Literature also shows that along with adjective, adverb also plays vital role in efficient scoring of sentiment analysis. Benamara, et al. (2007) showed that adverb can alter the sentiment value that are derived from adjective. Thus, a variable scoring algorithm must be used for sentiment analysis. "Sentiment expressed with regard to a particular subject can best be identified with reference to the subject itself" (Natsukawa and Yi, 2003).

Approach towards challenges

Pang and Lee (2008) provided new approach to face the problems that arise in sentiment analysis and to counterfeit the problems, the approach includes material on summarization for evaluative string and focused on issues regarding manipulation, privacy and economic impact that the development of "opinion-oriented information-access services" gives rise to. A discussion of available resources, benchmark datasets, and evaluation campaigns is also provided in same. It proposes machine learning approaches that considers parts of speech, unigrams, bigrams, combination of unigram and bigram, term presence along with term frequency, position of keyword in sentence, providing higher weightage to adjective, etc. Nasukawa & Yi (2003) provided initial approach to quantize the sentence and score based on quantized parts. This quantize parts collectively forms a complete sentence then can considered for sentence scoring. It is quite understood that more the quantization level, better the performance. However, it does **not** hold true directly while dealing with sentiment analysis. Here it must be used where, a domain understanding is required. Thus, in other words, it cannot be used, where a comment or opinion is correlated with adjacent sentence. This may appear in most of the case, as comments include multiple sentence and collectively all sentence represents a positive or negative opinion for subjected firm. For instance, speaker must have used multiple sentences in collaboration, to define the future financial stability of

the firm. Thus, in sentiment analysis, it is important to go for non-linear quantization rather than linear quantization of sentences. Thus, machine learning decision approach must include in algorithm that can identify the beginning of the blog and have sense to identify all linked chain blogs. Moreover, each such blogs must be at least quantised for all sentence in single opinion. A single opinion can have many sentences. Thus, single sentence quantization must contain entire opinion. Such quantization will help in effective scoring for sentiments and collectively all such linked quantization will help in ultimate scoring for referenced company.

Correlation between Finance and SA

Finance forecasting and volatility is been always a major concern in domain of financial research. Business analyst and financial analyst works out using statistics to identify the key risk elements in business and try to mitigate them by providing situation related solutions. The most common term that relates with finance forecasting is to find the future value (Barberis & Thaler, 2003). However, predicting or forecasting financial statement is not possible, however a prediction can be done that can at least support such analysts to take important decision related to finance. Regarding stock forums, major questions that needs to be answered are, is there any practical correlation with stock forum and real-time price of stocks? If yes, what would be the reliability ratio for such forums? Answers to this question shall lead to applicability of sentiment analysis in domain of stock forum (Melville, Gryc & Lawrence, 2009). To answer these questions, literature survey on historical forums and related work is required. This means that, once a proposed system is developed, sentiment analysis must be done on historical sentiments, up to a past value, say 2010. Based on the derived results, a pseudo prediction must be calculated for year 2013 or 2014. As per the situation, for machine tool, inputs are given for opinions up to 2010 only. Thus, 2013 and 2014 stock values predicted would be the actual prediction for it. However, in real sense, 2013 and 2014 is already a past and its actual values of stocks are known. Thus, comparing

both of this results must help to answer the questions that whether any such sentiment analysis can work in domain of finance forecasting or not (Bollen & Mao, 2011). If results seems to correct by at least 40 %, it answers yes to the question. Later on to increase the percentage of correctness would be a separate question.

Challenges in Financial SA

That was the survey that was done particularly for sentiment analysis. This showed existing challenges in domain of sentiment analysis and also provided the approach that people used so far to overcome subjected challenges and for better performance. Another survey that was done in correlation to sentiment analysis and finance. There are many such online forums that allow experts to post their reviews and onions about stocks and their predictive price (Kucuktunc, Cambazoglu & et. al., 2012). Major reliable forums that keep historical record are listed in Poon (2003). An extensive review of all such forums along with reference is provided. Engle (1982) presented ARCH model which was Auto Regressive Conditional Hetroskedasticity and presented GARCH model is the extension to this model. Generally GARCH models are used for predicting time sequence that have correlation among the samples at different time intervals. The major application that belongs to such category are finance (Connor, Korajczyk & Linton, 2006), behaviour of electoral (Vikmane & Kreituse, 2009), predictive price calculation for electricity based on inflation (Bosco, Parisio, & Pelagatti, 2007), and many more. Brummelhuis (2008) addresses the current technical development carried out for GARCH and ARCH. Nankervis & Savin (2012) showed dependency in sequel in domain of risk modelling in finance. Underlying assumptions were also studied for same and are examined. Distribution of asymptotes for GARCH and ARCH model's residuals were also studied and presented in Iqbal (2013). Two general models were identified classes of volatile models that are currently widespread (Engle & Patton, 2001).

"One formulating the conditional variance directly as a function of observables and the other formulating models of volatility that are not functions purely of observables" (Wang, 2001).

Scope of the Research

Based on the literature survey, scope of the project is defined. As aforementioned, Sentiment analysis itself have may challenges including sarcasm and many more. While defining the scope of the research, initial step would be to make use of manual samples of sentiments that directly indicate the correct score based on WordNet and lexicon. Study of sentiment analysis in stock forums for multilingual sentiments can be considered as the future scope. Another challenge that is listed in domain of sentiment analysis was for comments that have dependent sentence and irony. For such kind of comments, it is very important to develop a statistical model that define the probability of error for such wrong detection of the sentiment. However, Mukherjee & Bhattacharyya (2013) provided efficient model with derived mathematical equations for same. Thus, to consider this errors and while calculating the system performance, using this model is inevitable. In terms of performance of GARCH model, existing work that is carried out was surveyed and based on that system is proposed. Thus, using this model for online stock forums lies within the scope of the system. Based on the scoring and time vector as defined in GARCH, a decision support system must be defined where various parameters must be defined. Genetic Algorithm aids in defining the optimum solution where there are number of constrains are to be considered. The reason behind mentioning the use of genetic algorithm is that Kuo, Chen & Hwang (2001) showed an approach in which a decision support system was proposed for online stock forum by making use of artificial intelligence with combination of Genetic algorithm, fuzzy logic and neural network with Forwards feed, back propagation. Defining such parameters lies within the scope of the research while, making use of Genetic Algorithm for evaluating the defined parameters can be considered as future scope.

Artefact Design and Modelling

Functional Flow

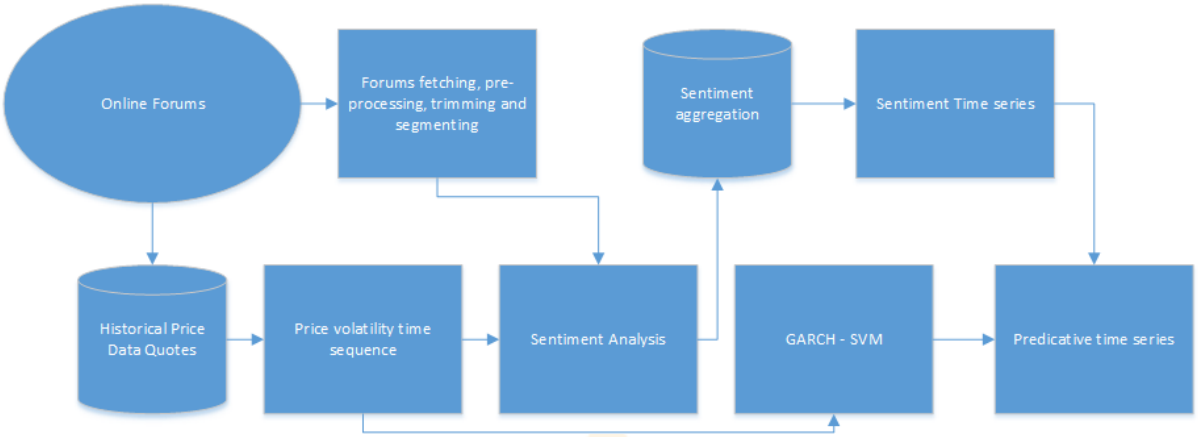


Figure 1: Artefact Model

Figure 1 shows artefact process flow from model point of view. The online forums mentioned here are the raw sentiments or can be considered as news articles. This news articles are fetched from the online forums like business news and share price listing. Share price listing have a unique format and thus, it becomes easy for machine learning tools to fetch the sentiments and interpret them accordingly. This forums need pre-processing like removal of unwanted comments and sentences, segmenting data in form of size and partitioning segment for different time duration (Kuo, Chen & Hwang, 2001). A price volatility time sequence is created for each company that are listed in news articles from all the (historical) price quotes (Engle & Patton, 2001). Here, older the data, better the accuracy in estimation of stock value. However, just for the prototype, price quotes will be fetched from year 2010 only. This price volatility as shown, will serve as input to sentiment analysis and GARCH – SVM algorithm (Iqbal, 2013). Sentiment analysis will perform the scoring for the posts and GARCH – SVM will process data for future prediction for time series. Both, sentiment time series and output of GARCH – SVM will serve as final output of the system, as predictive time series. This predictive time series, shows future value of stock for a company, where user need to select

the name of the company and select year and month for which price is to be estimated. That was the basic model of the proposed artefact from functional point of view.

Process Flow

Figure *2* shows top level flow diagram for the artefact from development's point of view. User is the person from whom a login is expected. Login provision is added to the proposed artefact to avoid misuse of the system. Once, user logs in to the system, news articles will be listed. This new articles are the raw sentiments that are fetched from various online forums of stock news and listing. For prototype, all the news listing from year 2010 for all live companies from date to date will be fetched. This serves as the base input to the proposed system. Selected news articles goes under pre-processing and makes raw sentiments a useful opinion.

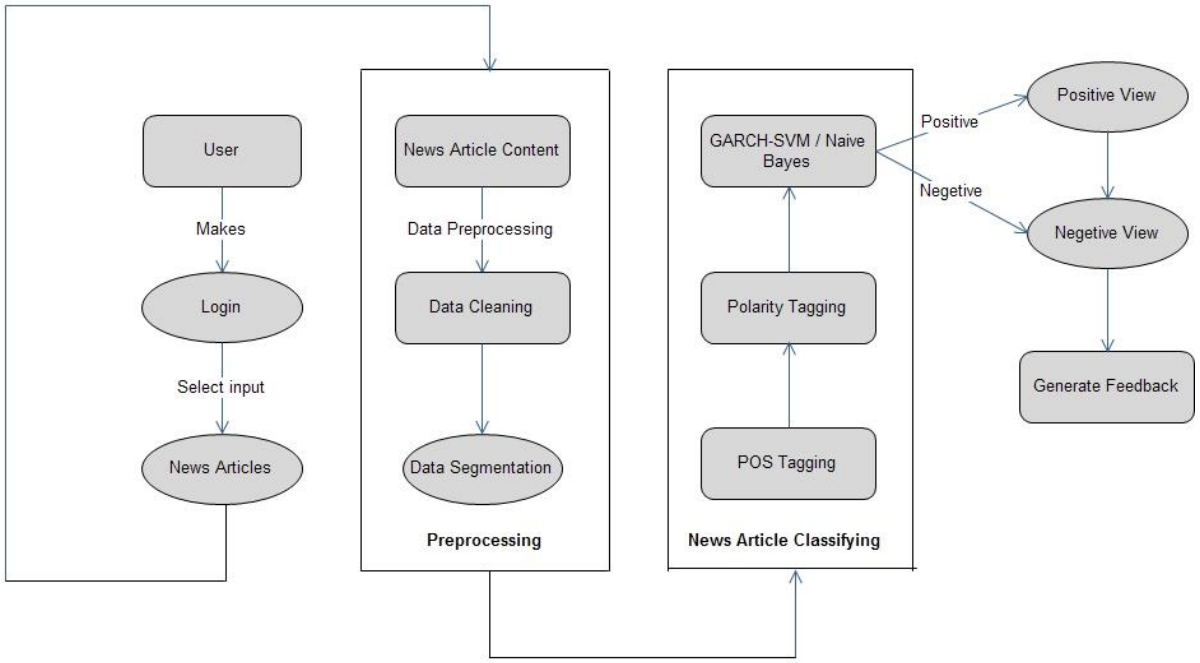


Figure 2: Top level Flow Diagram

Pre-processing

All the raw news articles goes under a process in which data is pre – processed. Here, all the articles that are not useful for sentiment analysis are removed. Moreover, all the manual process that is required to be done on sentiments are assumed to be done prior to this step. Thus, news articles that are not related with the financial domain and particularly for base requirements are removed. During this pre – processing, and removal of unwanted sentiments, the left over sentiments are segmented for different duration of time and for different companies. To refer to different companies, it makes use of predefined keywords that covers almost all companies listed in stock market (Melville, Lawrence, et. al., 2009). Once the data is segmented, initial segmentation is done based on size of data. Here data segmentation is done for quick processing of the data. Data segmentation is done based on size of the data. Further in process of pre- processing is to divide sentiments in n-grams. Figure 3 shown an example of 1-Gram approach (Wu, Zheng & Olson, 2014). The entire sentence is divided in to single words which is obtained by detecting blank space in the sentence. Similarly, n-grams refers to 'n' simultaneous words in single row. Basic advantages of n-gram approach is, it is language independent and can be applied to any languages including English and non – English languages (Wu, Zheng & Olson, 2014). As per the survey, it can even be used for Traditional and Simplified Chinese language. Moreover, as mentioned in literature survey, further grammar correction or parts of speech processing is not required. Word segmentation is already done in gram form thus also avoids situation of word segmentation. As the word is single, it would be easy to identify spelling mistakes and can be easily corrected using language correction tools provided by Microsoft. Also, no additional dictionaries are required to assume the incorrect word. However, it should be noted that, as 'n' in n-gram increases, there is exponential decline in the accuracy obtained by pre – processing (Wu, Zheng & Olson, 2014). Thus, for artefact, 1-gram method will be used in

pre– processing as show in Figure 3. However, it should not be forgotten that with decrease in order of ‘n’ there is linear increase in processing time and accuracy. Thus, the order on ‘n’ is manually selected and is configurable based on optimum value performing multiple iterations on repeated data input. Next step after 1 gram approach is to remove the punctuation marks, as they does not convey any meaning for sentiment analysis anymore. Thus, ASCII values of all punctuations are classified and all punctuations are removed from separated 1- grams (Wu, Zheng & Olson, 2014).

File	1-grams
This stock is not good.	This
	stock
	Is
	Not
	good

Figure 3: N-Gram approach

Feature weight determination

Further pre – processing includes removal of repeated words. However, it should not be forgotten that in compare to regular average, weighted average have better accuracy from statistical point of view. Thus, term frequency will be counted and a weight count will increment for particular word on repetition. Thus, it reduces size of data by removing redundant words, and still impact each term present is maintained (Wu, Zheng & Olson,

2014). Thus, if a sentence is defined in a sample space T, where T consist of {t1, t2, t3 … tn), then Tnew will be defined as Tnew = {w1, t1, w2, t2, w3, t3 … wn, tn}, where, 't' is the unique word in sentence and 'w' is the weight for which respective 't' is repeated in sentence (Wu, Zheng & Olson, 2014). Thus, w1 will be 5, if t1 is repeated for 5 times in sentence. Thus, all repeated words will be removed from the sentence and only unique words will be left with weighted score (Wu, Zheng & Olson, 2014).

Classification

LIBSVM (Wu, Zheng & Olson, 2014) classifier is selected in proposed artefact. Major role of classifier is to capture non – linear relationships by mapping data input vectors as defined in above section into feature space of multi dimension. To make use of such classification, lingual processing is required. As the sentences are already pre – processed, and divided in to n – grams, it becomes easy to identify the adjectives (Wu, Zheng & Olson, 2014). List of adjectives are available and machine learning tool identifies adjectives from segmented words. It should not be forgotten that adjectives play very vital role in sentiment analysis. Further, as mentioned in literature survey, POS i.e. parts of speech tagging is done on segmented words. POS also known as polarity tagging, each piece of word is converted and decomposed into a vector of keywords and each of them are assigned with a specific sentiment value. As mentioned above problems like domain dependency and world knowledge for polarity, to avoid such situation, an online tool known as "HowNet" is available that provide common sense knowledge that is required for inter sentence relation mapping and reviles conceptual relations within chained sentences. Based on polarity, adjectives, weight of the terms, i.e. term frequency and also considering term presence (Wu, Zheng & Olson, 2014), all scores are grouped together and final sentiment score is given that mentioned whether as positive or negative or neutral.

Score Indexing

Based on the score grouping, final sentiment index is required to be built. Thus sentiment value index points as positive or negative for particular company and particular day. Thus, for each and every company listed, a sentiment score index will be made for each day from the date, when historical data was available. Bullish index approach is used to aggregate the sentiment scores. As per the survey, it is the most robust technique that can be used regardless of language and input data type. This bullish index is obtained from total weight of bullish and bearish opinions. Bullish refers to opinions that believes rise in price and bearish refers to opinions that believes fall in prices. Thus, aggregated score, i.e. bullish index would be given by following:

$$BI = \ln[\frac{1+M^{bullish}}{1+M^{bearish}}],$$

Where, $M^{bullish}$ refers to total bullish opinions for particular stock and $M^{bearish}$ refers to total bearish opinions for particular stock (Wu, Zheng & Olson, 2014). Thus, intimately BI value would either a negative number or positive number. Thus, if value of BI is greater than (>) 0 implies that share price for particular stock will be high and if value of BI is less than (<) 0 implies that share price for particular stock will be lesser in future (Wu, Zheng & Olson, 2014). It should be noted that time stamp of particular opinions also plays vital role and must be recorded and mapped with this scoring. Timestamp refers to the time when particular opinion was given, thus it helps in time vector series mapping.

GARCH – SVM modelling

Price volatility means the standard deviation or in other words, variance in change of value in a specified time duration. This approach is most used while modelling time series particular in financial domain as it parades time fluctuating volatility. Bollerslev proposed the GARCH model which can be formulated as shown below (Bollerslev, Chou & Kroner, 1992):

$$yt = ut + Et$$

Here, ut is the deterministic mean with Gaussian distribution and zero mean, and Et is the stochastic process known as forecast error. Daily return is given by yt, which is sum of both the terms. GARCH model estimates using "conditionally Gaussian log-likelihood function" and maximizes it using BHHH iterative algorithm (Berndt et al 1974), as the function that is to be maximized shows nonlinearity in its arguments. Maximum Likelihood estimation is used when probability density function have Gaussian distribution for the sampled input data. Maximum likelihood estimation makes use of pseudo inverse using known training symbols and estimation coefficient is calculated. This calculated coefficient is the Et as mentioned in above equation. However, in case if probability density function (PDF) is not following Gaussian distribution, it must be pseudo converted to Gaussian distribution with zero mean and in such case, quasi – ML (Maximal Likelihood) estimation is used. However, it was found in Bollerslev and Wooldridge (1992) that the consistency of these estimation technique used, does not ensure that it is the best estimation technique for finite sample sets. SVM approach as mentioned above, helps in this case for efficient estimation. The vector selection of SVM using cross validation is used to adjust the GARCH parameters with the probability density function of the optimal solution using blind estimation technique. However, in case if the probability density function is not Gaussian, native ML estimation technique must be used.

Following are the mathematical terms used in statistical modelling of GARCH – SVM (Wu, Zheng & Olson, 2014).

Let U be the main set of users

U = {u1, u2, u3…….}

Let D be the main set of input documents (News Articles / Tweets)

D= {d1, d2, d3…….}

Let W be the set of Word Segments in Each news Article

W = {w1, w2, w3…….}

Let P be the main set of Positive Words in Each News Article

P = {p1, p2, p3……}

Let N be the main set of Negative Words in Each News Article

N = {n1, n2, n3……}

Let F be the feedback about Stock market based on News Articles

F = {f1, f2, f3……..}

If P is the main set of Processes given by,

P = {P1, P2, P3……}, where

P1 = {e1, e2, e3, e4} set of processes for pre-processing

Where,

{e1=i|i is to removing stop words}

{e2=j|j is to apply steaming on News Articles}

{e3=k|k Word Segmentation}

{e4=l|l Get Features}

P2 = {e1, e2, e3, e4} set of processes for Sentiment Analysis by GARCH-SVM

Where,

{e1= i|i is to make Stock market Predictions}

{e1: i|i Features Selected}

{e2: j|j Feature Reduction & calculate Feature Weight}

{e3= k|k is to predict stock market sentiment by using Polarity Tagging}

{e4= l|l Prediction of Stock Market}

It should be noted as, the solution for given polynomials are provided using GARCH – SVM algorithm, lies under non deterministic polynomial time as set of all decision problem's solution is provided into polynomial time using the GARCH - SVM algorithm. In the proposed artefact, Stanford-postagger-3.1.4, which is a free java library is used for Tagging News Article and its format is as shown below:

- Input: {News Articles / Online Forum Tweets}
- Output: {Feedback about Stock Market}
- Success: {Correct Feedback about Stock Market based on news Articles Content}
- Failure: {fail in case of Unstructured News Articles/ Tweets as a input}

Artefact Design

Technical design of the artefact is represented in multiple point of views. This helps in understanding the artefact development from multiple point of view. For example, stake holders can view use case diagram and can understand how it can affect their business. Database administrator can get idea about logical database requirements and can developed physical database from that. A reviewer at management level can make use of block diagram and activity diagram to understand artefact from management point of view. A co-developer can make use of class diagrams and understand packages and classes used and can revoke them wherever required. A system analyst can make use of sequence diagram and can get picture of artefact in terms of message flows and requests. Thus, different diagrams represents different information about the artefact.

Block Diagram

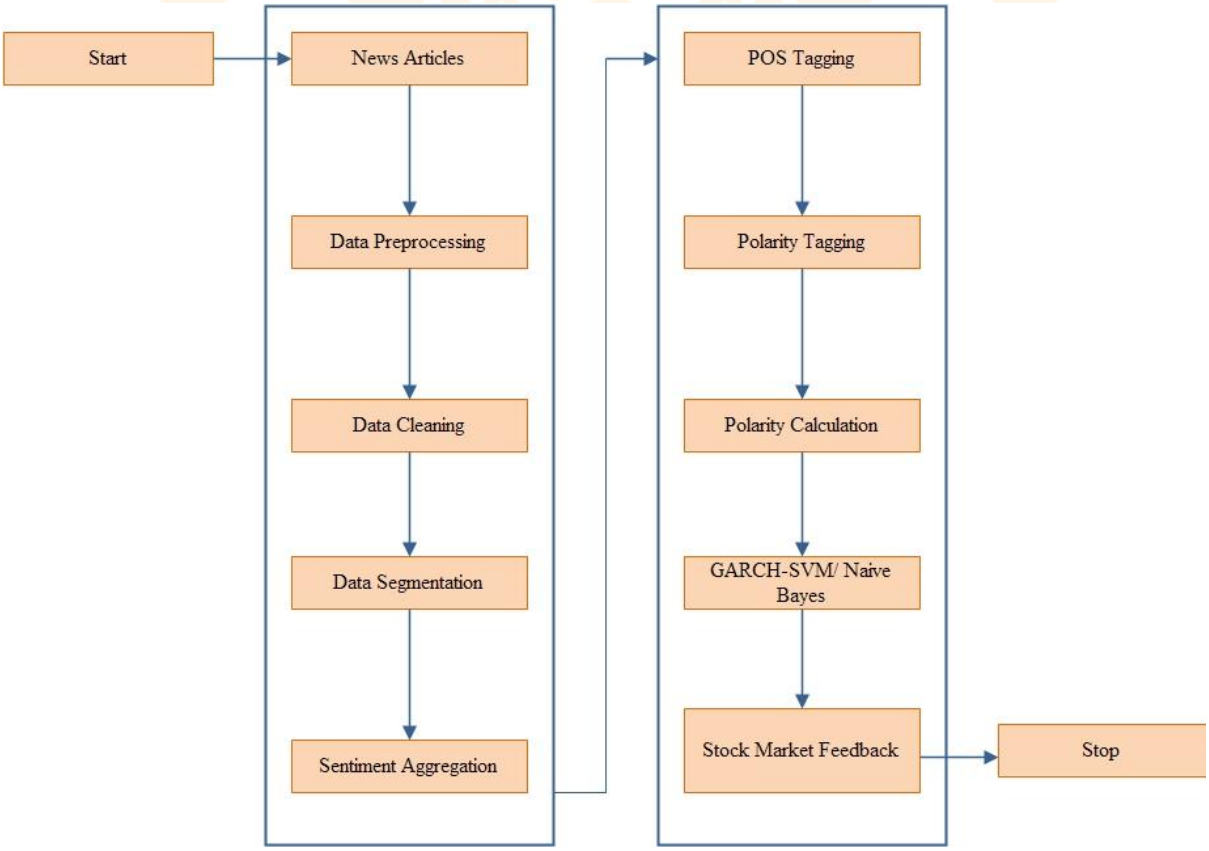


Figure 4: Block Diagram

Figure 4 shows technical block diagram of the artefact. The process starts by reading news articles. Then further data pre – processing is done as mentioned in above sections. Once pre – processing is done, all unwanted sentiments, punctuations and redundant words are removed in data cleaning process. Further, sentiments aggregation is done by the process mention in above section. To the aggregated sentiments, parts of speech tagging is done. This defines the polarity of the sentence. Further, polarity is calculated and final polarity results are given to GARCH – SVM algorithm (Wu, Zheng & Olson, 2014). There, it performs statistical analysis and out of algorithm provides stock market feedback. This was the actual artefact divided in functional blocks.

System Architect

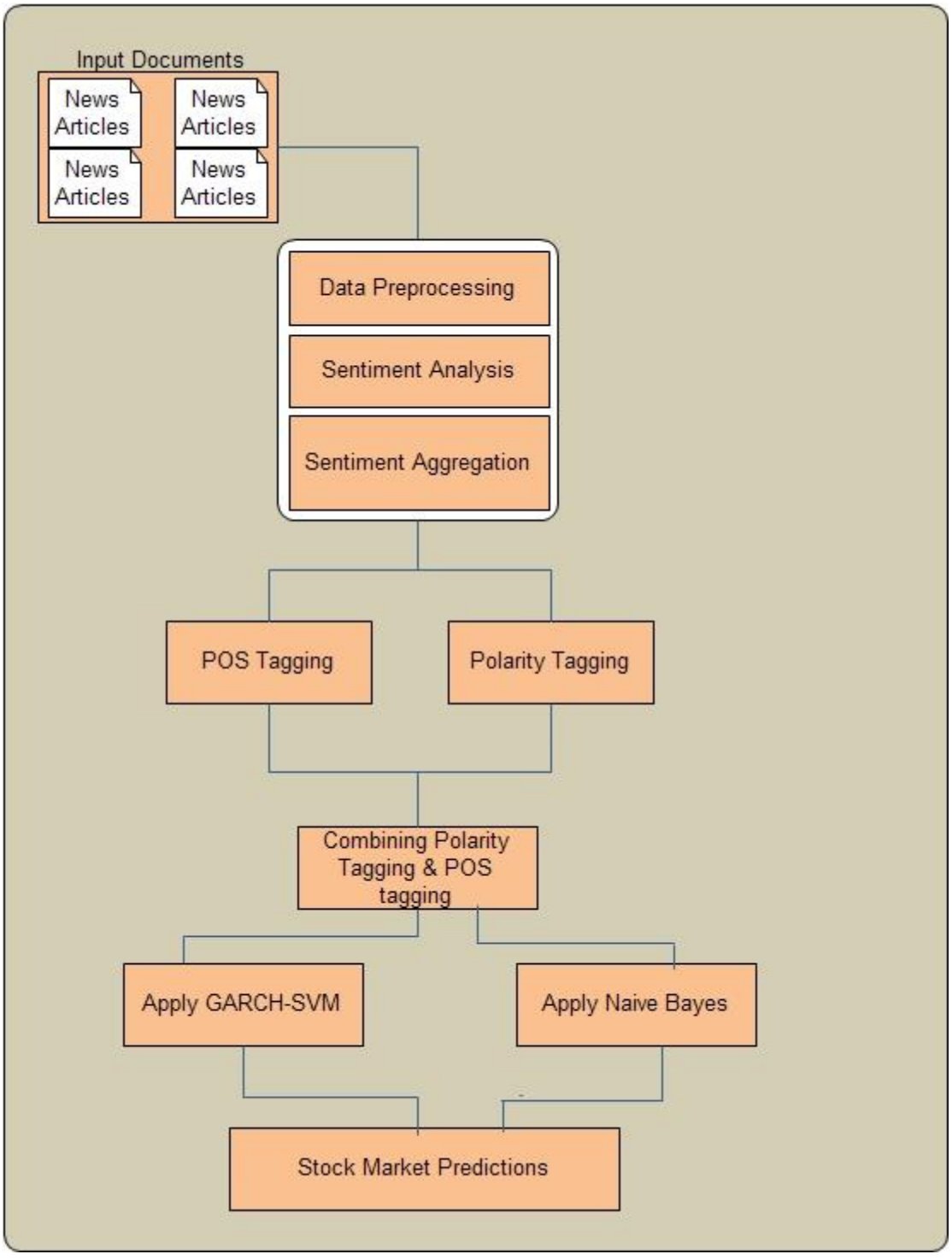


Figure 5: System Architect

Figure 5 depicts the architectural view of the proposed artefact. It shows the same process flow as aforementioned.

Use case diagram

Use case diagram shows interaction of user with the system (Rumbaugh, et. al., 2004). This helps in understanding multiple users involved in usage of the system at different cases and their roles. Figure 6 represents the use case diagram for the developed artefact. It shows that there is single user interaction and no simultaneous user interaction with the system. Moreover, a single user is supposed to interact with the system at different cases like, selection of news articles, data pre – processing, data pre – treatment, and same user expects the output result of estimated stock market prediction result. However, interaction of system within multiple cases at multiple instances is also shown. It also shows type of interaction between different use cases.

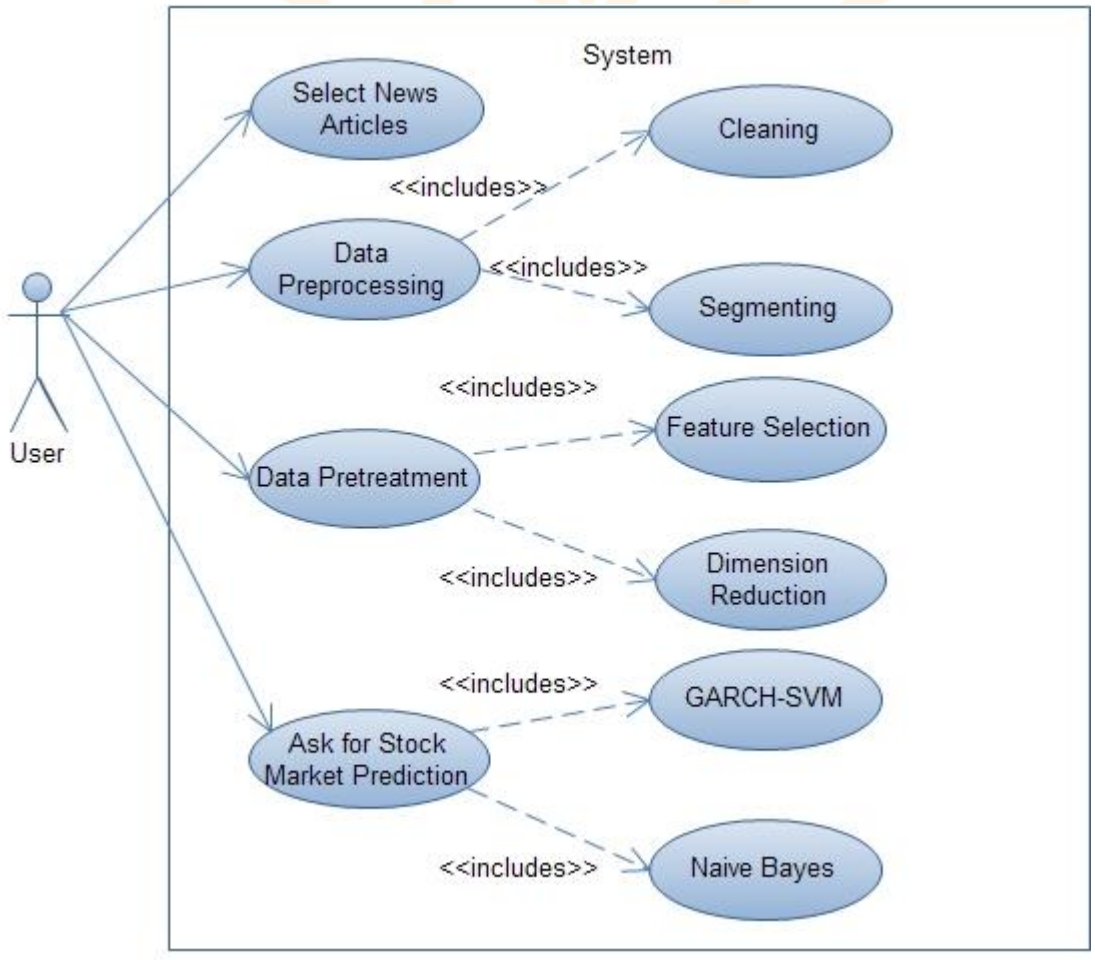


Figure 6: Use Case Diagram

Activity Diagram

Activity diagram in Figure 7 depicts the workflow or the activity flow of the developed artefact (Rumbaugh, et. al., 2004).

- Fist activity starts with user login.
- Second activity is to get news articles.
- Third activity is to pre – process the articles.
- Forth activity is to tag with parts of speech
- Fifth activity is to map polarity with the articles
- Sixth step if for parallel processing of scored sentiments with GARCH – SVM and Naïve Bayes processing
- Seventh step is to get the results of stock market prediction

Reason for choosing Naïve Bayes approach is for quality analysis of the developed artefact.

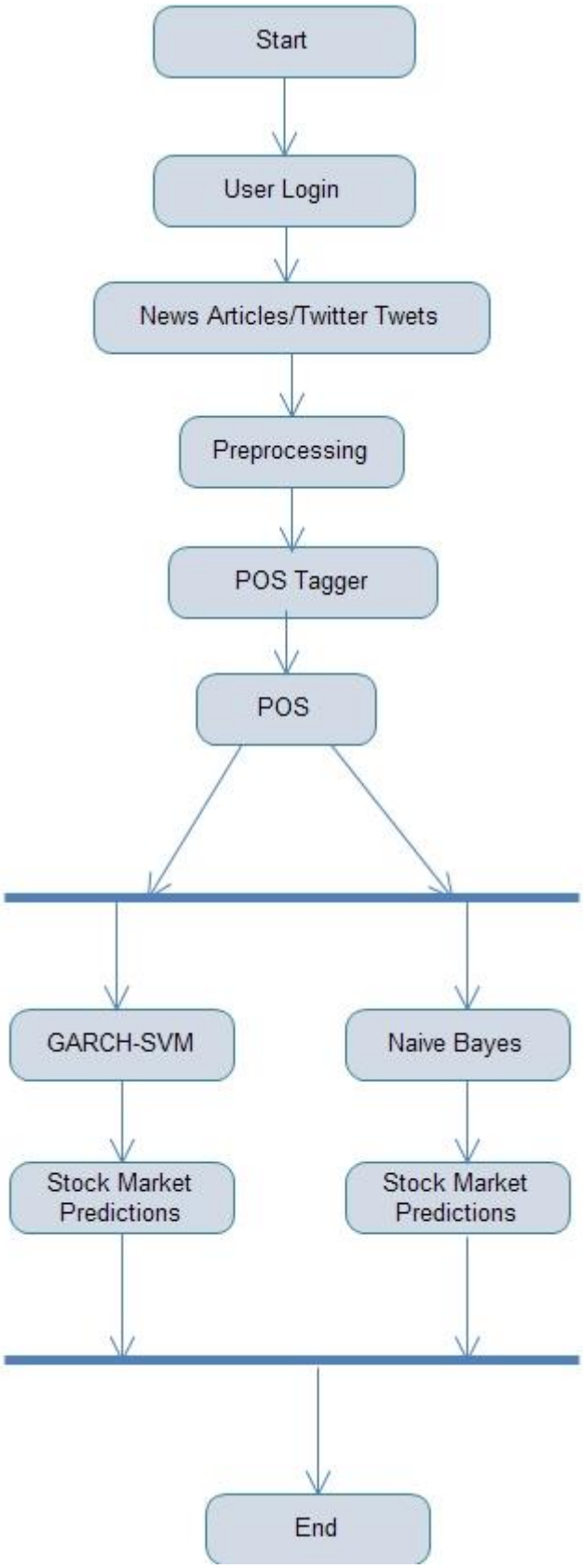


Figure 7: Activity Diagram

Class Diagram

Class diagram in Figure 8 shows the structure model of the developed artefact. Further in structural modelling, it shows inheritance and interaction within the classes (Rumbaugh, et. al., 2004). For example, frame login class needs username and password in form of string as data type and uses them as argument in the function call for checkLogin(). Thus, it shows the structure of the developed artefact from coding point of view.

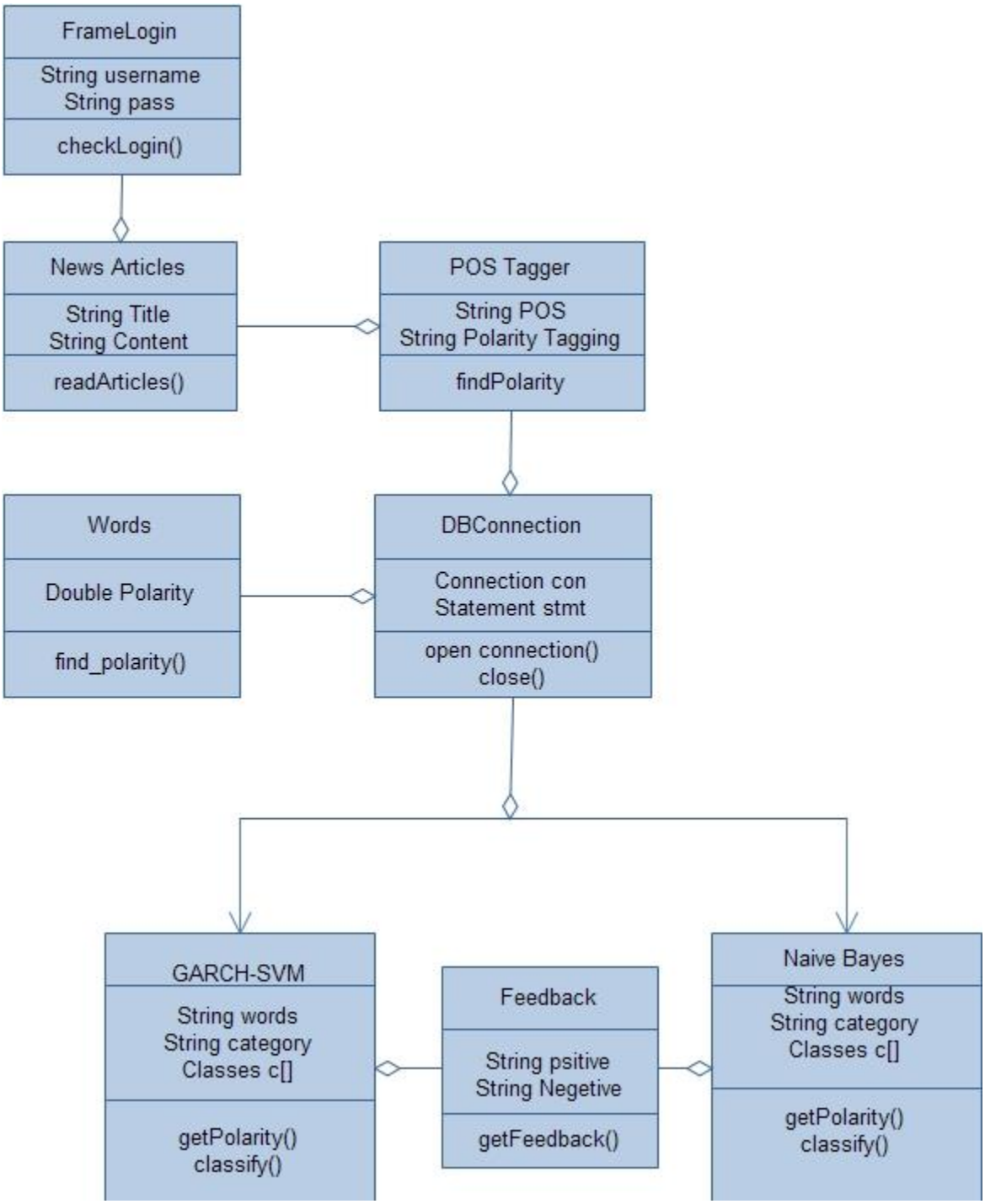


Figure 8: Class Diagram

Component Diagram

"A component diagram depicts how components are wired together to form larger components and or software systems. They are used to illustrate the structure of arbitrarily complex systems." (Rumbaugh, et. al., 2004). Component diagram as shown in Figure 9

represents developed artefact in form of components and its interaction within the components. It shows the structure of the arbitrary complex system in developed artefact.

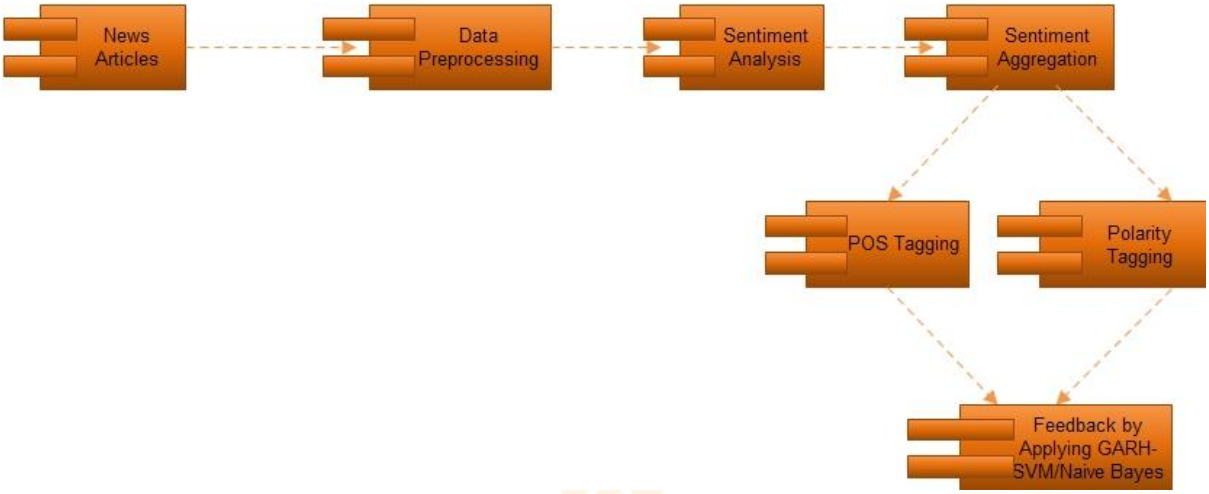


Figure 9: Component Diagram

Data Flow Diagram

Data flow diagram shown the actual flow of data within the artefact.

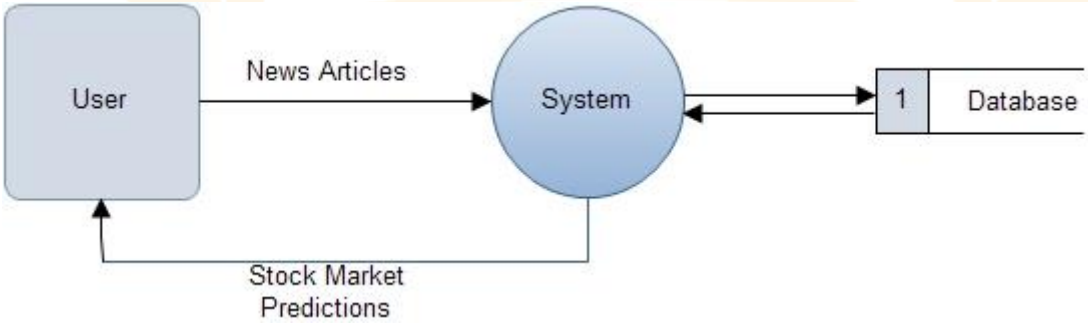


Figure 10: User - System DFD

Figure 10 shows data flow between user, system, i.e. artefact and database. System fetches raw data from the database, processes the data as per the requirements and stores back in the databases. Thus, system have bidirectional flow with data base. User provides news articles to system and expects stock market predictions. System expects such inputs and outputs to the user and accordingly interprets with the database.

Figure 11 and Figure 12 depicts data flow diagram with expanded system block diagram as shown in Figure 10. Thus, it represents the internal dataflow within the artefact, post to user inputs and prior to expect user output.

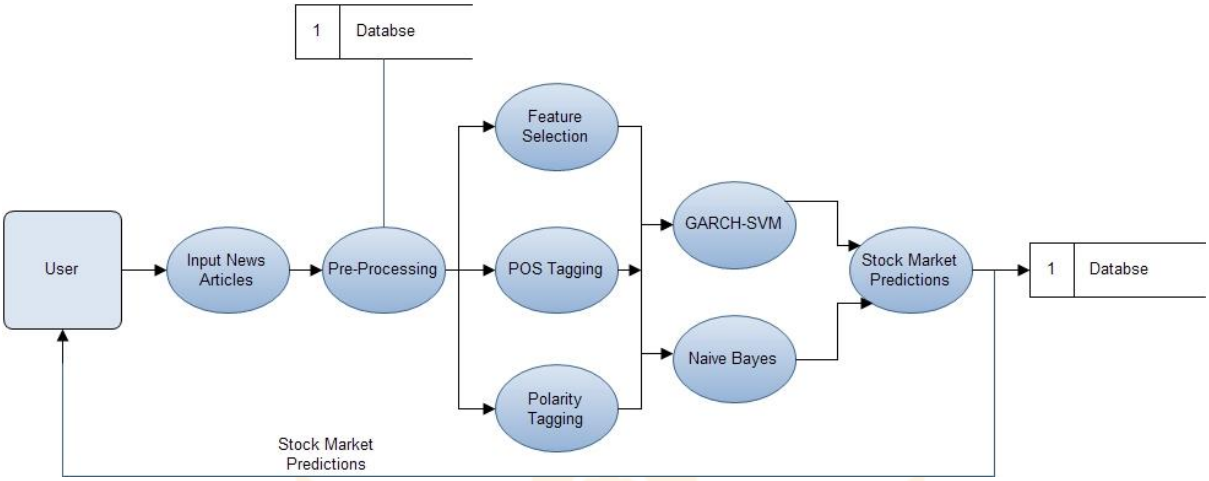


Figure 11: System Expanded DFD-1

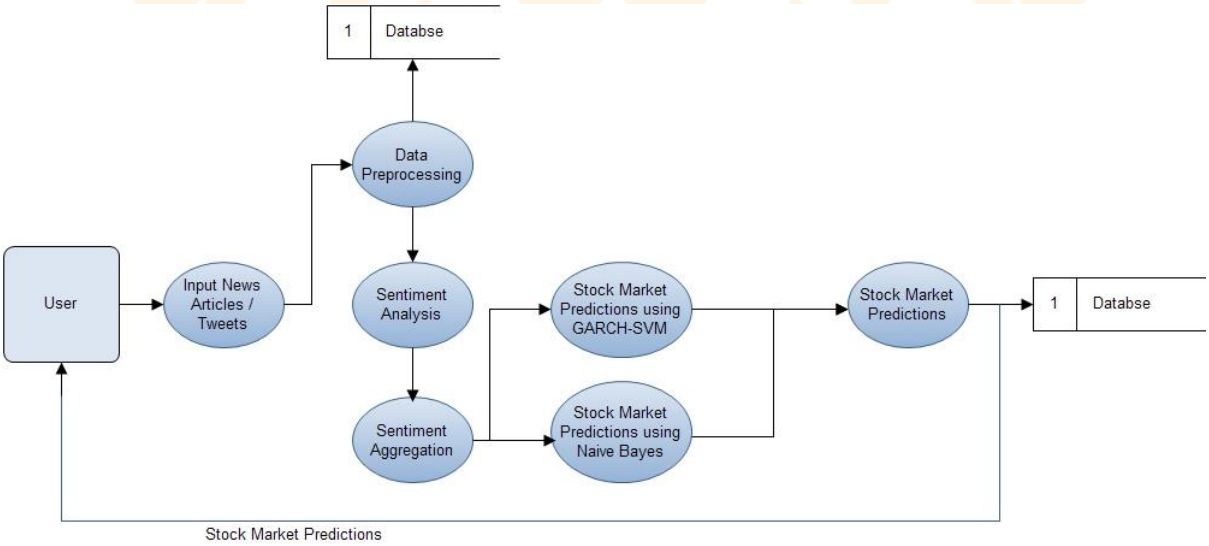


Figure 12: System Expanded DFD-2

Entity Relationship Diagram

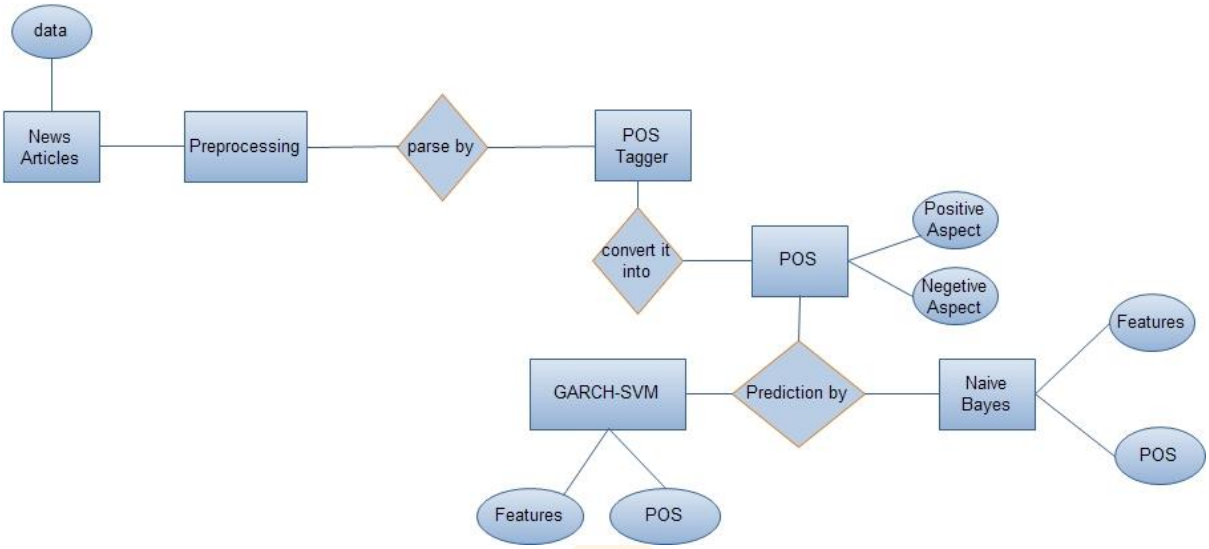


Figure 13Entity Relationship Diagram

Figure 13 depicts the relationships between different entities of the developed artefact in terms of procedural requirements (Rumbaugh, et. al., 2004). This is the logical form of database representation which help to develop physical form of the database.

Sequence Diagram

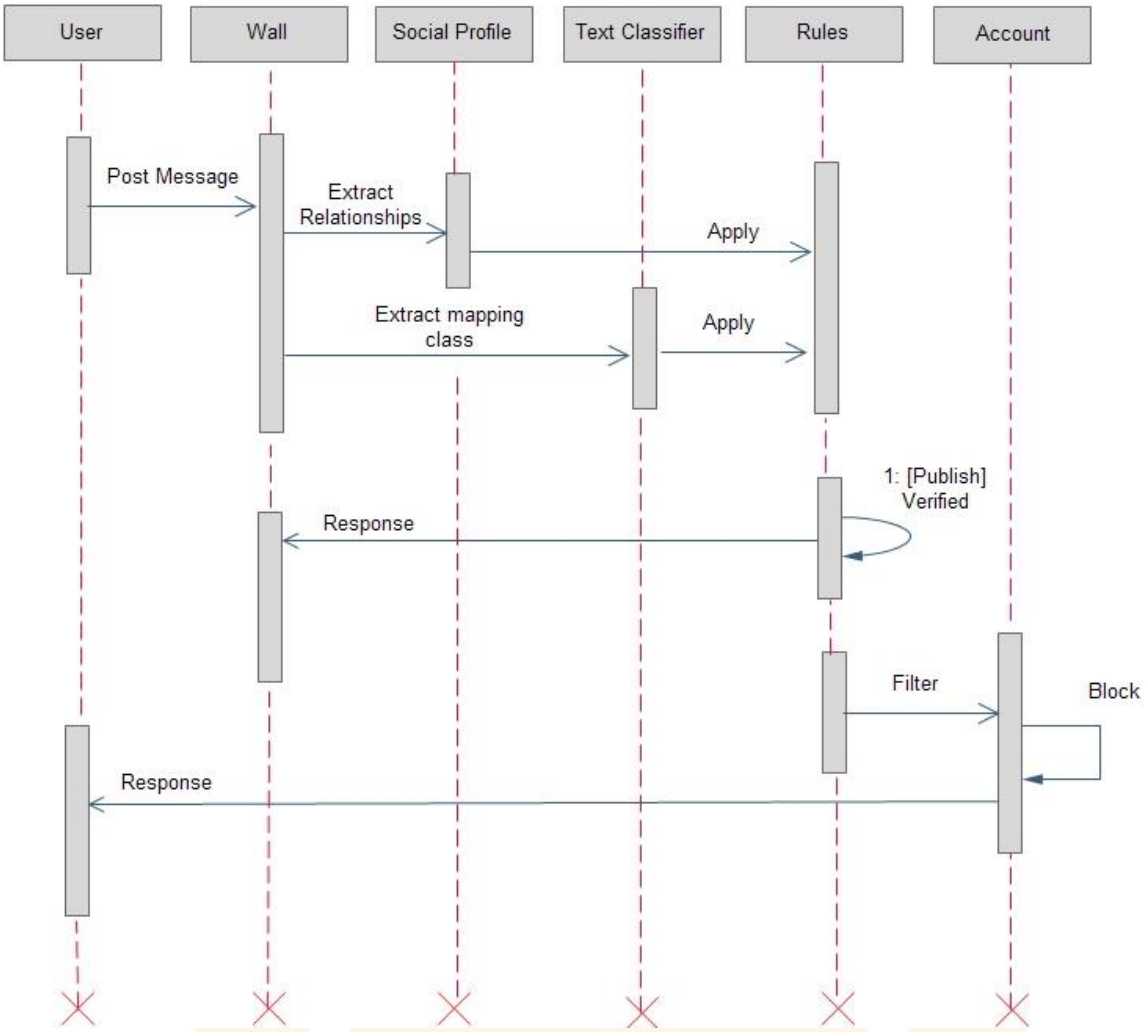


Figure 14: Sequence Diagram

"A Sequence diagram is an interaction diagram that shows how processes operate with one another and what is their order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios. A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live

simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner." (Rumbaugh, et. al., 2004). Figure 14 shows sequence of message flows within the developed artefact in terms of request and response. It helps to understand the message flows within the classes. It starts with user posting a message and ends with response of required prediction.

State Diagram

"State Diagram is specialization of behaviour and is used to specify discrete behaviour of a part of designed system through finite state transitions. The state machine formalism used in this case is an object-based variant of Harel state charts." Figure 15 shows state transactions of the artefact at different at different level and processes.

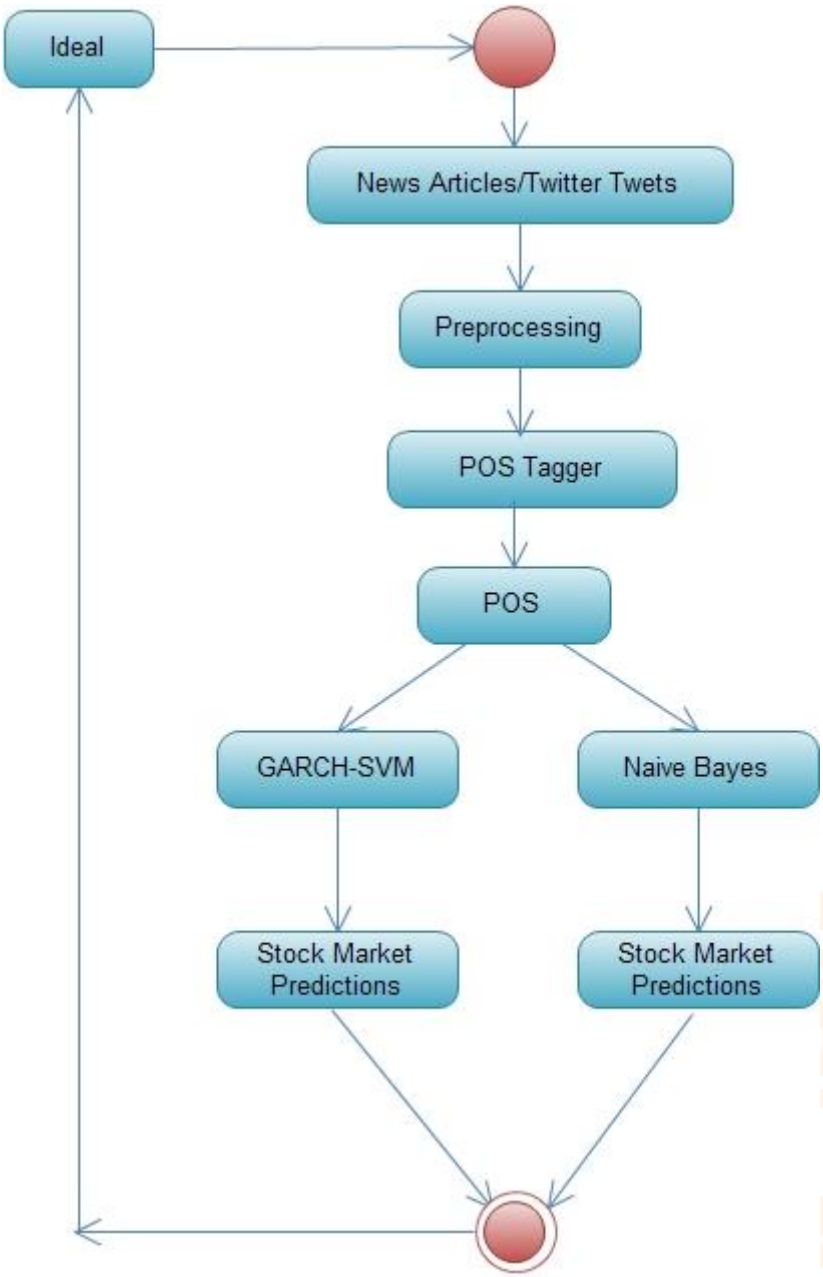


Figure 15: State Diagram

Design Details

This section shows list of packages their description and classes used in it.

Package: com.GUI

Description: It is Responsible for Graphical User Interface

Class	AccuracyFrame.java
Description	It is used to generate frame for last frame, i.e. accuracy frame which shows the accuracy graph and accuracy percentages. It also shows the number of times the prediction is tried by user. It has methods for fetching data from database to generate the graph.
Class	AllPredictionsTabular.java
Description	This class is used to generate the frame which shows all the predictions till date in table format. It has methods for loading data from Database into table.
Class	BaseApproach.java
Description	This class is used for applying Base Approach for today's recorded data. It has methods for calculating today's prediction for selected company. It has methods for getting data from database, and calculating the prediction. This class is called from Dashboard frame while program is running.
Class	CurrentBSE.java
Description	This class is used for getting stock data for a company that, is recorded. It fetches data from database and shows in table. This Class is called when double clicked on any company from Today's Gainers or Today's Looser companies list.
Class	Dashboard.java
Description	This class is responsible for displaying Dashboard to user. It is the class that is most working. This shows controls for running artefact, i.e. fetching data over Internet, showing the graphs, showing accuracy etc.
Class	DSSALauncher.java
Description	This class is the start point of artefact. While running this file it runs other

	java classes in cascading manner. When clicked on login button it calls the Login.java class and then Login frame opens.
Class	Fin_Analysis.java
Description	This class is responsible for generating the frame, when clicked on button from Dashboard frame named as Contribution. It has methods for getting data from Database. This class uses the Naïve Bayes theorem for the purpose of prediction. It shows the MAX and MIN values for sentiment as well as today's gain or loss. It also shows the prediction.
Class	GraphRenderer.java
Description	This class is responsible for generating graph, when clicked on buttons named as Price (Gain Or Loss) and Sentiment Index from Dashboard, this class reads data from database and generates the graph.
Class	Login.java
Description	This class is responsible for Login frame. After entering username and password it calls another named as Connector.java class for checking the user entered values against values from Database, and proceed to dashboard frame.
Class	PredictionGraphFrame.java
Description	This class is responsible for generating frame to Show Graph for all Predictions that have been tried till today. This class get called from Dashboard When Clicked on Button named as Show All Predictions Graph. This class fetches data from prediction info table from database
Class	Register.java
Description	This class is responsible for the registration frame, where a new user can enter his/ her details for registration purpose.

Class	StockMarketRenderer_BSE.java
Description	This class is responsible for displaying frame which shows all stock related data, i.e. all data recorded till date. This class is called from Dashboard, when clicked on BSE button under named as Historical Data.

Package: com.porterStemmer

Description: This package is responsible for processing comments which are downloaded from moneycontrol.com

Class	FileSentimentExtractor.java
Description	This class is responsible for processing comments and extracting sentiment index for each company. In this class, to process the comments that are financially contextual and important are processed and sentiment index is calculated and is recorded to database. This class is used for reading the comments which are downloaded and stored in files. Also it processes these files and calculates sentiment value for each company

Package: com.processArticles

Description: This package is responsible for processing the files. It is supportive and not of much importance.

Package: com.Test

Description: This package is used for the Testing purpose, i.e. if there is new code to write and test, it done here.

Package: net.OnlineDataExtractor

Description: This package is responsible for extracting data from online feeds

Class	Extract_BSE_NSE_from_Single_Page.java
Description	It is responsible for connecting to moneycontrol.com and fetching all stock related data, which consists of current stock price, current loss or gain, comments which are uploaded by online users for each company. This class gets called from Dashboard when clicked on button BSE under label Fetch Data, while fetching data over internet it writes the comments into files, and stock related data is recorded to database. After completion of fetching data for all companies, it will classify the companies which are gaining profit and companies which are losing for today. And update it into lists on Dashboard. And then Base Approach And Contribution on this data is applied.

Package: util

Description: This package has class Date Time which is used for getting Date and Time in Artefact

Package: com.graphs

Description: This package is responsible to plot appropriate graphs

Class	AllPredictions.java
Description	This class is responsible for generating graph for ALL PREDICTIONS that user has done till date. This class accesses the data from Database and renders graph.
Class	BaseContrAccuracy.java

Description	It is used for generating the graph for comparison between base approach and contribution
Class	CommentsGraph.java
Description	It is used for generating the graph showing volume of comments graph
Class	DailyGain_Graph.java
Description	It is used for generating graph for daily gain or loss. This class is called from Dashboard, when clicked on button Price (Gain/Loss).
Class	Sentiment_Graph.java
Description	It is used for generating graph for sentiment verses date. This class is called from Dashboard, when clicked on Sentiment Index

Package: com.GARCH

Description: It is responsible for calculations for Base Approach. Classes within this package are responsible for calculating variance and predicting market depending on today's market. Classes from this package gets called after a click on Base Approach button from Dashboard.

Package: com.DBConnection

Description: It is responsible for database connection and all database actives

Class	Connector.java
Description	This class is used for all database activities. It has methods for Connecting To Database through Java, Closing Connection, User login and User Registration
Class	TableGetter.java

Description	This class is responsible for getting list of all the tables in database
Class	UpdateStockValues.java
Description	It has methods for inserting the downloaded data from moneycontrol.com for all companies

Package: com.CLSFR

Description: This package is responsible for all classification related work, i.e. looking into history the patterns in the database.

Package: com.companyData

Description: Responsible for inserting all company related data to database

Class	Companies.java
Description	This class is responsible for listing companies who are Gainers and Losers for today. This class is called from Dashboard Class. A Lists for today's gainers and Today's losers is displayed on dashboard
Class	Prediction.java
Description	It deals with Prediction Related information for each company. This is called when clicked on Contribution button from Dashboard
Class	PredictionBase.java
Description	Deals with Prediction Related information for each company, this is called when clicked on Base Approach button from Dashboard

Testing and Evaluation

Testing and evaluation of the proposed artefact is very vital. As the proposal deals with future forecasting, if forecasting is found to be accurate even by 40% would be considered as success and its improvement can be considered as future scope. Testing is divided into two section, visually, User acceptance testing and performance evaluation testing. User acceptance deals with testing of GUI and objects that user interacts with like buttons pressed, etc. while performance evaluation testing involves logical testing of proposed approach and comparison with base approach which defines whether the developed approach is correct, etc.

User Acceptance testing – UAT

Sr. No.	**Module**	**What User Sees**	**User Input**	**Expected Result**	**Comment**
1	Login Form	GUI Frame for accepting Username & Password with Login button.	Leave Username and Password Fields Empty & click on Login button.	Error message must be shown as "Don't Leave Any of Fields Empty..."	Successfully tested
2	Login Form	GUI Frame for accepting Username & Password with login	Enter wrong Username or/& Password and click on Login button.	Error message must be shown as "Username Or Password Error".	Successfully tested

		button.			
3	Register Form	GUI Frame Which accepts User Information	Click on Register button with empty fields	Error Message must be shown as "Field Error"	Successfully tested
4	Register Form	GUI Frame Which accepts User Information	Click on Register button without name	Error Message must be shown as "Field Error"	Successfully tested
5	Register Form	GUI Frame Which accepts User Information	Click on Register button without Password	Error Message must be shown as "Field Error"	Successfully tested
6	Register Form	GUI Frame Which accepts User Information	Click on Register button without Confirm Password	Error Message must be shown as "Field Error"	Successfully tested
7	Main window	GUI of DSSA	Click on BSE button in	Current Stock Market frame	Successfully tested

			Historical Data Panel	must open.	
8	Current Stock Market Window	GUI Frame of Current Stock Market Window	Click on Get Companies and double click on Company name	Table Must loaded with values for selected company	Successfully tested
9	Main Window	GUI of DSSA	Click on Price Button under graph label	Graph frame must open.	Successfully tested
10	Graph Window	GUI frame for Graph Window	Double click on Company name	Panel Must loaded with graph for selected company	Successfully tested
11	Main window	GUI of DSSA	Click on Sentiment index Button under graph label	Graph frame must open.	Successfully tested
12	Graph Window	GUI frame for Graph Window	Double click on Company name	Panel Must loaded with graph for selected company	Successfully tested
13	Main window	GUI of DSSA	Click on Volume of Comments	Comments graph frame must open.	Successfully tested

			button		
14	Main window	GUI of DSSA	Click on Prediction Graph for Today Button	Today's Prediction frame must Open	Successfully tested
15	Today's Prediction	GUI frame for Today's Prediction	Double click on Company name	Panel Must loaded with graph for selected company	Successfully tested
16	Main window	GUI of DSSA	Click on Show All Prediction Graph button	Graph frame must Open	Successfully tested
17	Graph Frame	GUI frame for Graph Frame	Double click on Company name	Panel Must loaded with graph for selected company	Successfully tested
18	Main window	GUI of DSSA	Click on Predictions In tabular button.	Predictions in Tabular Frame must open.	Successfully tested
19	Predictions in Tabular Frame	GUI Frame for Predictions in Tabular	Double Click on company name	Table must loaded with the values.	Successfully tested

20	Main Window	GUI of DSSA	Click on Accuracy button.	Accuracy frame must be opened	Successfully tested
21	Accuracy window	GUI Frame For Accuracy	-	User must see the accuracy calculated and Graph of Accuracy.	Successfully tested
22	Main Window	GUI of DSSA	Click on BSE button under label Fetch Data label.	User must wait while Data is fetched, Cursor changes to busy cursor. After Fetching The data Lists from panel must be loaded, as today's gainers,	Successfully tested

				and today's losers.	
23	Main Window	GUI of DSSA	Click On Reset button	Lists for today's gainers and today's losers must be cleared. And Base Approach and Contribution Buttons must be Disabled.	Successfully tested
24	Main Window	Lists Loaded With today's Gainers and Losers	Double Click on Any item from list.	New frame with values loaded in Table	Successfully tested
25	Main window	GUI of DSSA	Click on Base Approach button	Base Approach Frame must be opened. Double Click on Company name, values must be	Successfully tested

				loaded into fields specified	
26	Main window	GUI of DSSA	Click on Contribution button	Contribution work Frame must be opened. Double Click on Company name, values must be loaded into fields specified	Successfully tested
27	Main window	GUI of DSSA	Select exit from Utilities menu	All windows must closed and Project must exit	Successfully tested

Performance Evaluation

To check the performance of the system, special functional testing is required.

Pseudo Prediction testing

First approach is to predict stock values for past values. In this approach, a date is selected as current date that is already a past date. All the input feeds are considered that are prior to the selected thus. Using these data, stock price is estimated for this selected current date. Now, in real situation, the current date selected is already a past date. Thus, price and stock value is already known for that. Now these pseudo predicted and actual prices must be compared and performance of the system can be calculated. This approach is known as pseudo prediction testing. To explain this scenario, let us consider an example. Let us assume that today's date is 2nd Jan, 2015. If current date is selected as 31st Dec, 2014, than 1st Jan, 2015 i.e. next day

would be a future date compared to current, which actually a past date too. Thus, if all the input articles up to 31st Dec, 2014 is considered as available input and if prediction of stock value is done for 1st Jan, 2015, this prediction would be actual prediction for the artefact, as no data was give prior to 31st Dec, 2014. But it should not be forgotten that 1st Jan 2015 is actually a past date and actual stock price is available for it. Thus, artefact predicted and actually available stock price can be compared and performance of artefact can be compared. Thus, approach is known as pseudo prediction. Moreover, same testing can be done, by predicting actually next day in future but to check the performance evaluation, a wait time of 24 hours would be required to get the actual stock price.

Table 1: Pseudo Prediction Testing-1

Pseudo Date duration	**Correlation with real price for actual current date**
Last year	10%
Last 9 Months	13%
Last 6 Months	15%
Last 3Months	35%
Last Month	59%
Last 15 Days	60%
Last Week	65%
Previous Day	80%

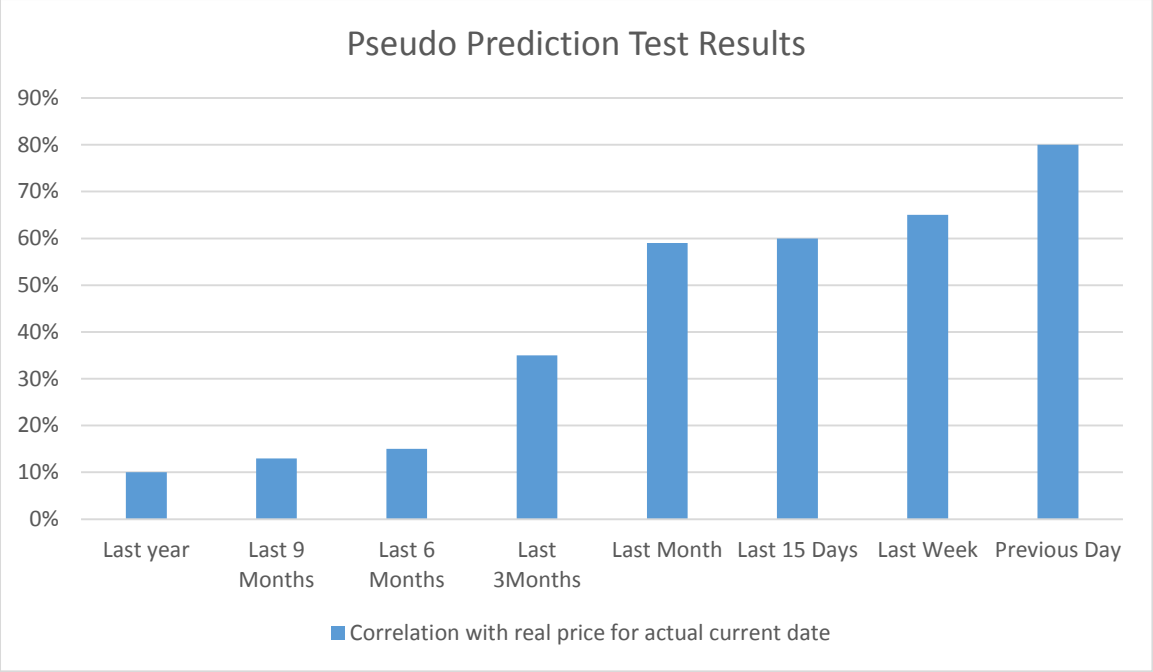


Figure 16: Pseudo Prediction Results for Actual Current Date

Table 1 shows test results for the artefact. During this test, the column Pseudo Data duration indicates the duration of selecting current date than actual current date. For example, if duration is 1 year, then if actual current date is 1st Jan 2015, then in artefact it must be select as 1st Jan 2014 as current date and prediction is calculated. It should be noted that, for such testing, inputs of data must also be limited. For instance, if testing is to be done taking current date as 1 year back than actual current date, then all the news articles must also be lesser than selected day. Then only, the real performance of the artefact can be calculated. Figure 16 shows results plotted in form of bar chart which shows that considering current date as 1 year back gives only 10% of accurate result. As the duration decreases, there is increase in performance for prediction of stock price for actual current date. Looking to the graph, it shows exponential rise distribution, however experiment is not self-sufficient to conclude same.

Table 2: Pseudo Prediction Testing-2

Pseudo Date duration	**Correlation with real price for pseudo current date**
Last year	88%
Last 9 Months	85%
Last 6 Months	83%
Last 3Months	87%
Last Month	78%
Last 15 Days	80%
Last Week	82%
Previous Day	81%

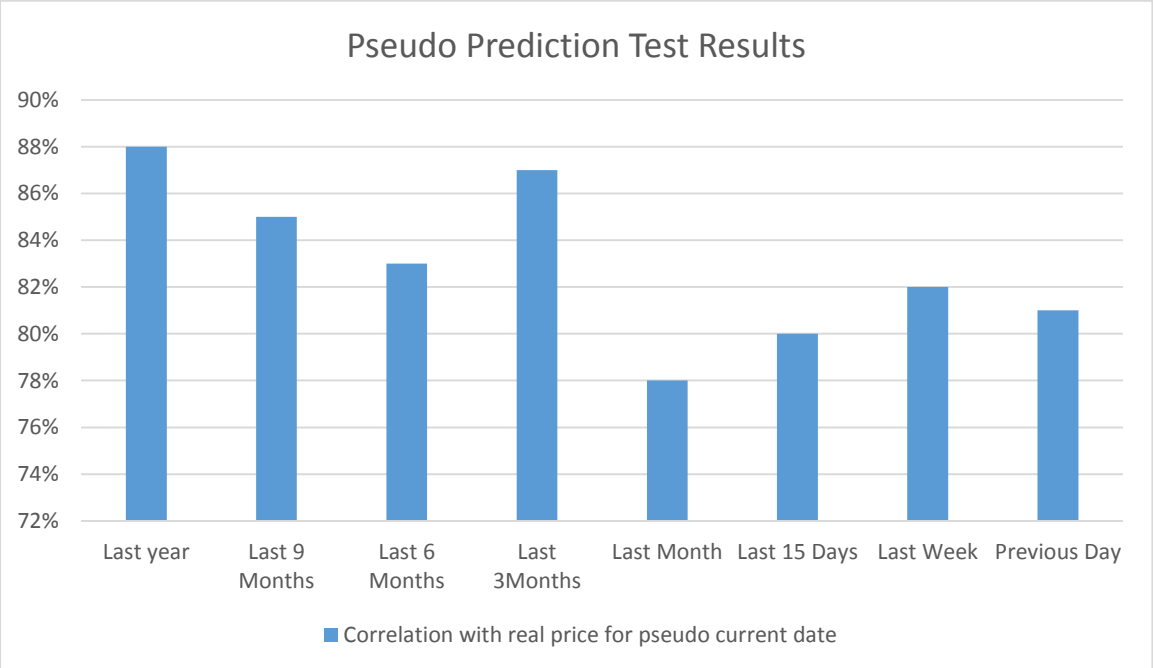


Figure 17: Pseudo Prediction Results for Pseudo Actual Date

Table 2 shows test results for predicting pseudo current date. The only difference of this results from Table 1 is that, in Table 1 prediction was done for actual current date and here prediction is done for pseudo current date, i.e. the next of the selected date. For example, if duration is selected as 1 year, than actual date would be next day to the date selected. Thus, for any past duration selected, the prediction would be for the next day to the selected date. Thus, it can be seen that almost all selected past duration have almost similar results and its value is somewhere near to 80%. Same can be seen in Figure 17. This simulation concludes that the price difference and performance degradation increases in the case when duration between data available and prediction date increases. It have very less relation with increase in duration between pseudo date and actual date when prediction is to be done for next day in both the case. For example, let current date is 15th Aug, 2015. If duration is selected as 1 year back, i.e. 15th Aug, 2014 and if prediction is to be done for 16th Aug, 2014, it will be almost same with a similar assumption of any duration selected. But, if duration is selected for 1 year back and if prediction is to be done for actual current date, performance varies exponentially.

Comparative testing

Proposed approach was tested with similar cases with Base approach. Naïve Bayes was selected as the base approach and proposed approach as GARCH – SVM.

Table 3: Comparative Testing

Approach	**Base Approach**	**Proposed Approach**
Total Number of Articles	43	43
Number of Records where difference is greater than 10	20	12
Accuracy in %	53	72

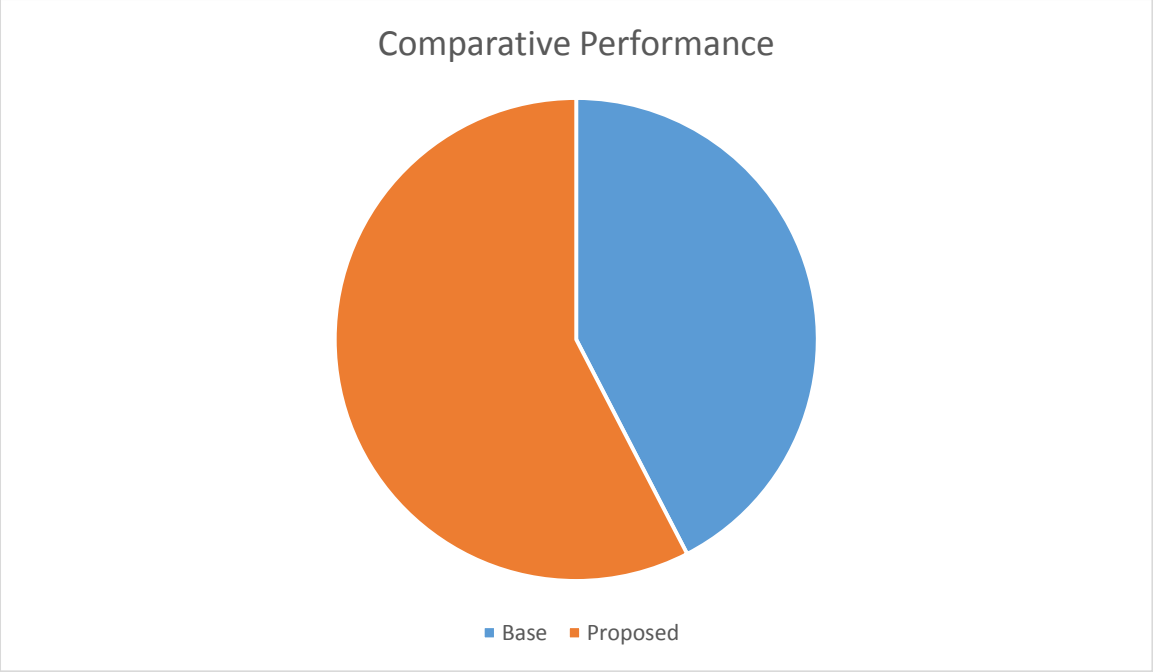


Figure 18: Comparative Performance Chart

Table 1 shows performance of proposed approach with traditional approach. Here, all the input news articles were taken as same and accuracy in prediction was calculated. Base approach was found to accurate up to 50% only, while proposed approach could provide accuracy up to 70% at least. Figure 18 shows same result in form of chart and it depicts that most of the portion of accuracy is covered by proposed approach. Thus, it concludes that proposed approach is better than base approach in terms of accuracy in prediction.

Conclusion and Future Scope

Conclusion

Artefact was developed to reduce the risk integrated in stock market. It made use of GARCH – SVM approach for prediction of stock values. Performance of the artefact was evaluated considering multiple situation and scenarios. Based on the result, it concluded few things after the artefact performance. The performance shows degradation when prediction date increase compared to current date. The performance degradation have exponential nature with respect to the increase in the duration. However, there is very little change in the performance degradation for predicting stock value for next day to the current date, with very little influence of month, quarter or season of current date. Another conclusion driven from the system is that, proposed GARCH – SVM approach is better than Naïve Bayes approach and it have performance gain of 20%. This means that proposed system is 20% better then base system. Looking to the simulation results, the performance of the artefact is much better than the expected one. Moreover, prediction of stock value is achieved up to 72% which can sure aid to mitigate the risk in stock market. This result is sufficient enough for investors and financial analyst to take decisive step pertaining to their benefits. However, it should be noted that, 72% of accuracy is a peak average and not the weighted average result. Thus, it can be said that a maximum of 72% of accuracy was achieved but still not limited to that. It also means that, it may be possible that for particular company and particular year or month, it may not succeed to achieve same percentage of accuracy.

Future Scope

For the proposed artefact, most of the opinions were selected from news articles. Thus, problems with language itself avoided. If sentiments were considered that contains mixed opinions from various sources like blogs, news feeds, tweets, it would arise the problem that

were listed in the literature survey, which are avoided for developed prototype. Thus, it can be considered to develop a generalized model which have capacity to handle such scenarios. Moreover, prediction of the stock price is done for today, whereas in proposal, it was proposed to predict stock price for any future date and time. This again could be considered as future scope. However, in that case, it may be possible that same algorithm may not work in that and must need to develop different algorithm. This assumptions are considered for the future scope. Moreover, it was concluded that proposed algorithm can achieve up to 70% of accurate prediction, thus performance tuning can be done by altering the training method to achieve even better performance. Again, this is considered as the future scope. Moreover, there are many approach that are used along with GARCH approach for prediction in time series. Using multiple combination and collective approach may be able to find even better performance for prediction accuracy. Moreover, if multiple combination of GARCH is taken, a pattern must be identified when particular approach shows better result, may be in reference to month or date or any other. Thus, learning that pattern, neuro based collective approach can be proposed that identifies the patter for individually best performance in such case and provided overall accuracy of 90% and above. Maximal Likelihood estimation used for training the model, have a disadvantage of processing complexity in multi dimension. Thus, a better suboptimal estimation is required to be proposed that can provide accurate results with less mathematical complexity. This can also be considered as future scope.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June), Sentiment analysis of twitter data, In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38), Association for Computational Linguistics.

Ahmad, K. (Ed.). (2011), *Affective computing and sentiment analysis: emotion, metaphor and terminology* (Vol. 45), Springer Science & Business Media.

Ahmad, K., Cheng, D., & Almas, Y. (2006, February), Multi-lingual sentiment analysis of financial news streams, In *Proc. of the 1st Intl. Conf. on Grid in Finance*.

Aiken, M., & Balan, S. (2011), An analysis of Google Translate accuracy, *Translation Journal*, *16*(2), 1-3.

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 28(1), 56-75.

Balk, E. M., Chung, M., Hadar, N., Patel, K., Yu, W. W., Trikalinos, T. A., & Chang, L. K. W. (2012), Accuracy of data extraction of non-English language trials with Google Translate.

Barberis, N., & Thaler, R. (2003), A survey of behavioral finance, *Handbook of the Economics of Finance*, *1*, 1053-1128.

Berndt, E. R., Hall, B. H., Hall, R. E., & Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. In Annals of Economic and Social Measurement, Volume 3, number 4 (pp. 653-665). NBER.

Bollen, J., & Mao, H. (2011), Twitter mood as a stock market predictor, *Computer*, (10), 91-94.

Bollerslev, T., & Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric reviews, 11(2), 143-172.

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. Journal of econometrics, 52(1), 5-59.

Bosco, B. P., Parisio, L. P., & Pelagatti, M. M. (2007), Deregulated wholesale electricity prices in Italy: an empirical analysis, *International advances in economic research*, *13*(4), 415-432.

Brummelhuis, R. (2008), Serial dependence in ARCH-models as measured by tail dependence coefficients. *Extremes*, *11*(2), 167-201.

Cenoz, J., Hufeisen, B., & Jessner, U. (Eds.). (2003). The multilingual lexicon. Dordrecht: Kluwer Academic Publishers.

Connor, G., Korajczyk, R. A., & Linton, O. (2006), The common and specific components of dynamic volatility, *Journal of Econometrics*, *132*(1), 231-255.

Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (pp. 507-512). IEEE.

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011, December), Combining technical analysis with sentiment analysis for stock price prediction, In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on* (pp. 800-807

Engle, R. F. (1982), Autoregressive conditional heteroscedasticity). IEEE. with estimates of the variance of United Kingdom inflation, *Econometrica: Journal of the Econometric Society*, 987-1007.

Engle, R. F., & Patton, A. J. (2001), What good is a volatility model, *Quantitative finance*, *1*(2), 237-245.

Gamon, M. (2004, August). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the 20th international conference on Computational Linguistics (p. 841). Association for Computational Linguistics.

Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 537-546). ACM.

Iqbal, F. (2013), Diagnostic checking for GARCH-type models, *Communications in Statistics-Theory and Methods*, *42*(6), 934-953.

Ivković, Z., & Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments. The Journal of Finance, 60(1), 267-306.

Ivković, Z., & Weisbenner, S. (2007). Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. Review of Financial Studies, 20(4), 1327-1357.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11, 538-541.

Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012, February), A large-scale sentiment analysis for Yahoo! Answers, In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 633-642). ACM.

Kuo, R. J., Chen, C. H., & Hwang, Y. C. (2001), An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network, Fuzzy sets and systems, 118(1), 21-45.

Lucey, B. M. (2003), Distributional aspects of Irish financial accounting ratios, *Available at SSRN 377220*.

Meijin, W., & Jianjun, S. (2004), Stock Market Returns, Volatility and the Role of Investor Sentiment in China [J], *Economic Research Journal*, *10*, 75-83.

Melville, P., Gryc, W., & Lawrence, R. D. (2009, June), Sentiment analysis of blogs by combining lexical knowledge with text classification, In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1275-1284). ACM.

Mukherjee, S., & Bhattacharyya, P. (2013), Sentiment Analysis: A Literature Survey, arXiv preprint arXiv:1304.4520.

Mukherjee, S., & Bhattacharyya, P. (2013). Sentiment Analysis: A Literature Survey. arXiv preprint arXiv:1304.4520.

Nankervis, J. C., & Savin, N. E. (2012), Testing for uncorrelated errors in ARMA models: non-standard Andrews-Ploberger tests, *The Econometrics Journal*, *15*(3), 516-534.

Nasukawa, T., & Yi, J. (2003, October), Sentiment analysis: Capturing favorability using natural language processing, In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77), ACM.

Pang, B., & Lee, L. (2008), Opinion mining and sentiment analysis, *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Poon, S. H., & Granger, C. W. (2003). Forecasting volatility in financial markets: A review. Journal of economic literature, 41(2), 478-539.

Raja, M., & Selvam, M. (2011), Measuring the Time Varying Volatility of Futures and Options, *International Journal of Applied Economics & Finance, 5*(1).

Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., & Vouros, G. A. (2009, September), Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach, In RANLP (pp. 370-375).

Rumbaugh, J., Jacobson, I., & Booch, G. (2004). Unified Modeling Language Reference Manual, The. Pearson Higher Education.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Vikmane, E., & Kreituse, I. (2009), The Impact Of The Socio-Economic Factors On The Electoral Behaviour In Young Eu Member States: Problem-Posing And Methodological Approach, The Case Of Latvia And Estonia, *European Integration Studies*, (3).

Wan, X. (2009, August). Co-training for cross-lingual sentiment classification. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 235-243). Association for Computational Linguistics.

Wang, F. A. (2001), Overconfidence, investor sentiment, and evolution, *Journal of Financial Intermediation*, *10*(2), 138-170.

Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. Systems, Man, and Cybernetics: Systems, IEEE Transactions on, 44(8), 1077-1087.

Appendix 1

Proposal Form

MSc Project Proposal Form
Academic Year:2014/2015.
Semester:3AAACIS000-6

Student Number	1330741
Student Name	Mohammed Selim
Degree Course	MSc. Business Information System
Supervisor Name	Dr.Ingo Frommholz 28/04/15
Title of Project	A Decision Support Approach for Online Stock Forum Sentiment Analysis
Description of your artefact	▪ Currently, in domain of stocks and share trading, schematic approach is followed. "Ami broker" is one of the popular tools that is used to provide buy and sell signals for online share trading. It follows a typical schematic algorithm as decision system for such signals. Many such tools are available which operate on more or less modified schematic algorithm. However, along with such tools, there are many blogs where experts who have good experience in trading, share their opinions. For example, "This time, HSBC bank is going to hit 90% of profit making share values touching sky in short period" Likewise, there are not only one, many such expert advisors post and share their opinions. People who have access to such blogs refer that and even if they refer, they have no idea on how much to rely on it. Thus, it arises two problems that make aim of this research approach. 1. Most of the expert advice is not being used and remains posted meaninglessly 2. Even if such advises are observed, people have no idea on exact reliability of advises. ▪ It makes requirement of a tool which makes use of such advices, analyse it statistically and provide help in domain of finance, share and trading. In other words, sentiment analysis tool is proposed for such problem.

Course Leader 30/4/15

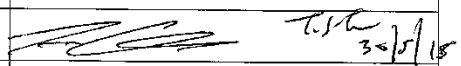
	▪ Currently, there is growth in sentiment analysis in multiple domains, but correlation between use of sentiment analysis and its use in share and trading is merely 0.1%. ▪ Aim: To develop a tool based on sentiment analysis in domain of finance, share and trading. ▪ Objectives: 1. To gather expert advice from multiple blogs and social networking sites 2. Identify language of the sentiments 3. Convent multilingual sentiments into English using "Google Translate" 4. Reform sentences in proper order in English 5. Fetch keywords to score sentence from -1, 0 or 1, where -1 indicates that something should not be done, 0 indicates neutral and 1 indicates positive, i.e. something must be done 6. Fetch names of companies using known keywords like "HSBC" and Map keywords with scores 7. Based on statistical analysis, results based on companies must be shown 8. Compare obtained results with results from schematic algorithms with justification and recommendation ▪ List of features 1. Data fetching from social networking websites 2. Support to non-English sentiments and its conversion to English 3. Ability to map companies with scores 4. Predict future of company based on sentiment analysis and statistics ▪ Added value 1. Use of expert advice and sentiments all over the world with statistical support in reliability ▪ Intellectual challenges involved 1. To gather such huge data

	2. To provide accurate score without any errors in sentimental analysis as sarcasm have highest impact on false analysis in sentiment analysis 3. Map scores to companies, where total companies are unknown and may be present in sentiments in form of spelling mistakes, leads to false analysis
What methodology (structured process) will you be following to realise your artefact?	▪ Methodology 1. Semi structured type of research methodology will be followed in following steps 2. To get introduced to sentiment analysis, its limitations and its reliability 3. To survey on existing sentimental tools, its working for different domains, also in domain of finance if available 4. Identify problems to it, or propose modification or addition of feature to it with justification of its requirement 5. Implement system as per proposal 6. Prepare test cases for system that includes, multilingual sentiments, sentiments with spelling mistakes, sentiments with irony 7. Compare results with existing schematic algorithm for all above test cases and find correlation ratio in two dimension. Cross correlation between different test cases for both proposed system and existing schematic approach. And autocorrelation for individual test case to system. This will provide two dimensional degree of analysis and correctness of result. ▪ Appropriateness and suitability of approach for realising artefact 1. Expert advice and sentiments remains unused 2. There is need of reliability in sentiments as, just on single sentiment it cannot conclude any point strongly 3. Use of sentiment analysis provides good conclusion along with reliability

How does your project relate to your degree course and build upon the units/knowledge you have studied/acquired	▪ Requires good analytical skills ▪ Requires thorough technical knowledge to develop such tools ▪ Requires wide angle thinking in terms of testing ▪ Requires adequate data related statistical knowledge	
Resources	▪ Windows OS, 4GB RAM, Dual Core Processor x64 ▪ JSP and Bootstrap CSS, Servlets, MySQL database, Tomcat 7, HTML, JDBC	
Have you completed & submitted your ethics form?	YES	

FACULTY OF CREATIVE ARTS, TECHNOLOGIES AND SCIENCE

Form for Research Ethics Projects (CATSethicsform)

1. Student Name	Mohammed Selim
2. Student Number:	1330741
3. Degree Pathway:	Msc Business Information System
4. Supervisor's name	Dr .Ingo Frommholz
5. Supervisor Signature	 30/5/15
6. Working title of project	A Decision Support Approach for Online Stock Forum Sentiment Analysis

IMPORTANT:
After the proposal form and ethics form have been signed off by both the Supervisor and Course Co-ordinator, the student must scan both signed proposal form and ethics form, then upload both of them on BREO in one file.
The original hardcopies of the proposal form and ethics from can then be submitted to the Faculty Office.
Failure to follow this process will result in the cancellation of the project and there will be no compensation for any time lost.

SECTION A Ethical Issues

Please summarise below the ethical issues involved in the research proposal and how they will be addressed. In any proposal involving human participants clear explanation of how informed consent will be obtained, how confidentiality will be observed, how the nature of the research and the means of dissemination of the outcomes will be communicated to participants must be provided.

Sentiments that are considered as inputs to the proposed system are the actual idea and suggestion of a person. Based on this sentiments, proposed tool will provide a suggestion. This sentiments will be collected from various blogs all over all internet with predefined keywords. This may include social networking as well as blogsites. Thus, will be no mechanism to acknowledge the owner of being its content used for educational purposed. Neither there will be any mechanism of requesting for permission for their content being used. As this information will be collected from publicly available sites, there will be no requirement for requesting of permission. However, ethical issue is in informing the owner for their content being used. Thus, to overcome this problem, instead of real sentiments or suggestions, dummy sentiments will be used, as purpose here is to demonstrate in form of prototype.

SECTION B Check List

Please answer the following questions by circling **YES** or **NO** as appropriate.

1. Does the study involve vulnerable participants or those unable to give informed consent (e.g. children, people with learning disabilities, your own students)?

   NO

2. Will the study require permission of a gatekeeper for access to participants (e.g. schools, self-help groups, residential homes)?

   NO

3. Will it be necessary for participants to be involved without consent (e.g. covert observation in non-public places)?

   NO

4. Will the study involve sensitive topics (e.g. obtaining information about sexual activity, substance abuse)?

   NO

5. Will blood, tissue samples or any other substances be taken from participants?

   NO

6. Will the research involve intrusive interventions (e.g. the administration of drugs, hypnosis, physical exercise)?

   NO

7. Will financial or other inducements be offered to participants (except reasonable expenses or small tokens of appreciation)?

   NO

8. Will the research investigate any aspect of illegal activity (e.g. drugs, crime, underage alcohol consumption or sexual activity)?

   NO

9. Will participants be stressed beyond what is considered normal for them?

   NO

10. Will the study involve participants from the NHS (patients or staff) or will data be obtained from NHS premises?

    NO

If the answer to any of the questions above is "Yes", or if there are any other significant ethical issues, then further ethical consideration is required. Please document carefully how these issues will be addressed.

Signed (student): [signature] Date: 28.04.2015

Countersigned (Supervisor): [signature] Date: 28/04/2015

[signature] 30/5/15

APPENDIX 2 : Artefact Code

Code For Fetching data from Internet

```
for(int k=0;k<link.length;k++)
    {
            fout = new FileWriter(path+"/"+path2+".txt");
            out1 = new BufferedWriter(fout);
            out1.write("");
            System.setProperty("http.proxyHost", "proxy.ssn.net");
            System.setProperty("http.proxyPort", "8080");
            System.out.println("Before Connecting");
                    Document doc=Jsoup.connect(link[k]).timeout(0).get();
                    System.out.println("Connected and Fetching ");
                    System.out.println("Current Link "+link[k]);

            int count1 = doc.select("div.PT18.PB18.brdb.iepb").size();
            System.out.println("COUNT "+count1);


          Element Info,Price,Time;
          for(int i=0;i<count1;i++)
          {
          if((Infodoc.select("div.PT18.PB18.brdb.iepb").get(i).select("div.info
a").first())!=null)
           {
              System.out.println("INFO : "+Info.text());
```

```
                out1.write(Info.text());
            }

if((Price=doc.select("div.PT18.PB18.brdb.iepb").get(i).select("p.gL_11.MT5.MB5").first())!
=null)
            {
                //price=Price_FILE.text();
                System.out.println("PRICE : "+Price.text());

              //out1.write(" "+Price.text());
            }

if((Timedoc.select("div.PT18.PB18.brdb.iepb").get(i).select("div.gL_11.PT10.PB10").first())
!=null)
            {
              System.out.println("TIME : "+Time.text());
              out1.write(" "+Time.text());
            }

            System.out.println();
            System.out.println();


            out1.newLine();
```

```
                out1.newLine();


              }
```

Code for sentiment Calculation

```
static int neg=0;
        static int pos=0;
        public static double sentiment;


        public static File fileExtractor(File f)
        {
                if(f.isDirectory())
                        {

                                fileExtractor(f);
                        }
                        else
                                return f;
                return f;
        }


        public static double fileHandler(String rootPath,String name)
        {
```

```
        System.out.println("Root path "+rootPath);
        System.out.println("Name  "+name);

        File root=new File(rootPath);
            File list[]=root.listFiles();
            for (int i = 0; i < list.length; i++) {
                File f=fileExtractor(list[i]);
                System.out.println("In File "+f.getName());
                if(f.getName().equals(name)){
                    sentimetCalculator(f);
                }
                else continue;
            }
            return sentiment;
}


static void sentimetCalculator(File f)
{

        System.out.println("File length "+f.length());
if(f.length()!=0)
{
```

```
                neg=0;
                pos=0;
                String str="";
                try {

                        FileInputStream fis=new FileInputStream(f);
                        BufferedReader br=new BufferedReader(new
InputStreamReader(fis));
                        /**
                         *
                         * Reading Commemts LINE BY LINE
                         *
                        */
                        while((str=br.readLine())!=null)
                        {
                        regExprNeg(f,str);
                        regExprPos(f,str);
                        }
                        calc();
                } catch (IOException e) {
                        // TODO Auto-generated catch block
                        e.printStackTrace();
                }
        }
}
```

```
public static void regExprNeg(File f,String str)
{
        Pattern pattern = Pattern.compile("sell*", Pattern.CASE_INSENSITIVE);
   Matcher matcher = pattern.matcher(str);
   // check all occurance
   while (matcher.find()) {
        neg++;
   }
}

public static void regExprPos(File f,String str)
{
        Pattern pattern = Pattern.compile("buy*", Pattern.CASE_INSENSITIVE);
   Matcher matcher = pattern.matcher(str);
   // check all occurance

   while (matcher.find()) {
        pos++;
   }
}

static void calc()
```

```
{
        System.out.println("In Calc ");
        sentiment=0.0;
        System.out.println("p "+pos);
        System.out.println("n "+neg);
        int flag=0;
        if (pos==0     &&  neg==0 ) {
        System.out.println("Neutral");
        flag = 10;
         sentiment=0;
        }

        if (pos==0 && neg > 0) {
                flag=1;
                 sentiment=neg;
                if(sentiment<-10)
                {
                        sentiment=-10;
                }
                System.out.println(" negative "+sentiment);
                }

        if (neg==0 && pos > 0 ) {
                 sentiment=pos;
                if(sentiment>10)
```

```
        {
                sentiment=10;
        }
        flag=2;
        System.out.println(" positive "+sentiment);
        }

if(flag==0)
{
        if (pos<=neg) {
                double t1=Integer.valueOf(pos);
                double t2=Integer.valueOf(neg);
                sentiment=-1*(neg/pos);
                if(sentiment<-10)
                {
                        sentiment=-10;
                }
                System.out.println("neg ratio "+sentiment);

        }else if (pos>neg) {
                double t1=Integer.valueOf(pos);
                double t2=Integer.valueOf(neg);
                sentiment=t1/t2;
                if(sentiment>10)
                {
```

```
                    sentiment=10;
                }
                System.out.println("  pos ratio "+sentiment);
            }
        }
    }
```

Code for dataset creation for historical data

```
public static String classify(double temp, double hum,String Company ) {

        //create_model();

        Instance instance;
        try {

            System.out.println(" 1 ");
            //System.out.println(temp+"\t"+hum);
            Attribute gain = new Attribute("gain");
            Attribute sentimentindex = new Attribute("sentimentindex");
            //Attribute pressure = new Attribute("pressure");
            System.out.println(" 2 ");
            FastVector fvClassVal = new FastVector(6);

    fvClassVal.addElement("Bearish_sentiment_index_tend_to_Fall_Tomorrow");

    fvClassVal.addElement("Bullish_sentiment_index_Tend_To_Increase");

    fvClassVal.addElement("Bullish_sentiment_index_No_increase_is_seen_in_Price_W
ait");
            fvClassVal.addElement("Bearish_sentiment_index_Hold_and_Wait");
```

```
                fvClassVal.addElement("Bearish_Sentimental_index_Wait");


        fvClassVal.addElement("Bullish_Sentimental_index_Rise_Can_be_seen");
                Attribute markettrend = new Attribute("markettrend", fvClassVal);
                System.out.println(" 3 ");
                FastVector fvWekaAttributes = new FastVector(3);
                fvWekaAttributes.addElement(gain);
                fvWekaAttributes.addElement(sentimentindex);
                //fvWekaAttributes.addElement(pressure);
                fvWekaAttributes.addElement(markettrend);
                System.out.println(" 4 ");
                Instances instances = new Instances("sentiment", fvWekaAttributes,
0);

                System.out.println(" 5 ");


                instance = new Instance(instances.numAttributes());
                instances.add(instance);
                System.out.println(" 6 ");
                instances.setClassIndex(2);
                System.out.println(" 7 "+Company);
                instance.setValue(gain, temp);
                instance.setValue(sentimentindex, hum);
                //instance.setValue(pressure, 37.92);
                instance.setDataset(instances);
                File f=new File("c:\\DSSA\\sentimentMODELS\\");
```

```
String s[]=f.list();
for (int i = 0; i < s.length; i++) {
        //System.out.println(s[i]);
System.out.println(s[i]+" : "+s.equals(Company)+" : "+Company);
if(s[i].trim().equals(Company.trim()))
{
        System.out.println(" 8 ");
FileInputStream fin = new FileInputStream(f+"\\"+s[i]);
ObjectInputStream io = new ObjectInputStream(fin);
Classifier classifier = (Classifier) io.readObject();


fin.close();io.close();

//      Classifier      classifier      =       (Classifier)
weka.core.SerializationHelper.read("NBP.model");

double result = classifier.classifyInstance(instance);
//System.out.println("result   "+result);
return instances.classAttribute().value((int) result);
}}

} catch (Exception e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
```

```
		}
		return null;

	}

	public static void create_model()
{
		try {
			Classifier cls = new weka.classifiers.bayes.NaiveBayes();
			FileReader fr = new
FileReader("c:\\users\\desktop\\sentimentARFFs\\SIEMENS.arff");

			BufferedReader br = new BufferedReader(fr);
			Instances inst = new Instances(br);
			inst.setClassIndex(inst.numAttributes() - 1);
			cls.buildClassifier(inst);

	weka.core.SerializationHelper.write("c:\\users\\desktop\\sentimentMODELS\\SIEME
NS.model", cls);
			fr.close(); br.close();
		} catch (FileNotFoundException e) {
			// TODO Auto-generated catch block
			e.printStackTrace();
		} catch (IOException e) {
			// TODO Auto-generated catch block
```

```
            e.printStackTrace();
        } catch (Exception e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
}
```

Code for getting data for companies from database

```
static ArrayList<String> bse_rise=new ArrayList<String>();

        static ArrayList<String> bse_fall=new ArrayList<String>();

        static ArrayList<String> nse_rise=new ArrayList<String>();

        static ArrayList<String> nse_fall=new ArrayList<String>();



        public static ArrayList<String> getBse_rise() {
                return bse_rise;
        }



        public static void setBse_rise(String c) {
                bse_rise.add(c);
        }



        public static ArrayList<String> getBse_fall() {
```

```
        return bse_fall;
}

public static void setBse_fall(String c) {
        bse_fall.add(c);
}

public static ArrayList<String> getNse_rise() {
        return nse_rise;
}

public static void setNse_rise(String c) {
        nse_rise.add(c);
}

public static ArrayList<String> getNse_fall() {
        return nse_fall;
```

```
}


public static void setNse_fall(String c) {
        nse_fall.add(c);
}
```

Code to connect to Database through Java Code.

```
public  Connection con;
        public Connector() throws Exception
        {

        System.out.println("In Connecter");
        /**
         * Register the driver
        */
        Class.forName("com.mysql.jdbc.Driver");

/**
         *Create Connection
        */
```

```
con=DriverManager.getConnection("jdbc:mysql://localhost:3306/features","root","root");
        st=con.createStatement();


        System.out.println("Conn2");
        }
```

Code to Register New User to System and check login for Registered User

```
public boolean registerDB(String newUserName,String pass) throws SQLException
        {
                System.out.println("In register DB..");
                rs=st.executeQuery("select * from user where
user_name='"+newUserName+"'");

                if(rs.next())
                {
                        /**
                         * if user already Exists..
                        */return true;
                }
                else
                {
                        /**
```

```
                * If New User for Registration
            */
                Connection
con1=DriverManager.getConnection("jdbc:mysql://localhost:3306/features","root","root");

                System.out.println(getMaxID());

                String query="insert into user values
('"+(getMaxID()+1)+"','"+newUserName+"','"+pass+"')";
                try{
                java.sql.PreparedStatement ps1=con1.prepareStatement(query);
                    ps1.executeUpdate();
                }catch(Exception e){System.out.println(e);}

                System.out.println("success..");
            }

            return false;
        }

        /**
         * check for user if its a authentic user
        */
        public boolean checkLogin(String user,String password) throws SQLException
```

```
        {
                System.out.println("user name : "+user);
                System.out.println("user name : "+password);
                rs=st.executeQuery("select * from user where user_name= '"+user+"'  and password='"+password+"'");
                if(rs.next())
                {
                        return true;
                }
                else return false;}
```

Code To get Company data from database.

```
public static ArrayList<String> getDatabaseMetaData(String to) throws Exception
  {
                Connection conn = null;
     try {
        conn=new Connector().con;
        DatabaseMetaData dbmd = conn.getMetaData();
        String[] types = {"TABLE"};
        ResultSet rs = dbmd.getTables(null, null, "%", types);
        while (rs.next()) {
                String bse=rs.getString("TABLE_NAME");
                if(bse.endsWith("bse"))
                {
                        BSE.add(bse);
```

```
        System.out.println("Company : "+bse);
                }
                else
                {
                        if(bse.endsWith("nse"))
                {
                        NSE.add(bse);
                System.out.println("Company : "+bse);
                }

                }
    }
}
    catch (SQLException e) {
    e.printStackTrace();
}
    finally
    {
            conn.close();
    }

    if(to.equals("bse"))
            return BSE;
    else
            return NSE;
```

```
}
```

Code to insert fetched Stock Market Data into database.

```
public static void updateRecentStockValuesBSE(String company,double bse_curr,double
bse_open,double gainORloss,double gainORloss_per, String bse_remark,double sentiment)
        {

                System.out.println("Company "+company);
                System.out.println("BSE CURR "+bse_curr);
                System.out.println("BSE OPEN "+bse_open);
                System.out.println("GAIN/LOSS "+gainORloss);
                System.out.println("GAIN/LOSS% "+gainORloss_per);
                System.out.println("REMARK "+bse_remark);
                System.out.println("********");
                Calendar cal = new GregorianCalendar();
            Date creationDate = cal.getTime();

            SimpleDateFormat date_format = new SimpleDateFormat("dd-MM-yyyy");
            SimpleDateFormat time_format = new SimpleDateFormat("HH:mm");
            String dt=date_format.format(creationDate);
            String tm=time_format.format(creationDate);
            System.out.println(date_format.format(creationDate));
```

```
System.out.println(time_format.format(creationDate));
    try {


                //,Sentiment_value varchar(5)
            String createTable="create table IF NOT EXISTS "+company+"_BSE"+" (date_of_REC varchar(15),time_of_REC varchar(10),bse_curr double,bse_open double,gainORloss double,PER_gainloss double,remark varchar(10),sentiment double) ";
            Connection con=new Connector().con;

            Statement st=con.createStatement();
            if(st.executeUpdate(createTable)>0)
            {
                System.out.println("table created..");
            }
            else
            {
                System.out.println("Table Not Created");
            }
            String insert_query="insert into "+company+"_BSE"+" values(?,?,?,?,?,?,?,?)";
            PreparedStatement ps=con.prepareStatement(insert_query);
```

```
                ps.setString(1,dt);
                ps.setString(2, tm);
                ps.setDouble(3, bse_curr);
                ps.setDouble(4, bse_open);
                ps.setDouble(5, gainORloss);
                ps.setDouble(6, gainORloss_per);
                ps.setString(7, bse_remark);
                ps.setDouble(8, sentiment);
                if(ps.executeUpdate()>0)
                {
                        System.out.println("inserted ");
                }
                else {
                        System.out.println("NOT Inserted ");
                }

        } catch (Exception e) {
                // TODO Auto-generated catch block
                e.printStackTrace()}}}
```

Code to get analysis by base approach.

```
public static String getAnalysis(double d,double s)
{
        String analysis="";
```

```
if ( s > 4 && d >= 0.05 ) {

        analysis="Bullish Sentiment good variance count Price Can Increase..!  ";

} else if(s > 4 && d < 0.05){

        analysis="Bullish Sentiment Falling variance, Price might Fall..!   ";

}else if((s <=4 && s >=-1 ) && d >= 0.05){

        analysis="Bearish Sentiment Good variance count Tend To Fall ..!  ";

}else if(s <=-1 && d < 0.05)
{
        analysis="Bearish Sentiment No Rise Might Be Seen...!  ";
}else if(s <=-1 && d > 0.05)
{
        analysis="Bearish Sentiment Fall Might Be Seen...!  ";
}
else  if(s >=-1 && s<= -4)
{
        analysis="Bearish Sentiment Price Might Fall  ";
}else
{
        analysis="Bearish Anlysis";
```

```
        }


        return analysis;
}
```


Code to calculate Variance.

```
public static ArrayList var_coll(LinkedHashMap<String,Integer> hm)
        {
        //System.out.println(hm);
        Set s=hm.entrySet();
                ArrayList<String> pair=new ArrayList<String>();
        Iterator<?> it=s.iterator();
        while(it.hasNext())
        {
        Map.Entry<String,Integer>  me=(Entry<String, Integer>) it.next();



        String query="select bse_open,gainORloss from acc_bse where date_of_REC=?";
        try {

                Connection con=new Connector().con;
                PreparedStatement ps=con.prepareStatement(query);
                ps.setString(1, me.getKey());
                ResultSet rs=ps.executeQuery();
                int i=0;
                double opening=0.0;
                double raise=0.0;
                double closing=0.0;
                while(rs.next())
                {
                        if(i==0)
```

```
                {
                        System.out.println("i==0");
                        opening=(Math.round(rs.getDouble(1))*100)/100;;
                        System.out.println("bse_open : "+opening);
                        System.out.println("gainORloss : "+rs.getDouble(2));
                }
                else if(i==(me.getValue()-1))
                {
                        System.out.println("i:::: "+(me.getValue()-1));
                        System.out.println("bse_open : "+rs.getDouble(1));
                        raise=rs.getDouble(2);
                        System.out.println("gainORloss : "+raise);
                        closing=(Math.round(opening+raise)*100)/100;
                        System.out.println("CLOSING price : "+closing);
                        pair.add(me.getKey()+"#"+opening+"#"+closing+"#");
                //      System.out.println("OPEN      CLOSE      :(MAP      )
"+op_close_prices);
                }
                i++;

        }


    //System.out.println(pairs);
```

```
        } catch (Exception e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
        }

        }
        return varCalculation(pair);}


static  ArrayList varCalculation(ArrayList<String> p)
        {
        ArrayList<String> al=new ArrayList<String>();
        try {
                Connection con=new Connector().con;

        LinkedHashMap<String , Double> lhm=new LinkedHashMap<String, Double>();

                String unv="";
for(int i=0;i<p.size();i++)
{
        if(i==0)
        {
        String s=p.get(i);
        String temp[]=s.split("#");
```

```
            String s1=temp[0];
            double op=Double.parseDouble(temp[1]);
            double cl=Double.parseDouble(temp[2]);
            double dR=Math.log((cl/cl));

            System.out.println("Variance For Date  "+s1+" :  "+dR);

            String sss=s1+"#"+dR;
            al.add(sss);
            unv=s;
}
else
{
            String prev[]=unv.split("#");
            String s_P=prev[0];
            double op_p=Double.parseDouble(prev[1]);
            double cl_p=Double.parseDouble(prev[2]);

            String s=p.get(i);
            String curr[]=s.split("#");
            String s_c=curr[0];
            double op_c=Double.parseDouble(curr[1]);
            double cl_c=Double.parseDouble(curr[2]);

            double dr=Math.log((cl_c/cl_p));
```

```
                System.out.println("Variance For Date  "+s_c+" : "+dr);
                /*String sql="insert into temporary value('"+s_c+"','"+dr+"')";
                Statement st=con.createStatement();
                st.execute(sql);*/
                lhm.put(s_c,dr);
                String sss=s_c+"#"+dr;
                al.add(sss);
                unv=s;
        }
        }//for return al;}
```

Code for generating graph for all prediction

```
private static CategoryDataset createDataset() {

    // row keys...
    String series1 = "base";
    String series2 = "contri";
    String series3 = "realtime";

    // column keys...
    String category1 = "Category 1";
    String category2 = "Category 2";
    String category3 = "Category 3";
    String category4 = "Category 4";
```

```
String category5 = "Category 5";


// create the dataset...
DefaultCategoryDataset dataset = new DefaultCategoryDataset();
String com=company;//'30:03:2015', 'tataconsultancyservices', 65, 35, 41
ArrayList<String> dates=new ArrayList<String>();
ArrayList<Integer> base=new ArrayList<Integer>();
ArrayList<Integer> cntr=new ArrayList<Integer>();
ArrayList<Integer> realtime=new ArrayList<Integer>();


        try {
            String sql="select date,baseapp,cntr,realtime from predictioninfo
where company='"+com+"'";
            Connection con=new Connector().con;
            Statement st=con.createStatement();
            ResultSet rs=st.executeQuery(sql);
            while(rs.next())
            {
                dates.add(rs.getString(1));
                base.add(rs.getInt(2));
                cntr.add(rs.getInt(3));
                realtime.add(rs.getInt(4));


            }
            System.out.println(dates);
```

```
                System.out.println(base);
                System.out.println(cntr);
                System.out.println(realtime);
            } catch (Exception e1) {
                // TODO Auto-generated catch block
                e1.printStackTrace();
            }

    for(int i=0;i<dates.size();i++)
    {
       dataset.addValue(base.get(i), series1, dates.get(i));
       dataset.addValue(cntr.get(i), series2, dates.get(i));
       dataset.addValue(realtime.get(i), series3, dates.get(i));
    }


    return dataset;}
```

Code for today's prediction graph

```
private static CategoryDataset createDataset() {

        // row keys...
        String series1 = "base";
        String series2 = "contri";
        String series3 = "realtime";
```

```
// column keys...
String category1 = "Category 1";
String category2 = "Category 2";
String category3 = "Category 3";
String category4 = "Category 4";
String category5 = "Category 5";

// create the dataset...
DefaultCategoryDataset dataset = new DefaultCategoryDataset();
String com=company;//
ArrayList<String> dates=new ArrayList<String>();
ArrayList<Integer> base=new ArrayList<Integer>();
ArrayList<Integer> cntr=new ArrayList<Integer>();
ArrayList<Integer> realtime=new ArrayList<Integer>();

        try {
            String today=new SimpleDateFormat("dd:MM:YYYY").format(new Date());
            String sql="select date,baseapp,cntr,realtime from predictioninfo where company='"+com+"' and date='"+today+"'";
            Connection con=new Connector().con;
            Statement st=con.createStatement();
            ResultSet rs=st.executeQuery(sql);
            while(rs.next())
```

```
                {
                        dates.add(rs.getString(1));
                        base.add(rs.getInt(2));
                        cntr.add(rs.getInt(3));
                        realtime.add(rs.getInt(4));

                }
                System.out.println(dates);
                System.out.println(base);
                System.out.println(cntr);
                System.out.println(realtime);
            } catch (Exception e1) {
                // TODO Auto-generated catch block
                e1.printStackTrace();
            }

for(int i=0;i<dates.size();i++)
{
   dataset.addValue(base.get(i), series1, dates.get(i));
   dataset.addValue(cntr.get(i), series2, dates.get(i));
   dataset.addValue(realtime.get(i), series3, dates.get(i));
}

return dataset;
```

}

APPENDIX 3 : POSTER