UNIVERSITY OF BRIGHTON

DEPARTMENT OF COMPUTING, ENGINEERING AND MATHEMATICS

THESIS

A Stochastic Framework to Fuzzy Environments through Case-based Reasoning and the Inverse Problem Methodology: An Investigation of Financial Bubbles

Author

FRANCIS EKPENYONG

Supervisors:

Dr Stelios Kapetenakis

Dr. Georgios Samakovitis

Prof. Miltos Petridis

October 8, 2021

Declaration

I certify that this work has not been accepted in substance for any degree, and is not concurrently being submitted for any degree other than that of the Doctor of Philosophy (Ph.D.) being studied at the University of Brighton. I also declare that this work is the result of my own investigations except where otherwise identified by references.

Acknowledgements

This research would not have been possible if not for the help of friends, families and people too numerous to mention.

I would like to express my profound gratitude to my academic supervisors Stelios Kapetenakis, Georgios Samakovitis and Miltos Petridis for their impeccable supervision, that despite their busy schedule managed to allocate time to read and comment critically on my thesis, their continuous encouragement, guidance and immense supports from inception made the completion of this thesis possible. I have to extend my thanks to all PhD students who helped and motivated me to keep going in my research. I would like to thank my family, none of this would have been possible without their love and encouragements, my children, Bernice, Derick and Anita and Davis. Most importantly is the support and understanding that Mary, my adorable wife gave me throughout the PhD candidature, her support sustained me through the stress and frustration or this research. My innermost gratitude goes to my brothers and sisters, Simon, Moses, Basil, Elizabeth and Patricia who are always there for me, I want to also acknowledge my in-laws, cousins and other relatives for their continuous support towards the success of this PhD thesis.

Dedication

I dedicate this thesis to God almighty for His grace upon my life and to my late parents, Hon. Barnaby Ekpenyong and Mrs Agnes Ekpenyong for their immense sacrifice in making me what I am today. Also to my supervisor and great mentor, Prof. Miltos Petridis of blessed memory whose immense contribution set the pace for this work.

Abstract

This thesis presents an approach to support the identification and potential investigation of abnormal asset performance in traded securities, often popularised as 'financial bubbles'. It uses an ensemble technique based on Case-based Reasoning (CBR) and Inverse Problems techniques (IPTs) that leverages capabilities of (CBR) in classification/prediction tasks especially in fuzzy domains and Strengths of (IPTs) in reconstructing ill-posed problems and identifying cause-effect relationships, to describe and model abnormal stock market fluctuations (often associated with asset bubbles) in time series datasets from historical stock market prices.

The thesis has developed and implemented a Machine Learning formative strategy called "IPCBR Framework" which is aimed to determine the causes of stock behaviour, rather than predicting future time series points in such fuzzy environment. By so doing, the research contributes to more robust strategies in investigating financial bubbles.

The IPCBR framework uses a rich set of past observations time series, and a geometric pattern description, and applies a combination of clustering techniques to derive a model that generalizes those patterns onto observations in the forward problem formulation. The derived result is then used as the input to the Inverse Problem formulation, the process of which is used to identify set of parameters that can statistically be associated with the occurrence of the observed patterns. The combined results is adapted to complete the CBR cycle.

This framework was implemented through the use of alliance machine learning methods.

The results of the implementation have shown that case retrieval accuracy can be achieved

using a simple yet efficient approach that is based on assortment algorithms.

The thesis has demonstrated that, the Inverse solution can be successfully integrated into the CBR cycle, and that, given a target problem, the IPCBR framework provides a computationally inexpensive description of abnormal asset performance.

The research has contributed significantly to knowledge in various ways; a novel and more effective case representation of time series data has helped in solving the case retrieval issue since the time series do not perform well with the traditional attribute value representation in the CBR.

Secondly, the use of alliance machine learning methods has demonstrated case retrieval accuracy can be achieved using a simple yet efficient approach that is based on assortment algorithms. This in effect steer the course of producing the optimum combinations of classifiers through combining the strength of individual classifiers to arrive at an accurate classification. That, given the target problem, the IPCBR framework provides a computationally inexpensive description of abnormal asset performance.

Another major contribution to knowledge is in successful adaptation of the Sentiment analysis results as input to the CBR adaptation phase. Thais has proven to enhance CBR's effectiveness in pattern matching.

This thesis has also demonstrated that the IPCBR framework can be successfully applied to the financial domain, which brings a novel perspective to the problem of asset bubbles investigation.

Contents

1 Introduction **1**

1.1 Aim of the Research: 4

1.2 Research Questions 5

1.2.1 Research Question 1 5

1.2.2 Research Question 2 5

1.2.3 Research Question 3 6

1.2.4 Research Question 4: 6

1.3 Objectives 7

1.4 Research Methodology 8

1.4.1 Stage 1 9

1.4.2 Stage 2 10

1.4.3 Stage 3 10

1.4.4 Stage 4 10

1.5 Methodology outline 11

1.6 Background Information on the Dataset 12

1.6.1 Drill operation dataset 12

1.6.2 Respiratory disorder dataset 12

1.6.3 Stock dataset 12

1.7 Research Contributions to Knowledge 13

1.8 Delimitation and scope of the research 14

1.9 Publications and research activities 15

2 Literature Review **18**

2.1 Introduction . . . 18
2.2 Background . . . 18
2.3 Review on Bubbles . . . 19
2.3.1 Tulipmania . . . 20
2.3.2 The South sea Bubble . . . 20
2.3.3 Dot-Com Bubbles . . . 21
2.3.4 Housing Bubble . . . 22
2.4 Financial Markets/Stock exchanges . . . 22
2.5 Stochastic Processes in Financial Markets . . . 23
2.5.1 Proportional Returns . . . 23
2.5.2 Log-return . . . 24
2.6 Bubble: Representation . . . 24
2.7 Characteristics of bubbles . . . 27
2.8 Case-based reasoning (CBR) . . . 29
2.8.1 Case Representation . . . 31
2.8.2 Case Similarity . . . 34
2.8.3 Adaptation . . . 34
2.9 Essence of Case-based Reasoning(CBR) . . . 34
2.10 Inverse Problems . . . 36
2.10.1 Inverse Formulation . . . 36
2.10.2 The Forward and Inverse Problems . . . 38
2.11 Methods of solving the Inverse Problems . . . 40
2.11.1 Least-squares Method . . . 41
2.11.2 Trial and Error Method . . . 41
2.11.3 Heuristics Approach . . . 42
2.11.4 Artificial Neural network (ANN) . . . 42
2.11.5 The Inverse Solution through Sentiment Analysis . . . 43

2.12 Case-based Reasoning and Inverse Problems 46
2.13 Concluding Remarks 48

3 Methodology **49**
3.1 Introduction 49
3.2 Proposed Framework 49
3.3 The Forward Problem 51
3.3.1 Case Formulation 51
3.4 The Inverse Phase 56
3.4.1 Data pre-processing 56
3.5 Chapter summary 57

4 Case Retrieval **58**
4.1 Introduction 58
4.2 Distance(Similarity) 58
4.2.1 Similarities in Time Series 59
4.3 Categories of Distance Measures 60
4.3.1 Minkowosi Distance 61
4.3.2 Hamming distance: 62
4.3.3 Euclidean distance 62
4.3.4 Cosine similarity 63
4.3.5 Pearson correlation distance 63
4.3.6 Longest Common Subsequence Similarity 64
4.3.7 Dynamic Time Warping 65
4.3.8 Principal Component Analysis (PCA) 67
4.3.9 Singular Value Decomposition 68
4.3.10 Fourier Transformation (FT) 69
4.3.11 Wavelet Transformation 69
4.3.12 Piecewise Linear Approximation 70

4.3.13 Piecewise Aggregate Approximation(PAA) 71
4.3.14 Adaptive Piecewise Constant Approximation 71
4.3.15 Symbolic Aggregate Approximation 71
4.3.16 Recurrent Neural Network (RNN) 72
4.3.17 Long Short-Term Memory model(LSTM) 72
4.4 Summary 73

5 Machine learning Interpolation Methods 75
5.1 Introduction 75
5.2 Interpolation Methods 75
5.3 K-nearest-neighbor KNN 76
5.4 Logistic regression 77
5.5 Support Vector Machine (SVM) 77
5.6 Linear Discriminant Analysis 79
5.6.1 Experimental Dataset 79
5.6.2 Data Pre-processing 80
5.7 Results and Evaluation 82
5.7.1 Testing with Data from a clinical trial comparing two treatments for a respiratory illness 84
5.7.2 Conclusions and future work 86

6 The Forward and Inverse Solution 88
6.1 Case Retrieval 88
6.1.1 Experimental dataset (Stock dataset) 88
6.2 Case Adaptation for Sentiment Analysis 96
6.2.1 Corpus Collection 97
6.2.2 Tweets Preprocessing and Cleaning 98
6.2.3 Imbalanced Data 99
6.2.4 Model building 101

6.2.5 Model Metrics and Evaluation 102
6.2.6 Model Performance 102
6.2.7 Results 103
6.2.8 Stock movement correlation 109
6.3 Validation and Evaluation 112
6.3.1 Summary 116

7 Research Summary and conclusion **117**
7.1 Research Overview 117
7.2 Research Contributions to knowledge 121
7.3 Limitations of the study 122
7.4 Future Research Directions 124

1 Introduction

Financial bubbles and crashes have long been dominant paradigms in economic history, causing major concerns for both policy makers and investors (Protter, 2016). Despite considerable efforts on bubble predictions, finding a universally acceptably empirical approach remains elusive. Bubbles are significant growths in the market that are not based on anything substantial, they usually escalate with no clear warning and later vanish. Studies have shown that these bubbles and crashes often result in recurring financial crises.

Interestingly, while majority of people suffer in the wake of a market burst, there are also reports of few people that are able to see the bubble forming in time to benefit from the burst that follows. Although each bubble instance varies in their commencement and certain details, there exist some level of similarities and patterns that reasonable assumptions could be derived from, which is part of what this study is about.

Detecting bubbles in real time is quite challenging (Contessi and Kerdnunvong, 2015), partly because the fundamental value is hard to ascertain, and partly due to the fact that a slight change to the expected growth rate could generate different fundamental earnings. An enormous and increasing number of papers propose methods of detecting asset bubbles (Kubicová and Komárek, 2011; Jiang et al., 2010; Jenny Freeman, 2018; Katja Taipalus, 2012). However, the dynamic nature of the market has made it difficult to have a clear prediction from the available brief literature, besides many machine learning algorithms have evolved over the years in an attempt to predict bubbles (Ince, 2014), with promising outcomes. Creating a more efficient and effective system can gives investors a competitive advantage over others as they can identify stocks with good performance with minimum

efforts as evident form the work of Dvhg et al. (2016).

Although predicting in a scenario of this nature is characterised by high level of uncertainty and much more, doing it in such a complex environment is ambitious, it is therefore imperative to build a foundation for methodological frameworks that will enhance our understanding of the root causes of asset valuation bubbles, and, in the long run, help markets identify and mitigate them.

Building of such financial model is capable of providing insightful and useful advice to business managers. Such models are designed to simulate how the system evolves over a time period, which traditionally follows a forward process.

While a model of this nature is beneficial to the decision managers by way of enabling them gain knowledge to the real market representation, and in effective and efficient design of an improved system, the decision makers often confronts the inverse problems of the original model. In a situation where the derivative of the original problem is a representation of the real market, the primary concern of the decision managers would therefore be to determine the optimal inputs of the model that will produce the given outputs. It is often common to introduce some constraints dynamically to the outputs while searching for the inputs that produced these sets of outputs. Solving the inverse constraints directly is computationally expensive and also often requires the use of different computational models which makes the process difficult, in some cases impossible to achieve. In such cases, the managers may resort to "trial and error"; by iteratively running the original model over and over and observing the variations in outputs and modifying the inputs accordingly for another run. These observations could then be added to the original data and stored as historical data as well that could be further used across board in making informed decisions. The data can be stored in form of predicate as

$$P(I_1, I_2, ...I_j; O_1, O_2, ..O_k)$$

where "I" refers to the various inputs and "O" the outputs.

Now, the question that arises from this is,

given some outputs patterns, O that is defined as a bubble, can a decision managers find the right sets of inputs, I which represent various indicators that cause these bubbles? (O being a function of I), or in a situation where a perfect match could not be found, can a near solution to the target be good enough to be adapted?

This work brings a solution to the above problem through the use of Case Based Reasoning (CBR) and the Inverse Problems (IP) which is termed the IPCBR framework. The CBR methodology is capable of solving problems by utilizing the specific knowledge acquired on the previously encountered problems and solved situations called cases (Aamodt and Plaza, 1994; Ince, 2014; H. and Elmogy, 2015). IP is a methodology employing process output information to recommend suitable input settings for the process concerned.

This ensemble, IPCBR framework has a particular application in financial domain where such models are used. And the use of such ensemble is necessary to build the foundation for methodological frameworks that can contribute to our understanding of how asset bubbles are created, hence potentially help in mitigating their occurrence.

Several successful CBR systems have been developed in various application fields; Industrial safety control (Su et al., 2019), Phishing detection (Abutair and Belghith, 2017), Production Planning (Seo et al., 2007), in accounting research and practice (Kapetanakis, Samakovitis, and Gunasekera, Kapetanakis et al.), and in engineering (Shokouhi et al., 2014). As sophisticated and user-friendly CBR shells emerge, CBR is more likely to be used to solve complex, real-world problems. This emerging and fast growing methodology has made significant contributions to the task of making predictions in a business process (Kapetanakis et al., 2010), it can be applied to a business process as a support to knowledge transfer as most of the previously developed methodologies lack actual guidance on the process design, thus, threatening the success of Business Process Relations (BPR) (Mansar and Marir, 2003).

Despite the numerous successes recorded through this methodology (Case-based reasoning), it is being faced with various challenges, some of which are the retrieval effectiveness

and sparse cases. Although various methods have been applied to address these challenges, the issues have not still been fully addressed as various operations are carried out in different domains.

In this approach, a suitable knowledge representation and similarity measures is developed and tested against a range of Machine Learning (ML) algorithms and the most suitable one chosen, then the Sentiments analysis is applied on the derived model to create a more effective distribution of cases across the model which will generate sufficient case for the knowledge-base. The results of which will then be used as a case base for standard Case-based Reasoning process and will be evaluated against a known episodic (real) data and human expert advice.

The Inverse problem theory that is embarked upon in this research work is confined to observations and questions that can be scientifically denoted. The observations of the world will be described with the real market data (prices, volumes, times, fundamentals). These properties will be referred to as the "model parameters". In the course of this research, we shall, by assumption, derive some specific method (usually a mathematical theory or model) for relating the model parameters to the data. The reasons being that the Inverse problem theory always demands that the physical model be specified beforehand. According to Ritter et al. (2003) "The success of an inverse parameter determination depends on how well the problem can be posed".

1.1 Aim of the Research:

The aim of this research is to enhance the capability for investigating stock behaviour that is often linked to asset bubbles, through a formative framework that uses a combination of Case-based Reasoning and Inverse Problem techniques.

1.2 Research Questions

This research is set to address the questions:

1.2.1 Research Question 1

> Can we have efficient representation of fluctuations using CBR?

Events of market crashes have entranced economists for centuries. Although many researchers have studied questions connected to collapsing asset price bubbles, there is not much consensus yet about their causes and effects. Advancing this search for a plausible solution in using the CBR, can we have a good representation of a typical bubble as a case in our knowledge base? This question arose because a typical case representation follows the attribute-value pattern which is not in line with the time series data of a bubble that is characterised to fuzziness and complexities.

1.2.2 Research Question 2

> Is it possible to identify effective similar measures of cases represented as bubble?

Although similarity measure is a common data mining task, it is subjective in real sense, extremely dependent on the domain and application. This measure has been a major research topic in time series data mining for more than a decade, which gives birth to various measures (Finnie and Sun, 2002; Seo et al., 2007; Salvador and Chan, 2018). Despite the existence of various methods of measuring similarity, the challenge of determining the best method for assignment of attributes' weight value in CBR still remains vague (Ji et al., 2010). Therefore in order that one can establish a match to some possible trends, the need to determine a proper measure of similarity between two periods has shown to be crucial.

1.2.3 Research Question 3

> Can we identify correlations between the historic period of fluctuations and stock sentiments using the Inverse Problem Methodology?

This is where a short window event is being carried out around the periods of fluctuations based on the results obtained from the forward problem. Information contained in News and tweets are harvested and sentiment analysed from the theory of Inverse Problems is performed, with the aim of finding some correlations if any between the short window events and the stock movement.

1.2.4 Research Question 4:

> Can the Inverse problem methodology enhance CBR's effectiveness in identifying patterns associated with bubbles?

Decision making in the financial sector is a crucial task for investors in which intricate decision-making problems are to be quickly resolved so as to minimize investment risks. The financial domain appears to be cloudy, making it difficult to articulate knowledge in the form of rules. This question arises because the concept of CBR is yet to be fully utilized in financial domain like the asset bubble that presents some unaccustomed features that is not supported by the traditional methodology. So, can the proposed framework successfully provide a reasonable explanation to a given machine tool selection problem in situations in which decision-makers have little information about the problem and are incapable of creating beliefs about the possible results and their probabilities.

Case-Based Reasoning techniques have been extensively used and have proven to be an effective problems-solving tool in the recent times.

Despite its extensive use, traditional approaches still focus on identifying similar problems without exploring the underlying information of subject during the problem-solving process, thus generating a growing interest towards integrating it with other reasoning paradigms.

Due to the recorded success of the Inverse Problems in engineering field, and characteristic easy use of a CBR system, an intuitive idea was proposed in this study that the Inverse Problem technique could be incorporated into a CBR cycle to arrive at a more robust knowledge system.

This became the driving force towards building the ensemble that uses the achievements of Case based reasoning and the Inverse Problem, and to assess their usefulness as business aids.

The ensemble methods were developed with reference to specific complete examples using historical stock data, and to guarantee that the vital detail was not ignored.

1.3 Objectives

To achieve the aim of this research, the following objectives have been set up as a guide to the study:

1. to perform extensive literature review on various Case-based reasoning applications and Inverse Problems with aim to propose a framework for the identification of financial bubbles.

2. to determine a suitable representation of time series for effective shape comparisons.

3. to study and identify which Machine Learning algorithms will be more suitable for identifying similar patterns in a bubble.

4. to determine correlations between sentiment analysis and stock indicators using combinational techniques to uncover causatives on the fluctuations.

5. to evaluate the CBR/IP framework through the use of historical stock data.

1.4 Research Methodology

In this section, the emphasis is mostly on the methodology that is being taken in the course of this research. The principle for each aspects of the work is elaborated, and the role in supporting the rationale of the research work is clearly presented. However, a more detailed methodology regarding the experimental set-up is presented subsequently in chapter 3.

Figure 1.1 depicts a graphical representation of the research outline, which is being segmented into four stages:

1. preliminary background research and methodology development
2. Design and development of the approach
3. Implementation and evaluation
4. Contribution to knowledge and future research direction

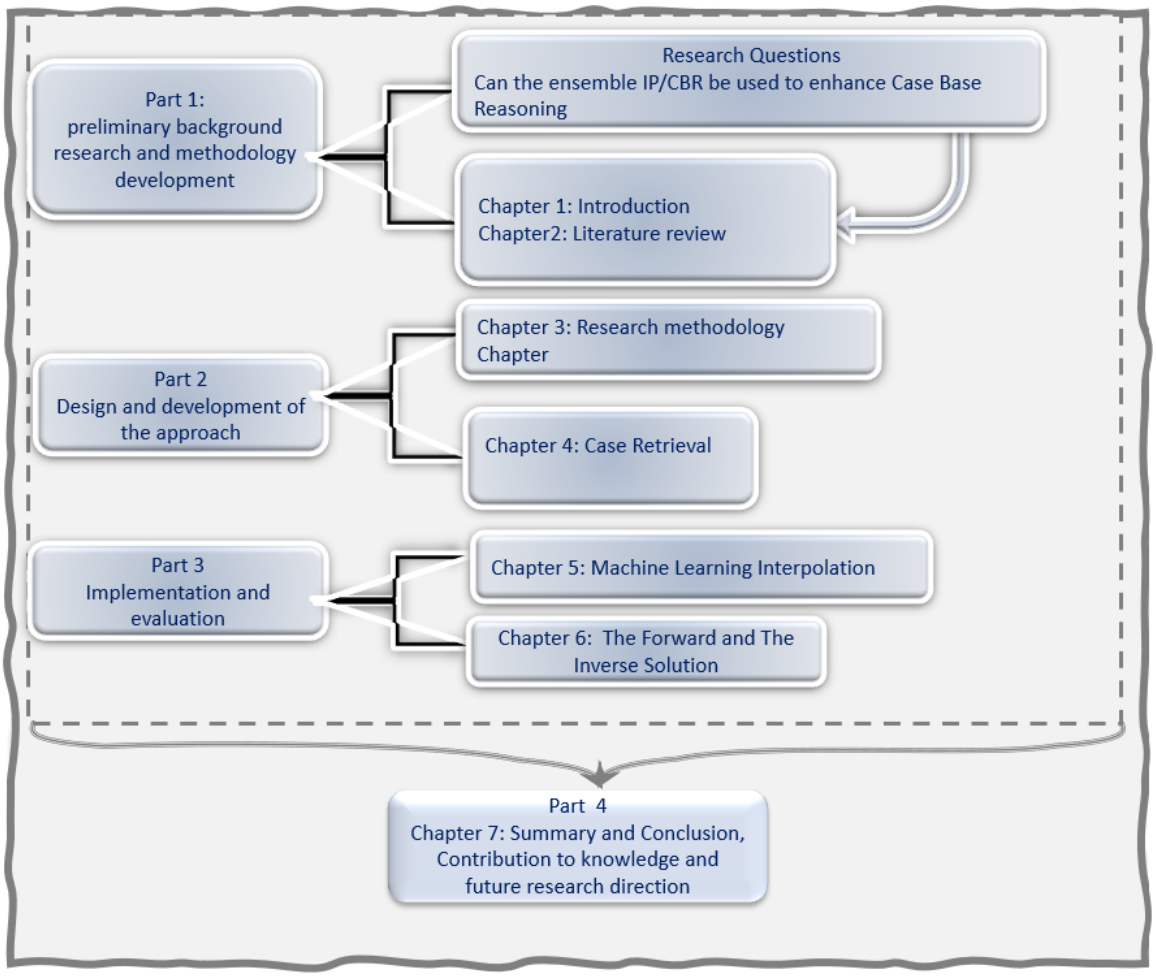

Figure 1.1: Roadmap of the research

1.4.1 Stage 1

The first task was to identify the core concern of this research and then carry out a more in-depth review on the past research relevant to this work. The main objective of doing this was to gather sufficiently useful background information as it is essential to identifying past and current research directions in the field and ascertain that the approach chosen will provide useful contribution to knowledge.

In view of this, it was therefore useful to explore in detail, financial market related business processes, Inverse Problems, and Case-based reasoning researches from a variety of literature for a more concrete definition of the research problem space. In addition to this,

it was also useful to identify the merits of using the Case-based Reasoning in some business models and to consider the specific problem to be addressed in using the Case-based reasoning methodology.

1.4.2 Stage 2

Sequel to a successful completion of the first stage, and with a clear definition of the research problem , the progression was to solve the forward problem and later examine if the inverse of the forward solution can be successfully derived. In view of this, a knowledge mining model was created, starting from developing a generic model which covers relevant information on historical stock transactions including the applied results.

In order to achieve this, it was again necessary to further review some relevant literature to ensure that the proposed methods for solving the sub-problems are original as well as searching possible existing work that might shed more lights on these problems.

1.4.3 Stage 3

After a successful conclusion of step two, with the forward problems solved and the derived inverse model to predict sufficient case for the knowledge base, it was then imperative to examine if the proposed methodology was able to answer the primary research questions.

The potentials of the IPCBR approach was tested on real episodic data to ascertain how efficient the framework is able to identify the underlying patterns using clustering method since this problem falls under unsupervised learning. The results were tested using available metrics to ensure that it would aid the real design problem, which would have ordinarily required considerable expertise and time.

1.4.4 Stage 4

With the recorded success of the new hybrid framework with respect to the experimented business applications, it was worthwhile to explore the potential application scope of the

framework in a wide perspective.

1.5 Methodology outline

In summary, the ensemble in made up of forward phase and the Inverse Problem phase. The idea was achieved firstly by solving the simplified forward phase through the case-based reasoning. This was done by clearly defining the observations of patterns that is classified as abnormal fluctuations which is otherwise termed “bubbles”

This observations was then represented in a case structure which is made of historical stock projections represented by set of points, where each point was given with the time of measuring and the equivalent stock volume. With this, the processes was represented as curves.

Further review of literature was done on previous attempts to address the similar problem as well as the suitability of CBR and the IP approach.

This search informed the decision to carry out some preliminary experiments using standard datasets to ascertain the performance of selected classifiers on traditional dataset before applying on the time series dataset.

Series of Pattern Matching experiments were carried out using datasets from the stock markets, where the new bubble model created was used to identify new instances that fit into the model through the use of appropriate similarity metric. The results of these experiments marks the end of the forward problem, while the solution was used as a seed for the Inverse Problem phase.

IP implementation required taking the newly identified structure from the retrieved case , extract the asset characteristics around the time of the occurrence to identify any correlation between such characteristics and the forward problem model parameters in order to derive stochastic description of the factors that accompany the said bubbles.

1.6 Background Information on the Dataset

For the purpose of this research, three sets of datasets were used; Simulated data (From Drill Reports), published traditional data (Clinical Test) and stock report(Time Series).

1.6.1 Drill operation dataset

This dataset was simulated based on the literature obtained from drill operations, where some relevant features were selected to form the sample dataset. Descriptions of situations in these reports contain the occurring problems and their proposed solutions resulting in 278,040 observations. This dataset was then used in the experiments to determine the goodness of the created model and used some statistical methods to make estimate on the accuracy of the models that we created.

1.6.2 Respiratory disorder dataset

This is a labelled dataset from a clinical trial that was used in comparing two treatments for a respiratory disorder dataset. It was originally published in (Everitt et al., 2017), which involves eligible patients that were randomly assigned to active treatment or placebo, each at four monthly visits.The trial recruited 111 participants (54 in the active group, 57 in the placebo group). Their respiratory status was determined as being “poor” or “good”, which constitutes the categorical output. Covariates such as centre, sex and age were also recorded. The goal was to evaluate the effect of the treatment on the respiratory status based on the output.

1.6.3 Stock dataset

This dataset was drawn from recorded stock market repository; New York Stock Exchange (NYSE) obtained from Yahoo finance. For the purpose of this research, the monthly stock prices of sixteen companies were considered; TOTAL, APPLE, ASPEN, BA, CAT, CSCO,

CVX, DIS,FDX, FRD, GSK, IBM,INTC, JPM, NSRGY, TM. The period under consideration is from year 2000 to year 2018. The procedure is to collect the daily index historical data(to cover as many details as possible), and to pre-process them for outlier, missing value, and standardization of the data using Python. THE focus is on the Adjusted price rather than the Closing price. Although they both provide different information that can be used for analysis, the closing price is the raw price and only indicates end of sales price whereas the Adjusted price mirrors stock value after adjustments for any corporate actions like all applicable splits and dividend distributions, which better reflects the assets' perceived value by investors (Ganti, 2020).

The first two dataset were necessary as a guide to ascertain how well the model will perform on a classical (labelled) dataset before carrying out the unsupervised learning task on the real (stock market dataset) which is a time series data, known to be characterised with noise and high dimensionality.

The experiments were designed and implemented using the open source software (Van Rossum and Drake Jr, 1995),(McKinney, 2010),Buitinck et al. (2013a).

1.7 Research Contributions to Knowledge

The research at hand has presented several original contribution to both, the theoretical and practical knowledge about the underlying factors and vital indicators of financial bubbles.

1. Firstly, the hybrid methodologies in the ensemble IPCBR, including qualitative and quantitative data with varieties of literature study have widened the volume of publicly existing, scientific literature in the perspective of Artificial Intelligence and Financial Bubbles.

2. This thesis has established a footing framework in IPCBR for imminent research around trends and future developments in investment behaviour using Artificial Intelligence. Also, based on the derived results, a starting point for building a framework

with the capability to study and identify patterns in financial data for future economic developments is defined.

3. This research is a novel work, at the time of writing, little evidence is present in the literature to demonstrate the existence of any similar approach to the one proposed in this work to the best of the author's knowledge. One of such works related to this is by Jenny Freeman (2018) titled "Early warning on stock market bubbles via methods of optimization, clustering and inverse problems" which used the theory of optimization, of inverse problems and clustering methods to an early-warning signalling for financial bubbles. The uniqueness of this research is that the concept of bubbles is dealt with using the CBR methodology; a methodology that has proven to handle vague domain like the financial bubble by the use of past knowledge.

1.8 Delimitation and scope of the research

Given that the concept of financial bubbles is a broad topic and has the possibilities of being analysed in different ways, it is important to define the scope of the research to the audience in order to generate useful information and trends based on this. There is need to bear in mind that the trends and developments generated through the stock analysis only display assumptions that are not objective facts.

The main essence of this study is to shed some lights on the predictability of financial asset bubbles and further the understanding of indicators that may be used to forecast them. Sequel to this, emphasis is made on indicators that predict an asset bubble to occur. Bearing in mind that this investigation is made on the back of the critical importance of understanding financial asset bubbles and what causes them.

Even though there are evidence of varieties of bubbles in different countries and all over the world at different points in time, only the most relevant ones are chosen for discussion during this research in order to get useful insights on the topic.

1.9 Publications and research activities

This research has witnessed its publication where the general concepts, research idea and some results were presented in conferences and also published as stated:

1. Ekpenyong Francis, Stelios Kapetanakis, Miltos Petridis, (2017). An ensemble method: Using CBR and the Inverse Problem to Identify Deviations in Business Process. 22nd UK Symposium on Case-Based Reasoning – UKCBR 2018

2. Ekpenyong, F., Samakovitis, G., Kapetenakis, S., Petridis, M. An Ensemble Method: Case-Based Reasoning and the Inverse Problems in Investigating Financial Bubbles. IEEE International Congress on Cognitive Computing, San Diego, USA (06 2019) 153–168 - ICCC 2019 *(Best Student Paper Award)*

3. Ekpenyong, F., Samakovitis, G., Kapetanakis, S. and Petridis, M. (2020) "Towards the Ensemble: IPCBR framework in Investigating Financial Bubbles", European Journal of Electrical Engineering and Computer Science, 4(4). doi: 10.24018/ejece.2020.4.4.193.

4. Ekpenyong, F., Samakovitis, G., Kapetanakis, S. and Petridis, M. (2020) Case Retrieval with Clustering for a Case-based Reasoning and Inverse Problem Methodology: An Investigation of Financial Bubbles. The 2020 16th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2020) *(Accepted)*

Overview of the Thesis

The thesis is divided into five parts - **Part I**: preliminary background research and methodology development, **Part II**: Design and development of the approach, **Part III**: item Implementation and evaluation, **Part IV**: Contribution to knowledge and future research direction, and finally **Part V**: References and Appendices.

Part 1: Preliminary background research and methodology development

This comprises the first three chapters. The first chapter provides a general overview of the background information of the research, followed by the main research problem and subsidiary questions. The objectives of this study and the methodology of the research are also presented jointly with the key contribution of this work to knowledge.

Chapter two gives a summary of what has been done and what has not, in the relevant fields with respect to this study as it presents a broad review of literature in financial domain while focusing on bubbles in particular. The overall aim is to furnish the reader with a broad view of the research problem and useful background information.

This chapter expands on a structural geometric representation of the framework, which is proposed to act as a base description for our CBR training and subsequent implementation.

It also reviews some publications in Case-based Reasoning and the Inverse problems that is relevant to this research, with the objective of ensuring that this research is not repeating the work of others.

Part 2: Design and development of the approach

This section is made of chapters three and four. Chapter three describes the approach that is adopted to achieve the aim of this research, which includes the overall concepts of the research methodology, data processing, feature extraction, case representation and insight to case retrieval method. It also provides an articulation of the overall framework to be used in which the Inverse Problem formulation component is discussed.

A simple stochastic asset bubble model is then proposed in Section 4 with a brief review of the rational bubble model, which is the theoretical backbone of rational bubble tests.

This section presents the core concept in IPCBR, formulation, to describe, identify and ultimately predict abnormal fluctuations in stock markets, widely known as bubbles.

It also outlines the class of asset bubble problems to be addressed in this research, the proposed relevant features/qualities of CBR that make it suitable, as well as the overall IP formulation approach.

Chapter four takes the concept of time series further with emphasises on case retrieval

which is the most important aspect of the CBR engine. Highlighting the various measures popularly used in searching the case base to select existing cases sharing significant features with the new case. This is aimed at comparing existing similarity measures and with a view to provide general guidelines for choosing the best fit similarity in conformity with the sort of analysed data as well for the required types of vigour.

Part 3: Implementation and evaluation

The section comprises two chapters, chapter five which gives account of some selected interpolation methods mostly used in case retrieval which serves as a guide on the effective choice in building the proposed framework. Some popular interpolation methods considered suitable to both linear and non linear model values are considered. The reason for these selection is to ascertain how these interpolation methods perform on classical traditional dataset before applying it on Time series dataset that is characterized with high dimensionality and noise.

Chapter six addresses the forward problem formulation and the inverse adopted in this research through the sentiments analysis. Machine learning technique is used to analyse news sentiments and correlates it with the output resulting from the forward problem to identify causes of some fluctuations. It also presents the clustering method used in the CBR shell in handling case retrieval. The results of combination of several unsupervised learning algorithms in determining stock pattern.

Part 3: Summary and Conclusion

The work closes with a critical discussion on what has been done in this study and the major contributions this work has delivered, and a set of relevant concluding observations in chapter nine. It also addresses the limitations of this work and achievable future improvements which in turn leads to the discourse of prospective future research ideas.

2 Literature Review

2.1 Introduction

This section presents a broad review of literature in financial domain while focusing on bubbles in particular. Also, a review of some publications in Case-based Reasoning and the Inverse problems that is relevant to this research is made, with the objective of ensuring that this research is not repeating the work of others. The review aims to understand the various models and methodologies that is used in the course of this research. It is intended that this chapter will provide the reader with a broad view of the research problem and with useful background information.

However, the issues that are connected to each specific technical facet of this research will be elaborated in relevant chapters.

2.2 Background

While the term "Bubbles" are still contentious issues in contemporary economics considering various definitions from many perspectives, crashes and financial crises continue to occur regularly through history. These facts, combined with the difficulty of identifying a bubble and a general misunderstanding of bubbles, warrants further study into their workings and what implications for policy should be reached in dealing with bubbles (Brice, 2007). But the overall implication is the existence of sizeable and persistent deviation between the fundamental value of an asset and its market value.

Detecting a bubble in real time is quite challenging because attributes to the fundamental value is difficult to pinpoint. Researchers have tried to develop several methods to detect bubbles in the financial markets (Milunovich et al., 2019). The noteworthy bubble detection tests include statistical attributes of bubbles (Press and Profit, 2017; Katja Taipalus, 2012) and recently, Machine learning approaches (Duan and Stanley, 2010; Hu, 2019)

Although every bubble varies in their initiation and particular details, there is a trend in similarity and pattern in which informed assumption can be derived. Besides, it is possible to identify bubbles in advance because they often leave some certain traces. The importance of bubble around many historic periods of economic downturn and instability, coupled with the difficulty of identifying a bubble and general misunderstanding of bubbles, warrants further research study. Gurkaynak (2005) reports that despite recent advances, econometric detection of asset price bubbles cannot be achieved with a satisfactory degree of certainty.

2.3 Review on Bubbles

This section only provides account of some of the most famous bubbles, providing the little necessary background details into the concept of bubbles for the purpose of this research. For full documentation on this, the reader is referred to Reinhart and Rogoff (2014); Naqvi (2019) and Kindleberger et al. (2005).

Variety of asset price bubbles have been specified in the academic literature, rooted specifically in its conceptual evolution and development. Based on the assumption that bubbles can occur from time to time in asset markets, and yet the human agents fail to comprehend. Their elusive nature make bubbles different in their initiation and specifics, but the recurring existence of similarities in patterns can provide useful insights and general assumptions can be made from them.

This section will take a brief survey of some famous historic bubbles in a bid to relate how they are intimately linked to the conceptual bubble definition that is presented in this work. This review is geared to assist in recognizing the leading indicators of an economy

prone to speculative behaviour which forms the basis of this study. However, the reader is referred to Garber (2000) and Cooper (2008) for more detailed documentation on this.

2.3.1 Tulipmania

According to literature, the Netherlands witnessed the first legitimately recognized creation, dissemination, and resultant crash termed bubble in history; Dutch Tulip Mania, started 1634 in Holland and burst in 1637 (Garber, 2018; Zeitschrift et al., 2019). As Dutch traders sold tulip (a flower native to east Africa) futures, prices quickly deviated from the core value of the actual asset and rose increasingly as traders assumed that wealthy foreign individuals would ever purchase bulbs of the fresh brilliant varieties, irrespective of the price. At the outbreak of virus that attacked the flowers, causing modifications in their appearance, the already scarce tulip became more hunted, keeping its price on continuous rise and eventually reaching a peak. Merchants later on realised how tough it would be for one to meet with the financial capacity to pay such huge prices for tulip bulbs, triggering the price to swiftly decline to its pre-bubble state. Consequently, On February 5, 1637, the flower market in The Netherlands came crashing down, while many had already purchased expensive tulip bulb futures (Chang et al., 2016a), pondering that they could resell the created bulbs three to seven years later, and the quick fortunes made by so many Dutch citizens were permanently lost. Luckily, the Dutch stock market did not involve itself with tulips, consequently, the tulip crash had not much impact on the Dutch economy(Chang et al., 2016b).

2.3.2 The South sea Bubble

The South Sea Company's history as recorded in Cooper (2008); Caroline Thomas (2003) is renowned for its bubble, a spectacular and almost unbelievable financial scheme that crashed nearly three hundred years ago, throwing anguish surfs in the eighteenth-century Britain. The Southsea history connects to the bubble case because it signifies seriousness of

investors actions. According to the literature, the company was formed in 1711 by Robert Harley, who needed partners to carry through peace negotiations and initially took over the national debt raised by the United Kingdom in War of Spanish Succession (1701-1714) a reward for a promise of a dominance of all trades to the Spanish colonies in South America (Dale et al., 2005; Oth, 2003). The control was centered on the hope of obtaining massive trading privileges from Spain in the peace pact. The firm had no assets but the claims of success of trading to South America was only a fictitious propaganda by the executives of the company. This incredible rise of rumor sky rocketed the price of the company's stock from £128 to £175. The result of which made the company bullish while the price continued on the increase leading to the pinnacle of £1050. When regular trading resumed, prices began to deteriorate, although the subscription was still strongly oversubscribed. Later on, prices fell quickly. In a short time, they almost declined to their starting level.

2.3.3 Dot-Com Bubbles

The dot-com bubble which began in 1995 (Chang et al., 2016a) with the transition of the traditional business model due to the introduction of the World Wide Web created an excited attitude towards business. As a result, hastening many aspirations for the affluent future of online commerce, flourishing quickly with new internet based companies evolving on daily bases. Stocks were trading at record multiples of earnings (Morris and Alam, 2012), many investors ignored the fundamental rules of investing in the stock market but rather, chase speculative behaviour of investors, resulting in internet related stocks and assets exponential acceleration from 775.20 to 2,505.89. These happened between January 1995 to January 1999, more than doubled from this point to its peak of 5,048.62 in March 2000 when the bubble burst (Leone and de Medeiros, 2015). With immense deficits, technology companies began to deteriorate and collapsed. Investors limited their portfolio exposure to the industry and sold most of their shares.

2.3.4 Housing Bubble

Between 2000 and 2012, America experienced a great housing convulsion that had all the classic features associated with real estate bubbles which grew up alongside the stock bubble in the mid-90s (Baker, Baker). Housing prices rose dramatically and then fell, leaving average real housing prices in 2012 no higher than they were in 2000 (Limjaroenrat, 2017). Since housing wealth is far more evenly distributed than stock wealth, the bursting of the housing bubble had global consequences and feedback causing financial crashes around the world than the collapse of the stock bubble.

These few classic cases present bubble regimes that resulted to similar patterns which were motivated by the nature of man and markets. ”History”, they say ”keeps repeating itself”. No matter how many bubbles happened, they will continue to happen again and again.

2.4 Financial Markets/Stock exchanges

A stock is an ownership share in a business/corporation, whereas, a financial market refers broadly to any marketplace where the trading of security occurs, including the stock market, bond market, forex market among others (Sikarwar and Appalaraju, 2018). The New York Stock Exchange (NYSE), Oslo Stock Exchange, and NASDAQ are a few typical examples of a physical financial market that are more specialized in certain types of corporations and industries.

A Stock Exchange is a stock market where brokers and traders buy, sell or exchange publicly listed financial instruments. Commonly, stock exchanges provide a way for brokers and traders to exchange financial instruments, Olden (2016). Traditionally stock exchanges were physical places, often referred to as the floor, where stock-brokers and -traders exchanged stocks for other stocks or money.

Presently, some smaller stock exchanges still have a floor where stocks can be traded, most of the stock trades take place through electronic communication, which enables the

trading of certain stocks and other financial instruments through a near instant electronic trading system.

Majority of the notorious stock exchanges use a continuous auction principle which includes an instant execution of stock orders as they are received by the market. By operating with the continuous principle and rapid electronic orders, modern day stock exchanges are driven solely by supply and demand.

2.5 Stochastic Processes in Financial Markets

Financial markets exhibit several properties that characterise complex systems (Gontis et al., 2016). A universally accepted assumption in financial theory is that these time series are erratic. This constitutes the hub of the description of price dynamics as stochastic processes. The concept of randomness is central to finance formalised by the mathematics of stochastic processes (a collection of random variables, representing the evolution of random values over time). A more formal treatment is given by Hidalgo (2011).

This property brings us to an interesting problem in finding a way of representing the underlying variation of the prices on the time scale of the regular measurement. Sequel to this, the econometric formula of Returns, Log Returns and Compound returns (Dunbar and Hall, 2016) is visited.

2.5.1 Proportional Returns

Statistically, it is preferable not to work directly with the raw price data but rather to convert them into series of returns as this makes them unit free. Proportional return otherwise referred simply to as return R is the profit on a particular investment (Miskolczi, 2017). It includes any change in the asset value, interest, commission or dividends and all other cashflows which investors receive or pay due to the investment. It defines a natural variation at

time t and is represented by

$$R_t = (s_t - s_{t-1})/s_{t-1} \tag{2.1}$$

The use of return metrics encourages consistent comparison among two or more securities even though their price sequences may differ by orders of magnitude.

2.5.2 Log-return

Log return has interesting property that they can be interpreted as continuously compounded return which is time additive and hence makes way for easy comparisons with other related financial time series. This is represented by

$$r_t = log(s_t/s_{t-1})$$

where s_t is a sequence of prices on a stock or a portfolio of stocks, measured at regular intervals. Say day-to-day, although there were many attempts to define stock pricing through a kind of a stochastic process, i.e. Brownian motion or geometric Brownian motion, for simplicity.

2.6 Bubble: Representation

Even though there are many opinions about bubbles in various literature, one thing is obvious, none of the authors seems to disagree about the theoretical determination of the fundamental asset price. An asset price bubble according to Nedelcu (2014) is defined as the difference between two components: the observed market price of a given financial asset, which represents the amount that the marginal buyer is willing to pay, and the asset's intrinsic or fundamental value, which is defined as the expected sum of future discounted dividends. In trying to give meaning to what a bubble is, this section defines what a fundamental value of an asset is. The representation is adopted from the concept given by Barlevy (2007) which starts with a case of an asset that yields a known and fixed stream

of dividends. In which case, d_t denote the dividend income paid out by the asset at date t, where t runs from 0 to infinity, and q_t denotes the current price of a bond that pays one dollar at date t. It stipulates the value any trader ties to the dividend stream from this asset, and is given by

$$F_t = \sum_{i=0}^{\infty} q_t d_t \tag{2.2}$$

here, F denotes the fundamental value of an asset. An asset bubble is therefore an asset whose price p is not equal to its fundamental value, meaning $P \neq F$. a bubble case would assume the asset price to sell above its fundamental value, in which case $P > F$. Also, when considered where dividends are uncertain, given a state of the world represented as states in a set Φ which denotes a set of all possible outcomes at a date t, given that $\tau_t \in \Phi$ refers to a particular state of the world at date t which all agents hope will occur with a probability say $Prob(\tau_t)$ which determines the value of a dividend at date t given by $d_t = d(\tau_t)$, assuming $q_t = q(\tau_t)$ denote the value agents assign to amount they receive at date t in a particular state τ_t, relating this with equation above, the fundamental values agents assign to the asset in this case is expectation that

$$F_t = [\sum_{i=0}^{\infty} q(\tau_t) d(\tau_t)] = [\sum_{i=0}^{\infty} \sum_{\tau_t \in \Phi} q(\tau_t) d(\tau_t)] \tag{2.3}$$

In this case, an asset would be considered a bubble if its $P > F$ as defined in the equation.

The equation above could be related to a descriptive bubble case adopted from the work of Asako et al. (2017), which represents a growing asset prize with respect to time t. shown in figure 2.1. Time here is considered continuous and infinite with periods $t \subset R$

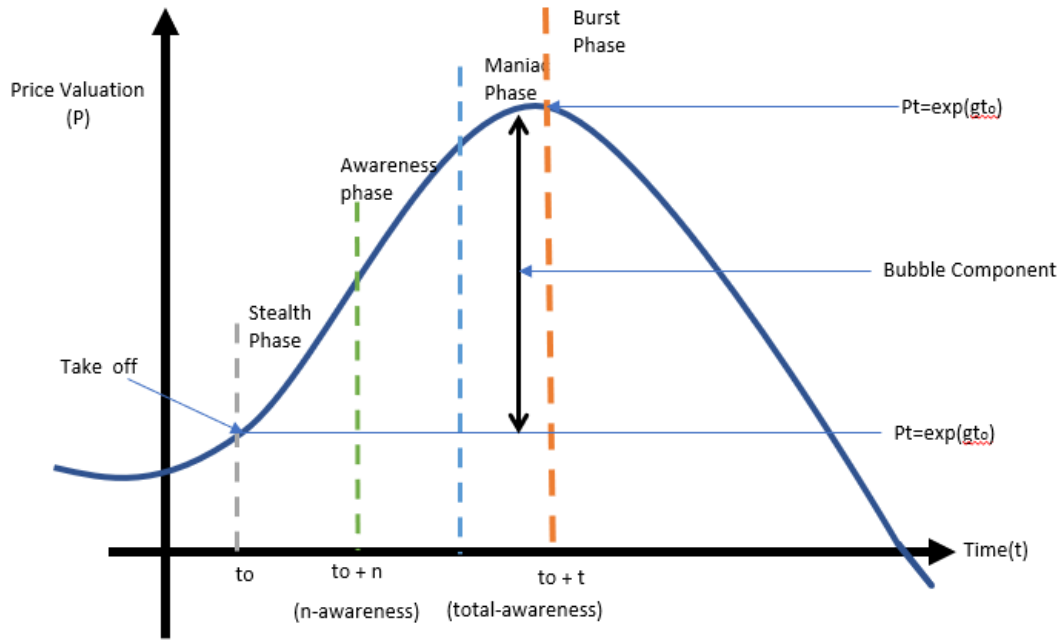

Figure 2.1: Bubble representation

Figure 2.1 shows lifespan of a bubble based on (Abreu, 2003). An initial constant steady growth in asset price based on the fundamental value at some random time t, At a point say t_o, (take-off point) the price driven by a bubble grows in time value with expectation $gt(t_o)$, From t_o the asset price pt grows exponentially at $g > 0$, denoting evolving price with a growing expectation given by $p_t = exp(gt)$. Hence the bubble component is denoted by $exp(gt) - exp(go)$ where $t > to$. The assumption is that the starting point of a bubble is t_o is discrete as $t_o = 0, \delta, 2\delta, 3\delta, ...$, where $\delta > 0$ and also obeys the geometric distribution with probability function given by $\psi(t_o) = (exp(\beta) - 1)exp(-\beta t_o)$ where $\beta > 0$. In this situation, the type of investors holding this asset are considered risk neutral investors that have a discount rate of zero, whereby, as long as they hold the assets, they have two choices; either to sell or retain the assets. But when $\alpha \in (0, 1)$ of investors sell their assets, the bubble bursts and the asset price drop to the true value. If fewer than α of the investors sell their assets at say time ρ after t_o, the bubble bursts automatically at $t_o + \rho$ but if the decision is

for the assets to be sold at t just before the bubble bursts, the investor receives the price in the selling period otherwise what he gets is the true value $exp(gt_o)$ below the price at $t > t_o$.

2.7 Characteristics of bubbles

During bubble, the market changes, exhibiting structural transformation to attain a completely new regime, which is most likely driven by sentiment, and no longer reflects any real underlying value (Sornette and Cauwels, 2014). This describes a situation in which the price of an asset has increased significantly in such a short period of time suggesting that the price is susceptible to an equally sudden collapse, with the potential to be hugely damaging to the financial system causing so many disturbing financial and economic concerns. Irrational exuberance (Shiller, 2003; Erber, 2010) is a term that has often been used to characterize the continued upward bidding of prices that goes significantly above what would rationally be explained as a fair price. What is rational is addressed differently by various economists, the acceptable view is defined by market fundamentals. Namely, the sum of the sequence of discounted dividends (Brice, 2007). Given by:

$$F_t = \sum_{i=0}^{\infty} q_{t+i} d_{t+i} \tag{2.4}$$

where q_{t+i} is the discounted factor $\frac{1}{(1+r)^i}$ and d_{t+i} is the dividend. The Rational Choice Theory (Brabham, 2019; Lovett, 2006), dictates that investors do not deviate with great margin from this price. But it is observed that in times of bubble, investors are willing to price far more than its fundamental value. By introducing a bubble term $b_t > 0$ which represents arbitrary price rise that increases at rate r so that $b_{t+1} = (1+r)_{b_t}$, the condition affects the price as represented

$$P_{t+n} = \sum_{i=n}^{\infty} q_{t+i} d_{t+i} + b_t = P^*_{t+n} + b_{t+n} \tag{2.5}$$

In this case P^* is the price obtained from the previous equation which could be considered as the fundamental fair price detected by the future dividend flow and b the bubble term or

overvalued price on the asset. Therefore, the obtained price P will diverge at a geometrical rate from the fundamental price P^* with time as the bubble term grows. What causes this jerk in price could be attributed to several factors; Psychological contagion that spreads from one person to another on observance of price increase accompanied by amplifying stories that may justify the increase attracting a larger investors otherwise known as the feedback loop (Brice, 2007) which fuels a spiralling growth away from equilibrium. This stage is otherwise referred to as the Displacement as explained by economist Hyman P. Minsky (Kindleberger et al., 2005; Friedman, 2012) which occurs when investors start to notice a new paradigm, like a new product or technology, or historically low interest rates or any juicy investments. This leads to speedy rise in price at first, then get momentum as more investors enter the market resulting in a boom. There is an overall sense of failing to jump in, causing even more people to start buying assets.

When sentiment changes and demand for the asset can no longer sustain the high price, investors sell the asset giving birth to "Crisis of Confidence". Now the feedback loops already mentioned operate in the opposite direction. Often, this is followed in short order by a market collapse resulting from the synchronisation of sell orders. As the price falls, investor sell more of their stock which cause further drop in price triggering even more sales. The bursting of the bubble occurs much more rapidly than the growth as investors' confidence is shattered. The crash occurs because the market has entered an unstable phase after a long maturation process associated with the inflation of the bubble. This mechanism is somewhat vague and much controversies then arose about the cause of the crash. In fact, the cause is quite obvious and is found in the preceding years of exuberant bubble dynamics that made the whole construct fragile

Bubble detection is regarded as a challenging task. There have been many studies using econometric approach, (Al-Anaswah and Wilfling, 2011; Yiu et al., 2013) and some recent studies have presented results using data mining techniques such as rule induction (Ince, 2014; Price et al., 2014), neural network (Magoulas and Vrahatis, 2006; Zhou et al., 2018), and combination of classifiers (Ou and Wang, 2009). CBR technique is one of the popular

methodologies in knowledge-based systems which solves a new problem by reusing specific knowledge from past experience. Its strengths have made researchers apply CBR to many areas with recorded successes, hence the reason to apply the methodology to unravel the causes of bubbles.

2.8 Case-based reasoning (CBR)

Case-based reasoning, a paradigm for combining problem solving and learning, has attracted the attention of researchers in recent times due to its simplicity and flexibility. This methodology has been applied in various application domain and it offers advantages over other AI based techniques in fields where experiential knowledge is readily available; such examples are available in areas of engineering (Romli et al., 2015; Shokouhi et al., 2014), planning and production (Marcela and Velandia, 2006; Lei et al., 2001; Seo et al., 2007), medicine (Bichindaritz, Bichindaritz; Merelli and Luck, 2004), commerce (Kaur et al., 2015; Kapetanakis, Samakovitis, and Gunasekera, Kapetanakis et al.) and others. Its operations is based on the concept proposed by Aamodt and Plaza (1994). The CBR cycle is illustrated in figure 2.2, which comprises of the four Rs (Retrieve, Reuse, Revise, Retain) as explained below.

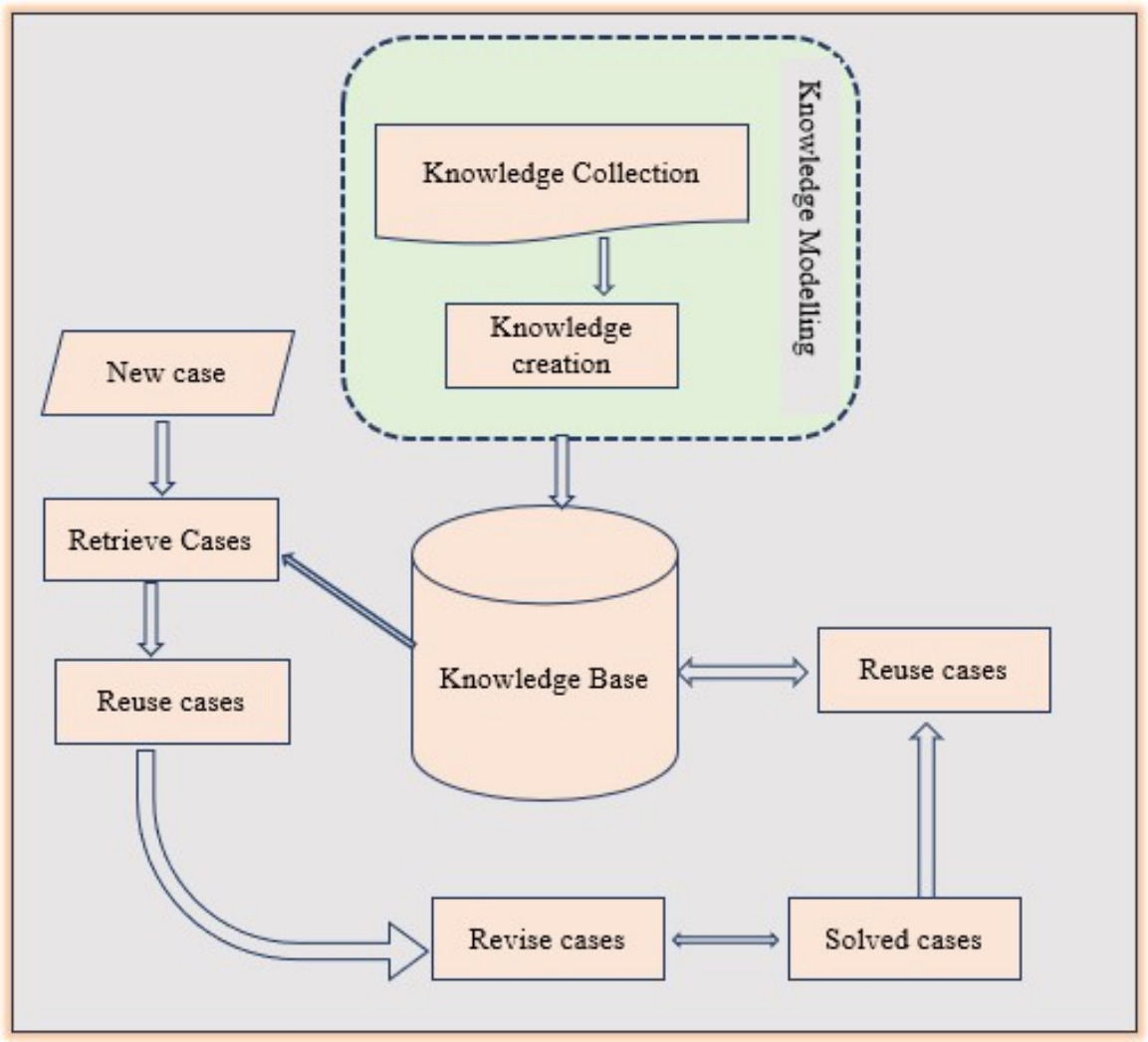

Figure 2.2: Case based Reasoning Cycle modified from Aamodt and Plaza (1994)

Retrieve:

In terms of case retrieval, the most corresponding case (or k most similar cases) to the problem case, is retrieved from the repository. A suitable similarity measure is selected, depending on the application domain and features used to define cases. This is then used to compute the relationship between a new case and previous cases stored in the case base.

Reuse:

In the reuse phase, otherwise referred to as case adaptation, modifications of the retrieved case is made to provide a better-adapted solution to the problem. This process usually maps the solution from previously stored cases to the target problem, which then becomes the immediate solution if it can be fittingly applied to the target case, and could as well be used for the latter solution.

Revise:

This phase maps previous solutions to the target situation in such a way that new solution can be applied to it. In this phase, the previous cases need to be further scrutinized to ascertain whether they are appropriate for the current situation and whether they can solve the target problem. The solution is tested for success, and repaired if failed. If revision is not possible, the system fails in finding a suitable solution to the target.

Retain:

In this phase, the knowledge acquired from this problem is incorporated in the system by modifying knowledge containers. This means that the obtained solution is evaluated, and decisions are made as to whether to retain the new solved case or not. This process involves dynamic procedures of adding and removing cases seeking to enhance the efficiency of the CBR model. Useful knowledge is retained for future reuse, and the case base is updated by a new learned case, or by modification of some existing cases.

2.8.1 Case Representation

Appropriate case representation has become a fundamental factor to CBR system because the methodology is heavily dependent on the structure and contents of its collection of cases. Case representation in CBR utilizes accustomed knowledge representation formalisms from AI to represent the experience embodied in the cases for reasoning purposes,

the capability of representing many different types of knowledge and store it in many different representational formats.

Case representation has posed some fundamental issues in Case-based Reasoning methodology due to existence of some unaccustomed features that do not exist or align with the processing of the traditional "attribute-value" data representation. A second issue arises from the difficulties with direct manipulation of continuous, high dimensional data in time series domain (Marketos et al., 1999).

From the available literature, cases are usually represented as two unstructured sets of attribute value pairs, that is the problem and solution features. This is composed of p attributes where each k attribute corresponds to either problem or solution attribute; N being total number of cases in case base.

In general, a case can be described by a set of attributes that corresponds to business process formulation specifications (e.g. materials, ratios and processing methods) and its associated final properties. Therefore, for the implementation of CBR system, the formulation parameters are coded into cases. Data regarding both the dimensions formulations to perform the operation process, and testing data showing the performance of those of processes are used to represent the case solution and the case problem respectively. Case problem's attributes consist of several physical properties measured for the operation, and the case solution's attributes correspond to the business formulation variables.

These cases have an attribute value representation in table 2.1 :

This becomes the preferred format because formulations are normally presented as a list of ingredients and processing conditions with their corresponding values. This type of representation is also used by other researchers (Segura et al., 2007), in formulation of cases.

Considering, for example a specific case i, the feature case vector c could be formally represented as:

$$c^{(i)} = \{c_1^i, c_2^i, ..., c_k^i\} \tag{2.6}$$

(Cases)	vector1	vector2	...	vector k	vector j
1	1,	2,	...	k	$j^{(th)}$ attribute
2	c_1^i	c_2^i	...	c_k^i	c_j^i
.	.	.	...	.	.
.	.	.	...	.	.
i	c_i^i	c_i^i	...	c_k^i	c_j^i
N	c_1^N	c_2^N	...	c_k^N	c_j^N

Table 2.1: Feature vector representation of cases

the sets of which constitutes the case base (CB) denoted by:

$$CB = \{c^{(i)}, ..., c^{(N)}\} \tag{2.7}$$

However, the decision of what to represent can be one of the difficult decisions to make especially when considering the nature of time series. No one representation of time series is superior for all tasks. There exists some level of obscurity on how to choose the best representation for the chosen problem.

In this research, the problem is addressed by forming a library pattern of observations where every group is considered as a case category. The adopted approach followed the concept proposed in Pecar (2002), where the entire Time series is split into smaller sequences of patterns, by decomposing the series into a sequence of rolling observation patterns or rolling windows in which case, every observation in the pattern constitutes the case which may attain a predefined upward, steady or declining patterns as shown in figure 2.3

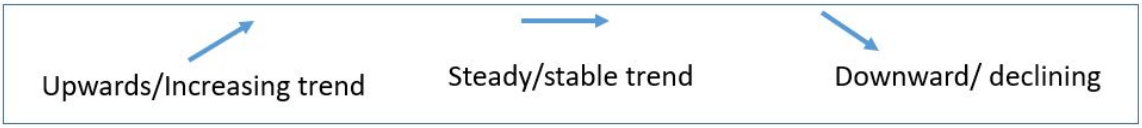

Figure 2.3: Sample case patterns

This also infers that an interval comprising a series of the three observation patterns

can be easily recognized as constituting a case. Further analysis and matching of all the similar cases using appropriately selected algorithm makes it possible to discover a specific relation to the pattern.

2.8.2 Case Similarity

Similarity assesses the right cases to be retrieved. In most of the CBR system, it considers the proper features and usually their importance for comparison between cases. Some researchers judge the similarity by taking into consideration the adaptation that needs to be carried out with respect to the new case. Detailed literature on similarity is presented in section 4

2.8.3 Adaptation

Adaptation is at the heart of the CBR process and plays a central role (Fuchs et al., 2000; H. and Elmogy, 2015; Qi et al., 2012). It is seen as one of the most challenging tasks (especially in a complex problem domain) in CBR (Liao et al., 2012). Adaptation relies on both the retrieval of proper cases that need less adaptations and the utilisation of appropriate domain knowledge. Traditional methods include substitution method that replaces some part of the retrieved case, and transformation method that transfer some part of the retrieved case to fit the constraint to a new situation (Craw et al., 2006). In most cases, adaptation is performed manually, although recent research employing heuristic methods have provided many promising prospects.

2.9 Essence of Case-based Reasoning(CBR)

In general, people use case-based or analogical reasoning in variety of situations, notably, making business decisions. Experts develop new design formulations by recapturing various kind of established or known knowledge, be it formal (through mathematical models) or informal (rules of thumb), and focus their search for solution through heuristic reasoning.

In other words, they tackle new problems based on their previous experience and perception when faced with problems that cannot be formalized. They are seen to recognize problems in isolation and have difficulties in visualizing common combinations of problems in most cases. All these however requires enormous training times. However, it is near difficult for the experts to be trained to recognize every combination of problems. Although they can apply various kinds of knowledge that they develop through experience, they lack the comprehensive theoretical knowledge base of financial bubbles. Also, various results from financial bubbles predictions have left investors with more questions than answers based on the assumptions of its "a random walk" theory (Degutis and Novickytė, 2014; Roofe, 2005).

The quest to proffer lasting solutions to these issues faced by experts informs the decision to apply CBR to financial bubble investigation. The CBR methodology is capable of solving problems from open (vast base of knowledge), dynamic (constant changes in knowledge base) and weak (high level of uncertainty) domains as the case with the financial bubble, that is characterised with fuzziness and lots of complexities.

Some other key reasons why CBR is considered more appropriate for this research is domiciled in the unique nature of the methodology and its reasoning powers in the following:

1. Domains with little or no body of knowledge, for instance, in financial market and bubble: A CBR system can build its knowledge incrementally as the cases are added and hence increase its efficiency over time.

2. Reasoning with imprecise data: Although various econometric techniques try to make predictions, the use of imprecise data leaves the solutions difficult to explain. Whereas, machine learning algorithms and tools can support CBR tasks to cope with uncertain and imprecise data.

3. Providing a means for explanation: Based on the similarity measure in use, CBR makes it easier to explain how a solution was arrived at and the reasoning involved

in such adaptation.

Marling et al. (2002) in their report however emphasized strong innovations for integrating CBR with other reasoning modalities and computing techniques to more accurately model the knowledge available in a problem domain, and compensate for the shortcomings of one approach by capitalizing on the strength of another.

2.10 Inverse Problems

The term "inverse problem" which appeared in the 1960s, has witnessed a great drift from its initial purpose. It was initially designed to designate the determination, through input/output or cause-effect in experiments of unknowns in the physics equations. It is now used as a contemporary "inverse problem" that designates the best possible reconstruction of missing information to estimate either the identification of sources, the causes, or the value of undetermined parameters (Argoul, 2012).

The definition of an inverse problem starts with a mapping between objects of interest, which is termed parameters, and acquired information about these objects, which is referred to as data or measurements (Bal, 2012). Stochastic driven data, probably imprecise prior information on model parameters, and a physical theory relating the model parameters to the observations are the rudiments of any inverse problem. Inverse Problems have registered numerous successes in science and engineering fields since many important real-world problems can be solved through its application (Sun et al., 2020; Sever, 2015; Iulian, 2013). The inverse problem provides a platform where related problems from different disciplines can be studied under a common approach with comparable results (Gomez-ramirez, 2003)

2.10.1 Inverse Formulation

The use of inverse analysis techniques (also called model inversion) represents a new research paradigm that arise in many fields, where one tries to find a model that typically

approximates observational data. Inverse problems are often formulated by assuming that the underlying phenomenon is a dynamic system characterized by mathematical equations, although no such assumption is always essential (Sever, 2015). Often, any inverse theory requirement is to relate a physical parameter *"x"* that describes a model to acquire observations making up some set of data *"y"*. Assuming there is a clear picture of the underlying concept of the model, then an operator can be assigned a relation or mapping x to y through the equation $F(x) = y$; formulated in each vector space setting. Figure 2.4 illustrates the dependency between x and y (input and output parameters) which can therefore be represented mathematically as follows (Iulian, 2013):

$$y_k, \ldots . y_m = f_k(x_1 \ldots x_n) \tag{2.8}$$

Figure 2.4: Dependency between input and output parameters

Where $i = 1$ to n and n is the number of input parameters and y_k is the output parameters; k ranges from 1 to m. The problem of estimating x (inputs) from a measurement of y (observation) is a model of an inverse problem expressed in the same vector space setting. As such, the idea of estimating x from a measurement of y is a prototype of an inverse problem. Given by:

$$x^{-1} = f(y_1, \ldots . y_m) \tag{2.9}$$

If the operator f is linear, the inverse problem is termed to be linear and the direct inverse is easy to find; otherwise it is a non-linear inverse problem, termed ill-posed problem which becomes quite difficult to determine the inverse.

2.10.2 The Forward and Inverse Problems

Considering the equation $y = f(x)$, the traditional forward function "f" (also called forward model) typically predicts a set of data that one is interested in. That is, f conducts simulations by entering different values for x and then examining the values of y that are generated as output. Diagrammatically, this could be visualised as in figure 2.5. It follows therefore that in a real situation, there may be many inputs (x) and outputs (y), and the nature of the function (f) may be complex (Tarantola, 1987).

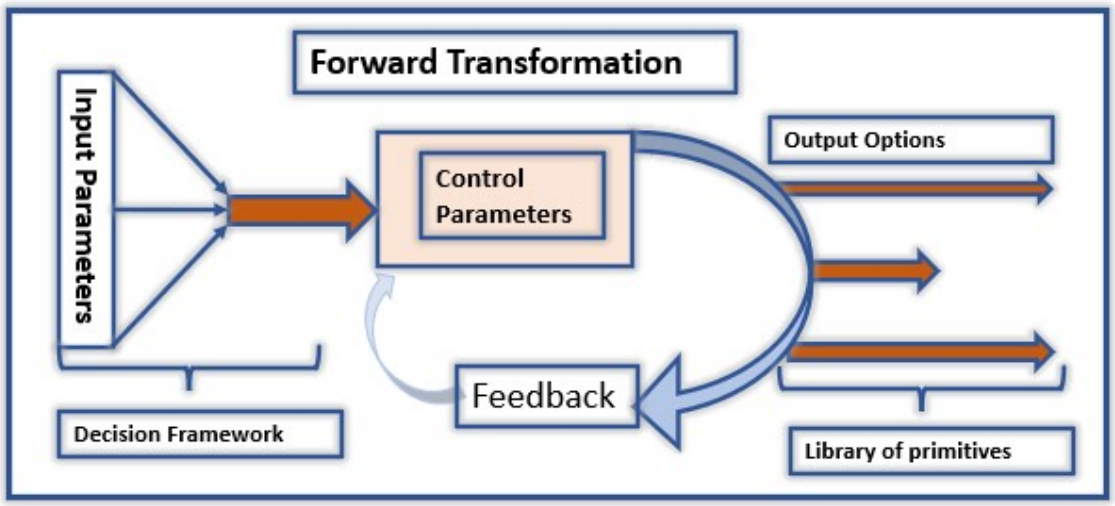

Figure 2.5: Forward Transformation

On the other hand, the inverse activity of the function f performs optimizations by placing a target value for y, then deciding or estimating the values of x that result in the target value for y, as illustrated in figure 2.6. In a situation where the inputs (x) and outputs (y) are of high volumes, this can increase the computation time of the model resulting in complex search procedures and time-consuming simulations, and in the worst case, may become computationally impractical (Tarantola, 1987).

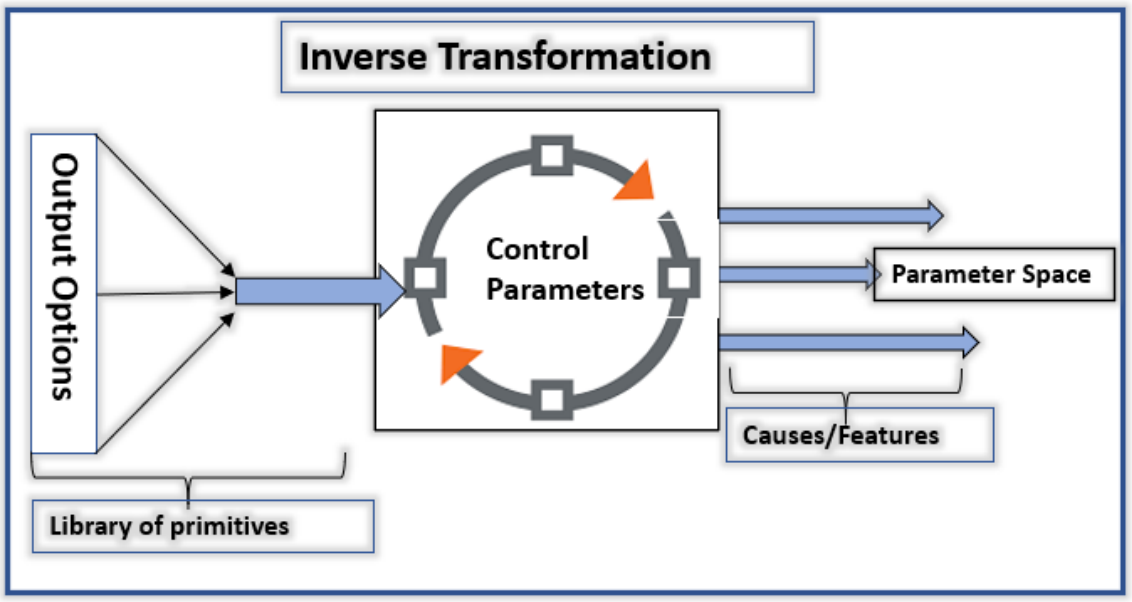

Figure 2.6: Inverse Transformation

Revealing cause from the knowledge of their effects (Argoul, 2012) is central to the idea of solving inverse problems. This is a crucial ingredient for the identifications of practical applications in different critical areas such as mine detection, medical imaging, remote sensing, non-destructive testing, and geophysical explorations. An "Inverse Reasoning" approach that goes beyond optimisation is adopted in this case, by taking into account the subjective knowledge (achieved through Sentiment Analysis) of the concept of asset bubble predictability.

The wholistic methodology begins with the forward modelling, which takes a cause C into an effect E, thus allowing us to make predictions on the value of observables, while the inverse problem uses those predictions to derive the values of the parameters that distinguish the system.

2.11 Methods of solving the Inverse Problems

Inversion methods theory involves a combination of mathematical techniques that operate on a reduced amount of data in a problem, with a primary objective of obtaining useful information relevant to the real physical system in question. There are very many solutions inverse problems in existence. Argoul (2012) classified various approaches to solving inverse problems into three main categories:

1. Regularization of Ill-Posed Problems: The solution obtained by regularized inversion will depend upon the data in a continuous manner, and will approach the exact solution (supposing it exist). The regularization of an inverse problem consists in rewriting the problem (Argoul, 2012).

2. Stochastic or Bayesian Inversion: In order to represent every uncertainty, all of the variables are considered to be random (Ulrych et al., 2001). Since for an under determined problem there are often several possible solutions, it is necessary to specify the confidence level that one can give to each solution.

3. Functional analysis: This is a decision-making approach in which a problem is broken down into its component functions. This approach has a strong impact on the development of inverse problems (Arridge et al., 2019) in a situation where the problem is ill posed. The problem can be modified into a well-posed problem by playing with the choice of the spaces that describe the variables, and the choice of their topology that allows the determination of the standard deviation or error. The choices of which are determined principally by physical rather than mathematical considerations.

Other approaches of solving the inverse problem are cited in Yao and Eddy (2014), a statistical approach, Domain decomposition (Jeewana, 2000), Neural Network (Elshafiey, 1991), and solving Inverse Problems with Piecewise Linear Estimators(Yu et al., 2012), Numerical approach (Al-Jamal, 2012). In all these, there is no generally acceptable approach or method that can solve all inverse problems. As such, what is being presented

in the subsequent section is a representation of some general methods to the solution of inverse problem.

2.11.1 Least-squares Method

This is perhaps the most widely used technique in geophysical data analysis (Estimation, 2008). It produces the estimated parameters with the highest probability (maximum likelihood) of being correct if several critical assumptions are warranted. When applying the Least-square Method to solve an inverse problem, the parameters not accounted for in the model may make the problem impossible to solve exactly.

For the least-squares methods, the sum of squares of the errors between the data recorded and the data that the model should have produced is taken as the measure of closeness. The basic problem is to find the best fit straight line $y = ax + b$ given that for $n\varepsilon\{1,2,...N\}$, the pairs (x_n, y_n) are observed, Miller (1992). The least square method can be used to determine the input parameters of a given inverse problem simultaneously.

2.11.2 Trial and Error Method

Trial and error is a basic method of learning that essentially all organisms use to learn new behaviours. This problem solving method involves using multiple attempts to reach a solution. Trial and error is trying a method, adjusting the input parameters, observing if it works, and if it does not work. If it does not work, a new method is then tried. This process is repeated until success or a solution is reached. The magnitude of adjustment on the input parameters may depend on intuitive judgement. He et al. (2010) applied this method to solve a typical ill-posed Bioluminescence tomography problem, which involves reconstructing the bioluminescent source distribution inside a biological tissue from the optical signals detected on the body surface.

2.11.3 Heuristics Approach

The notion of heuristics has played a crucial role in the A1 research. It is a science of problem-solving behaviour that focuses on plausible, provisional, useful, but fallible, mental operations for discovering solutions (Romanycia and Pelletier, 1985).

Heuristics is usefully applied in solving Inverse Problems. It is a problem-solving approach that is applied in optimization problems without exploring the whole space of solutions (Zhu et al., 2015). By so doing, the search space for finding the solution of the problem is reduced (Saha, 2002). It helps to discover the best and most practical ways to solve problems based on a simple principle of deciding which among the alternative course of action promises to be the most effective way to achieve some goals. The process makes it a mere decision strategies that sometimes overlook or discard some part of the available information, basing decisions on only a few relevant predictors.

2.11.4 Artificial Neural network (ANN)

Recent discoveries have witnessed novel algorithms using deep learning and neural networks for inverse problems based on the framework of the Artificial Neural Network (ANN). The ANN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. ANN topologies are composed of a large number of simple processing elements (neurons), in layers with highly modifiable weighted interconnections.

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. An ANN can be configured for a specific application, such as pattern recognition or data classification, through a learning process. Other advantages of ANN include:

1. Adaptive learning : An ability to learn how to do tasks based on the data given for training or initial experience. Magoulas and Vrahatis (2006); Moreira and Fiesler

(1995); Chaplot et al. (2016) used Artificial Neural Network to construct Adaptive Learner Model, which automatically models relationship between different concepts in the curriculum.

2. Self-Organisation: An ANN can create its own organisation or representation of the information it receives during learning time (Zgurovsky and Zaychenko, 2017; Schetinin, 1998) applied self-organizing the multi-layered neural networks in medical diagnostics.

3. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices being designed and manufactured to take advantage of this capability .

4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.(Kulakov et al., 2015)

Neural Network has proven to be an efficient tool for solving various inverse problems. Adler and Öktem (2017); Krasnopolsky and Schiller (2003) created a forward and inverse problems and applied the technique in Remote Sensing, Jin et al. (2017) used direct inversion followed by a Convolutional Neural Network to solve normal-convolutional inverse problems.

2.11.5 The Inverse Solution through Sentiment Analysis

Over the years, investors and scholars have devoted so much of time and resources into developing and testing models of stock price behaviour, which is exceptionally challenging because of the highly complex and dynamic nature of the markets. The efforts were greatly centred on the use of fundamental analysis and Technical Analysis based on unstructured data and structured data respectively. A third approach, the Sentiment Analysis (Kaushik et al., 2015) has also gained popularity in stock market analysis. The inverse approach used

in this research is based on the later; an approach which is used to extract some meaningful information from specific data.

This approach is of significance because there are specific patterns in stock behaviours which the Efficient Market Hypothesis (EMH) fails to explain (Degutis and Novickytė, 2014). Although this hypothesis proves that financial market movements depend on factors such as news, current events and product releases among others (Picasso et al., 2019). All these components have shown to have significant impact on the overall stock value, while it retained that the driving force behind price changes is the arrival of new information (Roofe, 2005). Available literature on bubbles and crashes also emphasise that the news media has an important relationship with investor beliefs, hence the need to incorporate qualitative element from news related to the stocks.

There are several research publications that report on applying Sentiments Analysis to predict stock movements like (Nisar and Yeung, 2018; Goel and Mittal, 2012; Jiawei and Murata, 2019; Bharathi and Geetha, 2017), predicting Bitcoin price fluctuation, Cryptocurrency Price Prediction (Abraham et al., 2018), Stock Price Returns (Ranco et al., 2015).

This inverse solution is set to explore and identify patterns that could explain the presence of this bubble. The question we seek to address which arose from the inverse problems is "In event of fluctuation, are there certain events, news that correlates with this type of behaviour?". Sentiment Analysis also known as Opinion Mining (Kaushik et al., 2015) is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract meaningful information in documents. This is a process of automatic extraction of features by mode of notions of others about specific product, services or experience (Kaushik et al., 2015). Sentiment analysis has been predominantly applied to obtain useful insight in business across industries, most especially in digital marketing and financial analysis. Sentiments Analysis also shows significant prominence in social media, law, policy making, sociology and even customer service and more importantly, stock markets. The study approached these by performing a collection of correlation and regression analyses to compare daily news with the period of fluctuations to identify the causes.

Using this Machine learning based approach to develop a classification model, (the "learning" part refers to choosing an "optimal" parameter given some training data) which is trained using the pre-labelled dataset of positives and negatives (Kaushik et al., 2015), This pipeline is illustrated in figure 2.7

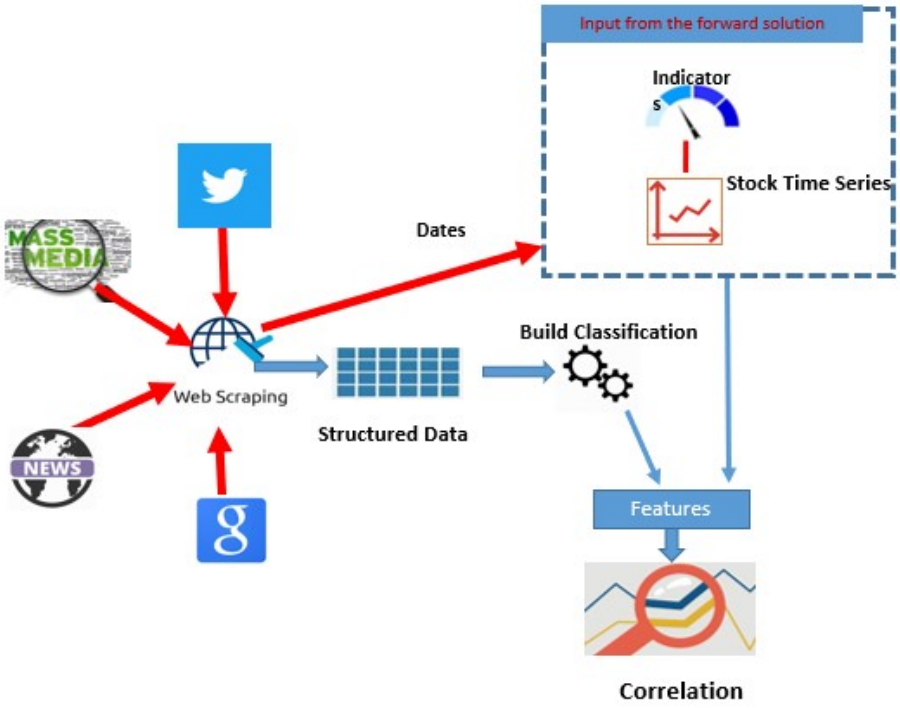

Figure 2.7: Inverse Sentiment Pipeline

As depicted in figure 2.7, the inverse process is accomplished through the use of different data sources scabbed from the web; financial news, social media, twitter and google. The data is then collated and stored in structured form which is then used to build a classifier. The result from this classifier is then merged with that gotten from the time series pressing to derive a new data-frame according to dates. Further processing is done on the resultant data frame to ascertain the degree correlation between the stock information and the sentiment analysis.

2.12 Case-based Reasoning and Inverse Problems

There are quite a few publications specifically on Case-Based Reasoning and Inverse Problem in financial domain, although some have been conducted using the ensemble methodology. Woon et al. (2004) presented a work on the ensemble to improve usability of numerical models. Their work showed that the CBR model is capable of giving fast answers to questions difficult or impossible to formulate through numerical models.

Other works are presented as stand-alone use of the techniques, some of which are the work of Sever (2017) on Machine Learning algorithm based on Inverse Problems for Software requirements selection which shows the effectiveness of multi objective inverse model (IM) approach to software requirement specification. Here, a multi objective inverse approach algorithms was used to manage conflicting objectives without the artificial adjustments included in traditional single objective optimization methods.

Michael V (2000) discussed the inverse problem in classification systems by mapping a case to its unknown category (inverse), thereby reclassifying the case as a member of a (different) preferred category, given a set of prototype cases representing a set of categories, a similarity function, and a new case classified in some category.

A steady state inverse radiative transfer problem in a one-dimensional participating media was presented in the work of Acevedo and Roberty (2010). Ulrych et al. (2001) presented a tutorial on the Bayesian approach to the solution of ubiquitous inverse problems with some special emphasis on the concepts and approach to solving inverse problems.

A study on applications of methods from the theory of inverse problems to pattern recognition was done by Sever (2015) where a new learning algorithm derived from a well-known regularization model was generated and applied to the task of reconstruction of an inhomogeneous object as pattern recognition.

Search et al. (2008) meshed a clear connection between regularization theory for inverse problems and statistical learning with recent acceleration techniques for classical gradient methods to introduce a new Machine learning algorithm on a real-world experiment concerning brain activity interpretation through the analysis of functional magnetic resonance

imaging data.

Another interesting work on inverse problem is reported in Iulian (2013). He applied inverse problem in biometric technology, by generating synthetic (false) biometric information as a way of training biometric systems and preparing them against attack from false data.

Floyd and Esfandiari (2009) presented how cases can be automatically generated in an inexpensive manner when learning by observing an expert, a method that incorporates active learning with learning by observation, a way to overcome the problem that arise from generating cases that do not contain a representative distribution of the problem space.

"A methodology for automatic generation of a quality case-base using genetic algorithm (GA)" was proposed by Manzoor et al. (2012) which was evaluated using the examination scheduling problem. This approach used the GA to generate initial case for the Case-based system by using a fitness function to evaluate a particular individual in the existing population and allowed reproduction of the accepted candidate before applying the conventional CBR cycle to the examination scheduling problem.

Xing et al. (2012) proposed a hybrid parameter estimation algorithm for predicting the strip temperature during laminar cooling process which combines Hybrid genetic algorithm with grey case-based reasoning. The model uses the GA to optimise the weight vectors of retrieval features in the CBR. The results of which show the effectiveness in improving the estimation of the strip temperature during laminar cooling process.

Two similar works that are closely related to this research idea is a report from Jenny Freeman (2018) that developed an early-warning signal for timely detection of bubble. His work uses the minimum-volume covering ellipsoids clustering method and Random transform which tackles the bubble concept geometrically by determining and evaluating ellipsoids. Radon transform which emanated from the theory of the Inverse Problems was applied to create visuals of the ellipsoids. The study reveals that the volumes of the ellipsoids gradually decrease and, correspondingly, the figures obtained by Radon transform becomes clearer which signals a stronger warning at the approach of bubble-burst time.

Another is the work by Woon et al. (2004). His work considers how CBR can be used as a flexible query engine to improve the usability of numerical models. Which helps to solve inverse and mixed problems, and to solve constraint problems.

Although this research is closely related to the work by Jenny Freeman (2018) and Woon et al. (2004), they both differ from this IPCBR framework both in approach and in concept. Not only does the IPCBR utilizes stock data in its frameworks, it is also made up of the CBR and the IP phases. The CBR phase uses clustering to retrieve similar bubble structure while the IP phase uses sentiments analysis to provide some explanations to the bubble occurrences.

2.13 Concluding Remarks

This chapter presents Bubbles with several defining characteristics; a period of over-optimism that results in the price deviating largely from fundamental valuations resulting in crisis of confidence that exposes an environment of fraudulent activities which results in a period of pessimism and economic downturn. The literature contains numerous varieties of learning techniques such as Case-bases Reasoning (CBR), Inverse Problems, and Sentiment Analysis. It presents various components CBR methodology as a fast-growing research methodology that is extremely successful in a wide range of application domains over the last few decades. The Inverse solution approach which relates to creating IPCBR framework is also presented. The chapter also presents Case based Reasoning as the chosen methodology because of its ability to handle problems with such fuzziness as the financial time series. Its also presents Inverse Problems technique as the ensemble component due to its successes in identifying input parameter from the output. These problems will be surveyed comprehensively in the remaining chapters of this thesis.

3 Methodology

.

3.1 Introduction

This chapter describes the approach that is adopted to achieve the aim of this research, which includes the overall concepts of the research methodology, data processing, feature extraction, case representation and case retrieval method.

3.2 Proposed Framework

Consequent to the complexity of the problem at hand, it was first approached by defining and solving its simplified forward form, and then with a clear definition of this, the solution of which will then be an input to the inverse problem phase. Intrinsically, the ensemble is made up two sections: The Case Based Reasoning Model and the Inverse Problem Model. First of all, the CBR model evaluates the potential indicators of all the stocks and output with potentially high yielding stocks with respect to the predefined criteria as a preselected stock set. Secondly, the results of this stock set, together with its corresponding indicators is fed into the inverse Problem Model, thus serving as the input to the inverse phase. The holistic framework is detailed in figure 3.1.

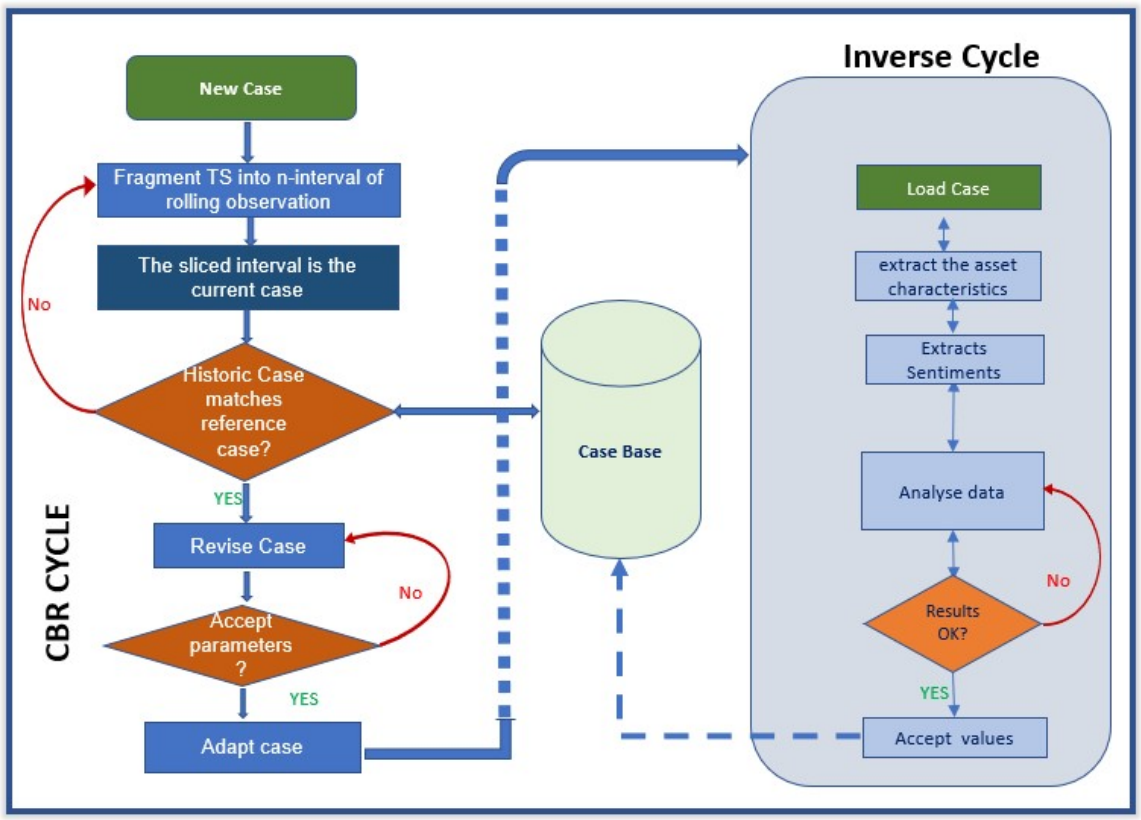

Figure 3.1: CBR/IP Model framework

With all the afore stated assumptions that define the descriptive model, The aim is to arrive at a representation of the descriptive model by calibrating the model parameters of the seed model through Case-base knowledge, which will be used to initially populate our case base. This involves representing our bubble model in a case structure which is made of historical stock projections represented by a set of points, where each point was given with the time of measuring and the equivalent stock volume. It follows from this that these processes could be represented as curves.

Then follows a Pattern Matching phase which entails the process of automatically mapping an input representation for an entity or relationship to an output category.

This involves using the new model perform pattern recognition to identify new instances that fit into the model with the use of appropriate similarity metric.

Since most classic data mining algorithms do not perform or scale well on time series data Lin, Jessica, Sheri Williamson, Kirk Borne (2011), a combination of algorithms was used for this purpose. If a perfect match is found, then the complete cycle of the CBR will be adopted and solutions adapted, otherwise, a new problem case will be reformulated. The output of this phase signifies the end of the forward problem and the solution then used as a seed for the Inverse Problem phase as shown in figure 3.1.

IP implementation requires taking the newly identified structure from the retrieved case and extract the asset characteristics around the time of the occurrence. It then requires that we identify any correlation between such characteristics and the forward problem model parameters in order to derive stochastic description of the factors that accompany the said bubbles. The output of this phase will as well be stored in the Knowledge base for easy recommendation.

3.3 The Forward Problem

To enable us to perform suitable case retrieval on the CBR model, series of experiments were conducted, stock prices analysed to identify stocks that are fluctuating in similar ways. Clustering techniques were used to pick out stocks with similar patterns of variation, as clearly defined in section 2.6, “Bubble: Representation”. In a bid to make sense of the overall bubble pattern. The necessity to use this unsupervised learning algorithm is informed by the nature of the dataset which is unlabelled.

3.3.1 Case Formulation

CBR is seldom used in time series domains mainly because it introduces some unaccustomed features that do not exist in the processing of the traditional “attribute-value” data representation and secondly, direct manipulation of continuous, high dimensional data which involves very long sequences with variable lengths is extremely difficult (Marketos et al., 1999).

In this approach, the case is created by forming a library pattern of observations and treated every group as a case category (Pecar, 2002). In which case, the entire time series is split into smaller sequences of patterns, treated individually as a case. By so doing, we closely follow the concept used in He et al. (2006), and Chen et al. (2007), where the data is created by sliding a fixed-length time window from time t_b to t_e.

resulting in $N = t_e - t_b$ time series created with a specified window length w_{tr}.

$s_1 : p1, p2, ..., pwtr$

$s_2 : p2, p3, ..., pwtr + 1...$

$s_N : pN, pN + 1, ..., pwtr + N - 1$

where $s_1, s_2, ...s_N$ represents the stocks, $pi(i = 1, 2, ..., w_{tr} + N - 1)$ are stock prices at time i. Resulting in an N by w_{tr}. matrix or a data set with N data records and w_{tr} attributes of continuous values making is easy for direct application of our chosen data mining methods.

Algorithmic Approach to Case Matching

Consequently, the concept of the CBR proposed by Aamodt and Plaza (1994); López (2013)and H. and Elmogy (2015) is aligned with the task of pattern recognition. The algorithm illustrated in figure 3.2:

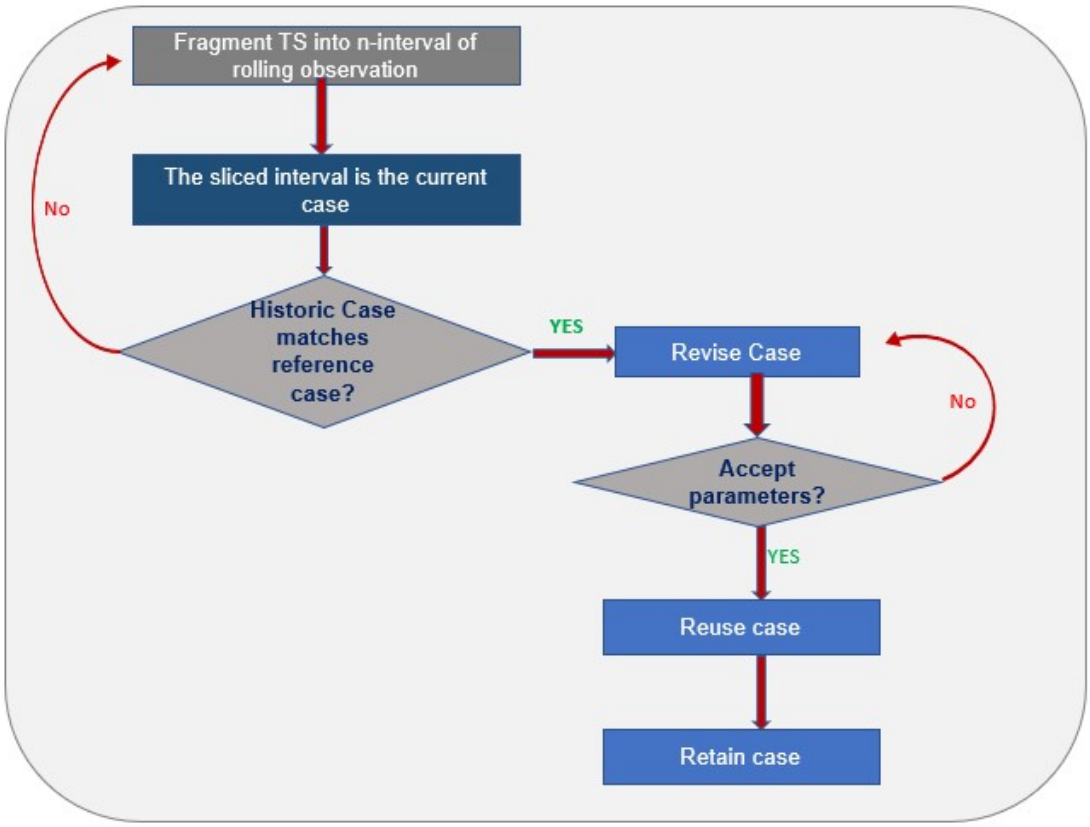

Figure 3.2: Case Matching in bubble structure

With this method, a fragment of the series is created and divided into n-interval patterns consisting of sequential rolling observation which are being stored as historical cases of n-interval patterns. This sliced interval in the series of a new observation now makes the current case to be referenced during the retrieval process. All cases satisfying the pattern matching conditions are retrieved. If there is no perfect match as the case may be, the closest cases are retrieved, otherwise the fragmentation process is done again with the new case. The retrieved cases would then be revised and tested against the stated parameters, if satisfied, then the case is then reused and the solution Retained in the database.

Retrieval Method

Similarity measure is the most essential ingredient of time series clustering and classification systems, and of great importance because one of objectives in this thesis is the task of finding concealed patterns or similar groups in data. To determine whether observations (i.e. time series) in the data are similar, however, it is a practice for one to first decide on a measure to quantify this similarity. Iglesias and Kastner (2013) states that case retrieval is one of the first decisions to be made in order to establish how the distance (similarity) between two independent vectors must be measured . As already mentioned, computing the distance (a term referred to as similarity) is sufficient for the case based retrieval. When the pattern is predefined in its analytical form (as the case with our bubble in section 2.1) one very intuitive and simple distance measure can be used.

Because of its importance, numerous approaches in view of estimating similarity have been proposed. Among these are Longest common subsequence (LCS) (Sengupta et al., 2012; Khan et al., 2013) Histogram-based similarity measure Lin and Li (2009) Cubic Spline (Wongsai et al., 2017), dynamic Time Wrapping (Zhang et al., 2011; Phan et al., 2017) have been extensively used.

Similarity in real sense is subjective, highly dependent on the domain and application. It is often measured in the range 0 to 1 [0,1], where 1 indicates the maximum similarity while 0 implies no similarity.

Similarity between two numbers x and y can be represented as:

$$nSim(X,Y) = 1 - \frac{|x-y|}{|x|+|y|} \tag{3.1}$$

When comparing two time series $X = x1,...,xn, Y = y1,...,yn$, the measures that are given according to the following equations:

mean similarity defined as:

$$MSim(X,Y) = \frac{1}{n}\sum_{i=1}^{n} nSim(x_i,y_i) \tag{3.2}$$

Root mean Square similarity defined as:

$$RMSim(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} nSim(x_i,y_i)^2} \tag{3.3}$$

Peak similarity defined as:

$$PSim(X,Y) = \frac{1}{n}\sum_{i=1}^{n}[\frac{|x_i-y_i|}{2max(|x_i|,|y_i|)}] \tag{3.4}$$

Despite the sporadic introduction of various methods of measuring similarity, the challenge of determining the best method for assignment of attributes' weight value in CBR still needs to be addressed (Ji et al., 2010).

Euclidean distance is by far the most popular distance measure in data mining, and it has the advantage of being a distance metric. However, a major demerit of Euclidean distance is that it requires that both input sequences be of the same length, and it is sensitive to distortions, e.g. shifting, along the time axis. Such a problem can generally be handled by elastic distance measures such as Dynamic Time Warping (DTW)

In this research, the dynamic time warping (DTW) is the preferred distance measure because it has shown to overcome some limitations of other distance metrics by using dynamic programming technique to determine the best alignment that will produce the optimal distance. DTW is extensively used technique in speech recognition (Cassidy, 2002; Xihao and Miyanaga, 2013) and many other domains, including Time Series analysis (Phan et al., 2017; Zhang et al., 2011).

3.4 The Inverse Phase

This is where the results in the forward phase is adapted to serve as input to the Inverse Phase. This is achieved by conducting a Sentiment analysis on the sliced portion of data representing the period of fluctuation.

Sentiment Analysis, which depend on techniques utilized by Natural Language Processing (NLP) (Devika et al., 2016; Paper, 2021), builds models capable to computationally identify and categorize opinions expressed in a piece of body work with the aim to determine whether the writer's attitude towards a chosen subject, has positive, negative, or neutral tone.

From the Machine learning perspective, Sentiment Analysis is a supervised learning task, where a group of phrases is equipped with labels of their respective sentiments to the machine learning model, and the model is tested on unlabelled phrases. It is considered "Inverse" because, the input parameters is derived from results of the forward phase and the solution is used therefore to identify the possible cause of the effect of the forward problem. This follows the inverse pipeline as illustrated in figure 2.7

In this regard, the IP implementation requires taking the newly identified structure from the retrieved case and extracting the asset characteristics around the time of the occurrence. It is then required that we identify any correlation between such characteristics and the forward problem parameters in order to derive stochastic description of the factors that accompany the said bubbles. The output of this phase will as well be stored in the Knowledge base for future recommendation, consequently this completes the CBR cycle.

3.4.1 Data pre-processing

Experiments performed to calculate distance or similarity between vectors are done directly on the raw data, these however work fine in various domains, but considering that one is dealing with time series, it is often reasonable clean the data, and then transform the time-series sequences, to have an insight into the general nature of the data, reduce noise

highlight trend-related information. This is addressed in subsequent sections accordingly because the research were done using a combination of datasets.

3.5 Chapter summary

This chapter has presented the methodology that is adopted to achieve the aim of this research, which includes the overall concepts of the research methodology, data processing, feature extraction, case representation and case retrieval method.

It also details a suitable case formulation for the representative candidate object which has specified 'bubble' characteristics. The neighbour network is built based on the similarity of time series objects which is measured by suitable similarity metrics.

Also addressed is case formulation, bringing to focus how best the complex case nature as bubble is formulated and aligned with pattern recognition tasks for easy retrieval.

Furthermore, the inverse formulation based on the sentiments analysis is also introduced.

4 Case Retrieval

4.1 Introduction

This chapter describes the retrieval process of the framework. The chapter begins by addressing the concept of distance with respect to time series. It also gives insights of some popular distance/similarity measures, identifying their merits and drawbacks. With these insights, the most suitable similarity measure for the shape/pattern matching tasks is selected.

4.2 Distance(Similarity)

The term distance is inferred when one is measuring a length or interval between two points by considering the complete path of motion from one point to another. Displacement is the vector measure of an interval between two locations and describes the shortest path connecting them. But when a distance is done with respect to time series, it presents a different scenario, The shape of input vectors entails features that are arranged in time, whereas in univariate time series, an input vector is typically the sequence of values that a definite variable takes throughout a specific time space.

Let $d(x,y)$ represent the distance between time series (observations) $x = (x_1,...,x_n)$ and $y = (y_1,...,y_m)$. In which case, time series x and y can be seen as numerical vectors with dimensions n and m, such that $x \in Rn$ and $y \in Rm$. where m and n can thus differ in value. The value of $d(x,y)$ for every pair of x and y is a single non-negative number that depends

on the distance measure that is used. The lower the value of $d(x,y)$, the closer the two observations are depending on the chosen distance measure. The distance between two items depends on both the representation used by the feature vector and on the distance measure used.López (2013). The similarity measure(s) can be gotten from distance measure "d" from the simple equation given by $s = 1 - d$. The total number of time series will be indicated by N.

4.2.1 Similarities in Time Series

Time series analysis presents a very interesting research domain with a great number of applications in business, economics, finance and computer science (Zhang et al., 2011). A time series is a sequence of data points obtained from consecutive observations or measurements over a specific time interval (Guo et al., 2016), it tracks the movement of the chosen data points, such as the stock price over a specified period of time with data points recorded at regular intervals.

The main aim of time series analysis is to study the path observations of time series and build a model to describe the structure of data and predict the future values of time series.

A usual representation is a set of points, where a point is an ordered pair (x,y). Very often the pairs are (t,v) where t represents time and v represents a value at the time t. A point in such cases can be a result of an experiment, a result of measuring, or a value of some stocks at a given time etc. When the data is given in this way (as a set of points) it can be represented graphically. Furthermore, if the points are connected they represent a kind of curve.

A distance measurement between time series is often needed to determine similarity between time series and for time series classification. The major challenges, as with any CBR system, is how to select the best measure of similarity, because Time Series data has features that are not completely independent of each other.

4.3 Categories of Distance Measures

The function used to measure the similarity or distance between two time series objects is one key component in the time series clustering algorithm (Zhang et al., 2011), hence the reason for existence of various similarity (distance) measures. While this gives rise to varying choices, finding concealed patterns or similar groups in data still remains a common objective. This section gives an overview of the most commonly used distance measures with a view of determining which is best suited measure for the IPCBR framework. This presentation is based on Esling and Agon (2012) and adopted from the work of Roelofsen (2018), that put the measures into four different categories viz.

i Shape-based distances

   Lock-step measures

   Elastic measures

ii Feature-based distances

iii Edit-based distances

iv Structure-based distances

The shape-based distances makes a comparison of the complete shape of time series based on the actual (scaled) values of the time series. In this category, two subcategories are also identified viz lock-step measures and elastic measures. Whereas the lock-step require both time series to be of equal length $(n = m)$and compare time point i of time series x with time point i of time series y, the elastic measure shows more flexibility. It uses dynamic programming to align sequences with different lengths, and allow one-to-many or one-to-none point matchings. The feature-based distances first extracts features from the time series and then measures the distance between these features. Feature-based distances are often used in a situation where the aim is to obtain a reduction in both dimensionality and noise (Qian et al., 2015). The edit-based distances is used to compute the dissimilarity

of two time series based on the minimum number of operations that are needed to change or transform one series into the other series. In the last category, structure-based distances, the dissimilarity between time series is obtained by comparing higher level structures that are obtained by modelling or compressing the time series.

Our major focus is on shape-based and feature-based distance measures. This is because these measures are greatly used in time series clustering and therefore, we expect these methods to be most promising.

Lock step Measures

Considering that $X = \{X_1, X_2, ..., X_n\} and Y = \{Y_1, Y_2, ..., Y_m\}$ are two time series of equal length $n = m$ and elements X_i and Y_i are vectors in some d-dimensional space.

The lock step measure is often referred to L_p-norms meaning that the i^{th} instance from one series n is compared to the i^{th} instance of another series m, therefore the lock-step distance measures require both time series to be of equal length $(n = m)$. Although this these measures can also be used for non-time series clustering assignments. The only requirement is that all observations are numerical vectors of equal length. The following are some popular metrics that fall under this category.

4.3.1 Minkowosi Distance

Minkowosi Distance metric is a form used for the calculating distance for multidimensional data.It generalizes a wide range of distances such as the Hamming and the Euclidean distance (Merigó and Casanovas, 2011). The normalized Minkowski distance distance D between two parts X and Y in its generic form is given by

$$D(XY) = \left(\sum_{i=1}^{d} |x_i - y_j|^{\frac{1}{\lambda}} \right)^{\frac{1}{\lambda}} \tag{4.1}$$

where x_i and y_i are the ith arguments of the sets X and Y and λ is a parameter such that $\lambda \in (-\infty, \infty)$. y

4.3.2 Hamming distance:

Hamming distance, a natural similarity measure (Norouzi et al., 2012) is a possible choice of measure if the feature vectors are binary. Formally, the hamming distance $d(x,y)$ between two vectors $(x,y) \in F^{(n)}$ (n is the number of binary digits) is the number of coefficients which they differ. In other words, the Hamming distance between two binary sequence of equal length could be determined by computing the length of positions for which the corresponding symbols are different. For example, the Hamming Distance for $F^{(8)}$ $d(01110110, 11100101)$ is 4. In case of a string, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other. Widely used search engines, such as Google, Baidu, and Bing, use Hamming-distance search in their image content-based search engines (Tang et al., 2015).

4.3.3 Euclidean distance

Euclidean distance (Berthold and Höppner, 2016; Liberti et al., 2014), also known as simply distance between points M and N is the length of the line segment connecting them (MN) . This distance between two points is given by the Pythagorean theorem that theorem applied to p dimensions rather than the usual two dimensions. Equation and expressed as:

$$D_{ij} = \left(\sum_{i=1}^{d} |x_{il} - x_{jl}|^{\frac{1}{2}} \right)^2 \tag{4.2}$$

In other words, it is a special case of Minkowski distance at norm value of 2. This metric has been receiving increased attention in scientific research and found applications in psychometrics, crystallography, machine learning, wireless sensor networks, acoustics, and more (Dokmanic et al., 2015)

4.3.4 Cosine similarity

Cosine similarity (Garcia, 2015), is a popular vector based similarity measure in text mining and information retrieval. The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that computes cosine of the angle θ between them by this, its judgement is based on orientation rather than magnitude (Sitikhu et al., 2019). The metric finds normalized dot product of the two attributes, say A and B.

$$\overline{A}.\overline{B} \tag{4.3}$$

now, to build the cosine similarity equation is to solve the equation of the dot product for the $\cos\theta$:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta \tag{4.4}$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|} \tag{4.5}$$

Here the cosine similarity given in equation 2.5 will generate a metric that determines how two documents relate by looking at the angle instead of magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors.

4.3.5 Pearson correlation distance

Correlation is a technique for investigating the relationship between two quantitative, continuous variables and gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. The Pearson correlation distance takes into account the linear association between two vectors or variables. It does so by using the Pearson correlation coefficient, which is defined as follows:

$$\rho(x,y) = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{4.6}$$

where x and μy are the means of x and y and ρx and ρy are the standard deviations of x and y, respectively. It is worthy of note that the coefficient requires $n = m$ and that it is invariant for scaling. The values of ρ lie within the range $[+1, -1]$ and, where $\rho = 1$ evidences a seamless positive relationship between x and y, $\rho = -1$ indicates a perfect negative relationship between x and y and $\rho = 0$ indicates no relationship between the two variables. The magnitude of the relationship of other values that lie within this range varies based on the value of ρ, where values closer to -1 and 1 indicate a stronger negative or positive relationship, respectively. Pearson correlation has recorded wide applicability in clustering gene expression data (Monks et al., 2004; Barido-Sottani et al., 2019).

Elastic measures

The elastic distance measures allow one-to-many or one-to-none point matching, which is an edge over the lock step measure. This ability makes it possible for elastic distance measures to warp in time and be more robust when it comes to handling outliers for instance. This also has a demerit of an increase in time complexity. This subsection discusses some elastic distance measures:

4.3.6 Longest Common Subsequence Similarity

The longest common subsequence (LCS) similarity measure problem has been extensively studied for several decades and has many practical applications; speech recognition (Guo and Siegelmann, 2004; Ko et al., 2010) text/string pattern matching (Dash and Nayak, 2013), and Time series classification (Guo et al., 2016) e.tc. Its major task is to find the longest subsequence common to the input sequences (usually two sequences) (Hasna and Potolea, 2016). The advantage of the LCSS method is that some elements may be unmatched or left out (e.g. outliers), while in the Euclidean distance and DTW, all elements from both sequences must be used, even the outliers.

For a formal representation, let $C = (c_1, c_2, \ldots, c_m)$ and $P = (p_1, p_2, \ldots, p_n)$ be two sequences of length m and n, respectively. Let L(i, j) denote the length of longest common sub-sequence $(c_1, \ldots, c_i) and (p_l, \ldots, p_j)$. $L(i, j)$ The recursive algorithm could be represented as :

$$IF c_i = q_j THEN$$

L(i, j) = 1 + L(i - 1, j - 1) ELSE L(i, j) = max L(i - 1, j), L(i, j – 1) And the distance/dissimilarity represented as

$$LCSS(C, Q) = \frac{m + n - 2l}{m + n} \tag{4.7}$$

where l is the length of the longest common subsequence. Intuitively, this quantity determines the minimum (normalized) number of elements that should be removed from and inserted into C to transform C to Q. Like the case with dynamic time warping, the LCSS measure can be computed by dynamic programming in O(mn) time (Fakhrazari and Vakilzadian, 2017).

4.3.7 Dynamic Time Warping

Dynamic Time Warping is a time series distance measure that is introduced in the field of data mining to overcome some of the disadvantages of the Euclidean distance (Keogh and Ratanamahatana, 2005). This technique finds the optimal alignment between two time series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis (Salvador and Chan, 2018). The algorithm, originally conceived by Vintsyuk (1968), has since been applied in many fields including spoken word recognition (Technologies et al., 2013), gesture recognition (Choi and Kim, 2018), pattern recognition in stock market(Zhu, 2011; Lee, Suk Jun, Jeong, 2012), behavioural perception (Iizuka et al., 2013; Switonski et al., 2019). This measure does not obey the triangular inequality and thus has resisted attempts at exact indexing (Keogh and Ratanamahatana, 2005).

The DTW minimizes distortion effects due to time-dependent movement by using an elastic transformation of time series data to recognize the similar phases between different patterns along time. In event of any deformation relationship between two different sequences of time series data, the DTW determines the most similarities between them. The main feature of this distance measure is that it allows to recognize similar shapes, even if they present signal transformations, such as shifting and/or scaling The algorithm gives not only the information of how alike they are, but also the best correspondence among their data prints. In addition, the procedure is quite flexible, so it can be easily adjusted to the type of data that one intends to operate in, hence the decision to apply this in our framework.

The mathematical procedure can be translated into a succession of steps that results in the comparison of two different sequences. To illustrate this, consider two time series T = $\{t_1, t_2, ..., tn\}$ and S = $\{s_1, s_2, ..., sm\}$ lengths of x and $y \in N$ respectively, an alignment by DTW method exploits information contained in a n x m distance matrix:

$$\mathrm{d}istMatrix = \begin{pmatrix} d(T_1,S_1) & d(T_1,S_2) & \dots & d(T_1,S_m) \\ d(T_2,S_1) & d(T_2,S_2) & \dots & ,d(T_2,S_m) \\ \vdots & \vdots & \ddots & \vdots \\ d(T_n,S_1) & d(T_n,S_2) & \dots & ,d(T_n,S_m) \end{pmatrix} \tag{4.8}$$

where distMatrix(i, j) corresponds to the distance of ith point of T and jth point of S $d(Ti, Sj)$, with $1 \leq i \leq n$ and $1 \leq j \leq m$.

The DTW objective is to find the warping path W = $w_1, w_2, ..., w_k, ..., w_K$ of contiguous elements on $distMatrix(withmax(n,m) < K < m+n-1$, and $w_k = distMatrix(i,j))$, such that it minimizes the following function:

$$DTW(T,S) = min\left(\sqrt{\sum_{k=1}^{k} w_k}\right) \tag{4.9}$$

And the warping path is subject to several constraints. Given $wk = (i, j)$ and $w_{k-1} = (i', j')$ with $i, i' \leq n$ and $j, j' \leq m$:

- Boundary conditions. w1 = (1,1) and wK = (n, m).
- Continuity. i–$i' \leq 1$ and j–$j' \leq 1$.
- Monotonicity. i–$i' \geq 0$ and j–$j' \geq 0$.

The warping path can be efficiently computed using dynamic programming. By this method, a cumulative distance matrix ρ of the same dimension as the distMatrix, is created to store in the cell (i, j) the following value

$\rho(i,j) = d(T_i,S_j) + min\rho(i-1,j-1j),\rho(i-1,j),\rho(i,j-1)$

The overall complexity of the method is relative to the computation of all distances in distMatrix, that is O(nm). The last element of the warping path, w_K corresponds to the distance calculated with the DTW method.

For this purpose, the algorithm Dynamic Time Warping (DTW) is used, which along with some pre chosen indexes, will assist in the gathering of the numerical results.

Feature-based distances

Feature extraction and representation is essential in machine learning research. This measure calculates the degree of likeness or proximity between two time series based on the number of common features between the series. The central idea of this approach is to improve classification accuracy. The preceding section discusses some of the feature extraction techniques.

4.3.8 Principal Component Analysis (PCA)

Principal component analysis (PCA) is an effective method of multivariate statistical data analysis. It is a linear transformation that projects the original (Gaussian distributed) data to a new coordinate system on the premise of little or no loss of the raw information (Bankó et al., 2011), it can converse complex data which has correlation to less number and unrelated integrated indicators to each other by linear combination, and thus it will achieve

the goal of dimension reduction (Cao et al., 2015). For instance, considering a dataset using an m x n data matrix whose ij element is the $j'th$ sample of the $i'th$ variable. If the m-dimensional data disperses along chosen directions of variability, but does any distinctive clusters, then one would consider the principal component analysis (PCA), as a better approach. PCA is widely applied in many fields, where it goes by a variety of alternative names and used in time-series forecasting as reported in Cornillon et al. (2008),

4.3.9 Singular Value Decomposition

Singular Value Decomposition (SVD) has been initially successfully used for indexing images and other multimedia objects and recently has been proposed for time series indexing (Chan and Fu 1999; Korn et al. 1997).

The SVD is a generalized form of matrix diagonalization which postulates that any *mxn* rectangular matrix A can be factorized into the product of three matrices as

$$A = UDVT$$

where the columns of U and V are orthonormal corresponding to m and n-dimensional rotations, respectively, and the matrix D is diagonal with positive real entries.

SVD differs from DFT and DWT in the sense that the later are local, which implies that they examine each data object at a time before applying a transformation which makes these transformations completely independent of the rest of the data. Whereas, SVD uses a global transformation mode by examining the whole dataset and then performs rotation such that the first axis has the maximum possible variance, the second axis has the maximum possible variance orthogonal to the first, the third axis has the maximum possible variance orthogonal to the first two, and so on.

The SVD is useful in many tasks; Schanze (2018) applied the method to simulated and biomedical signals, gene expression profiles (Allanki et al., 2017; Tsz et al., 2010), Khoshrou and Pauwels (2019) reported that SVD can also be successfully applied to identify periodic patterns (profiles) in time series.

4.3.10 Fourier Transformation (FT)

Virtually everything in the world can be described via a waveform such as sound waves, electromagnetic fields, as well as stock elements versus time. The Fourier Transform presents a unique and powerful mathematical tool that shows a way of visualizing these waveforms and to deconstruct the waveform into its sinusoidal components.

The Discrete Fourier Transform (DFT), which is normally computed using the counterpart Fast Fourier Transform (FFT), has revolutionized modern society. The Fourier analysis methods are nowadays frequently implemented into an algorithmic trading as a Technical Analysis tool for a directional forecasting of a market price development (STADNIK et al., 2016), especially in noise filtering (Huang et al., 2014). To perform the dimensionality reduction of a time series T of length n into a reduced feature space of dimensionality N, the Discrete Fourier Transform of T is calculated. For instance, if there are 32 data points, the FFT would first split the data into two sets of 16 points, then split the data into four sets of 8 points, then split that data again into eight sets of 4 points, and so on; at each stage of the splitting, the FFT would calculate the DFT and the results would all be combined in order to calculate the final FFT of the data. However, one of the setbacks is that the FT gives only the frequency information of the series, but not the time information simultaneously which therefore implies that FT is suitable for stationary series whose frequency does not change in time, but not for non-stationary series.

4.3.11 Wavelet Transformation

Wavelet analysis is a mathematical model very suitable for non-stationary data that transforms the time domain signal into a different domain for analysis and processing. With functionality, it can decompose the signal in both time and frequency domain simultaneously as compared to the FT.

The basic idea of wavelet transform is to decompose a signal to the number of basis

signals which can be written to the wavelet basis functions as follows:

$$\sum_{j,k} a_{j,k} 2^{j/2} \psi(2^j t - k)$$

where $a_{j,k}$ is the coefficient, $\psi(2^j t - k)$ is wavelet which is scaled by the indices j and is shifted by k, j is the scaling index, and k is shifting index. Function f(t) is any signal that varies with time or other variables.

It is much similar to DFT. However, one important difference is that wavelets are localized in time, i.e. some of the wavelet coefficients represent small, local subsections of the data being studied. This is in contrast to Fourier coefficients that always represent global contribution to the data.

The limitations of DWT is the requirement of length of the dataset to be in power of 2, and the dependency of the generated output on origin of the signal being analysed. This makes it difficult to align the transformed signals with time because a small shift in origin impacts remarkably on the outputs generated.

4.3.12 Piecewise Linear Approximation

Piecewise Linear Approximation (PLA) is a well-established tool to reduce the size of the representation of time series by approximating the series by a sequence of line segments while keeping the error introduced by the approximation within some predetermined threshold. This representation has numerous advantages, including data compression and noise filtering, and has been used by various researchers to support clustering, classification, indexing and association rule mining of time series data (Stone and Clinton, 2016).

PLA is widely applied to numerous fields; wireless sensor networks (Berlin and Van Laerhoven, 2010), biomedical (Lee and Singh, 2012) etc., because it has a good ability to compress natural signals, fast linear algorithm exists, supports weighted measures etc. The main disadvantage of PLA is that it is not indexable by any data structure (but the sequential scan very fast).

4.3.13 Piecewise Aggregate Approximation(PAA)

Piecewise Aggregate Approximation (PAA) is an approach of average dimensional reduction, which divides the time sequence equally and take the mean value of each segment as representation (Zhang et al., 2019) Though highly promising, PAA requires estimating the number of segments required for the representation a priori. The process works by approximating a time-series X of length n into vector $\overline{X} = (\overline{x}_1, ..., \overline{x}_m)$ of any arbitrary length $M \leq m$ where each of $\overline{x}$ is derived from:

$$\overline{x} = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{m(n/M)i} x_j$$

meaning, to reduce dimension from n to M, first divide the original time-series into M equally sized frames and then compute the mean values for each frame.

PAA has recently been applied in many fields like Biomedicine Zhang et al. (2019), mainly because of its numerous advantages. Some of them are: extremely fast to calculate, supports queries of arbitrary lengths, supports any Minkowski metric and non Euclidean measures, supports weighted Euclidean distance simple, intuitive, etc.

4.3.14 Adaptive Piecewise Constant Approximation

As an extension to the PAA representation, Adaptive Piecewise Constant Approximation (APCA) is introduced in Chakrabarti (2002). This representation allows the segments to have arbitrary lengths, which in turn needs two numbers per segment. The first number records the mean value of all the data points in a segment, and the second number records the length of the segment.

4.3.15 Symbolic Aggregate Approximation

One of the most interesting and promising representation methods that allows for dimensionality reduction and indexing with a lower-bounding distance measure is the Symbolic

Aggregate Approximation (SAX) (Lin et al., 2007). SAX is as good as well-known representations such as Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT), while requiring less storage space as added advantage and behaves like an aggressive form of smoothing. Given a time series of some length n, SAX reduces it to a string of arbitrary length w where $w \leq n$. For making comparisons, this is extremely efficient, since it greatly reduces the dimensionality of the original sequence. There is great potential for extending and applying the discrete representation on a wide class of data mining tasks.

4.3.16 Recurrent Neural Network (RNN)

Is a special case of neural network where the objective is to predict the next step in the sequence of observations with respect to the previous steps observed in the sequence of arbitrary length. Ideally, the idea behind RNNs is to make use of sequential observations and learn from the earlier stages to forecast future trends. As a result, the earlier stages' data need to be remembered when guessing the next steps. In RNNs, the hidden layers act as internal storage for storing the information captured in earlier stages of reading sequential data. More formally represented, given a sequence

$$x^{(1:n)} = (x^{(1)}, x^{(2)}, ..., x^{(t)}, ..., x^{(n)})$$

the RNN updates its recurrent hidden state $h(t)$ by

$$h^{(t)} = g(Uh^{(t-1)} + Wx^{(t)} + b)$$

where g is a nonlinear function. Despite its popularity, it can be difficult to train them to learn long term dynamics, likely due in part to the vanishing and exploding gradient problem (Hochreiter and Schmidhuber, 1997a).

4.3.17 Long Short-Term Memory model(LSTM)

Long Short-Term Memory model (Hochreiter and Schmidhuber, 1997b) has brought lots of transformations to machine learning field. LSTM networks are specifically designed

to learn long-term dependencies and are capable of overcoming the previously inherent problems of RNNs, which is, vanishing and exploding gradients (Hu et al., 2018; Lee et al., 2017). It has been particularly successful in language translation, text classification tasks and widely used applications in complex prediction problems. Liu et al. (2015) used LSMT model for Chinese word segmentation, Hu et al. (2018) Deployed ANN and LSTM network models for simulating the rainfall-runoff process based on flood events, Stock Market Prediction (Zhou et al., 2018), as well as stock price forecasting reported in Chen et al. (2017)

4.4 Summary

A survey of literature reveals that in spite of plethora of similarity measures, they generally have certain drawbacks (Sengupta et al., 2012), traditional lock step distance measures are widely used for typical clustering tasks, most predominantly, the Euclidean distance and Pearson correlation distance. While these give better performances for a wide variety of clustering applications, using them on time series often give inferior performances or are often outperformed by other distance measures. This is due to some limitations that lock-step measures have, being unable to handle cases of shifts in time and also may not process noise and outliers in a desirable way. The elastic measures such as DTW and LCSS are often used in to counter such limitations. Both distance measures warp the time series to account for shifting in time. LCSS also accounts for noise and outliers, depending on the threshold setting (Guo et al., 2016), which makes LCSS more robust than DTW under noisy conditions. however, the Euclidean distance is also used as the local distance measure in DTW, while LCSS makes use of the Manhattan distance for similar purposes. Despite the good performance of the elastic measures in finding similarity between time series, they are computational more expensive because they come with a significant increase in calculating time. According to Lin and Li (2009) the classification accuracy difference between the Euclidean distance and elastic distance measures converges to zero as the

number of time series increase. Although this also affect the lock step measures when the dataset is considerably large, this problem is significantly handled through the use of feature based measures to perform dimensionality reduction. Besides reducing the number of dimensions, the feature-based distance measures are also capable of reducing noise by eliminating certain levels of detail in the approximations.

5 Machine learning Interpolation Methods

5.1 Introduction

This chapter presents a brief survey of some selected interpolation methods mostly used in case retrieval and from the result of the experiment, suggest the ones that would be most suitable for the model. Four interpolation methods considered in this experiment: k-nearest Neighbour, Support Vector Machine, Logistic Regression, and Linear Discriminant Analysis (LDA), are suitable to both linear and non-linear model values. The rationale behind the choice is to ascertain how these interpolation methods perform on classical traditional dataset before applying them to Time series dataset that is characterized with high dimensionality and noise. Similar experiment were done using the Respiratory Disorder dataset that was originally published in Everitt et al. (2017). All these were done to determine the quality of the created model with the use of some statistical methods to make estimate on the accuracy of the models.

5.2 Interpolation Methods

Interpolation is a process for estimating values that lie between known data points, the process of which are well studied in real domain. It has proven to give good results from relatively sparse datasets being typical in CBR systems (Knight et al., 2010). Interpolation

tools can provide solutions to unknown problems by adapting solutions from other problems already solved. The interpolation method is then used to select an appropriate solution value from a solution domain. The type of interpolant to use depends on the characteristics of the data being fit, the required smoothness of the curve, speed considerations, post-fit analysis requirements, and so on. for instance, the linear and nearest neighbour methods are fast, but the resulting curves are not very smooth. The discussion on several interpolation approaches is given in the proceeding section.

5.3 K-nearest-neighbor KNN

The k-NN algorithm is a non-parametric method, is perhaps the simplest of all algorithms for predicting class of a test example (Sutton, 2012; Kumar and Kumar, 2015) which is usually used for classification and regression problems because of its simplicity and ease of implementation. To perform classification of an unknown instance represented by some feature vectors as a point in the feature space, the k-NN classifier applies one of the distance functions (usually the Euclidean distance), calculates the distances between the point and points in the training data set, and then assigns the point to the class among its k nearest neighbours (where k is an integer). Formally represented, the k-nearest neighbour method uses the k observations in the training set closest to the point on the background space grid to form

$$\hat{Y} = \frac{1}{k} \sum_{i=1}^{n} y_i \tag{5.1}$$

where y_i is the ith case of the examples sample and $\hat{Y}$ is the prediction (outcome) of the query point. In contrast to regression, in classification problems, KNN predictions are based on a voting scheme in which the winner is used to label the query.

The dataset can be represented as a matrix , containing scenarios where each scenario contains features . A vector with length of output values accompanies this matrix, listing the output value for each scenario. The algorithm can be used for approximating both discrete-valued target and continuous-valued target function.

5.4 Logistic regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. It is very much like linear regression and represented by the equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{5.2}$$

it accepts some input values (x) and combines them linearly using weights or coefficient values (represented with a symbol) to predict an output value (y). But the difference is that the output value being modelled is a binary value (0 or 1) unlike linear regression which outputs continuous number values. For it to map predicted values to probabilities, the sigmoid function is used

$$g(z) = \frac{1}{1 + exp(-z)} \tag{5.3}$$

that maps any real value into another value between 0 and 1. In our machine learning, the sigmoid is used to map predictions to probabilities.

How it works: Let us consider a model that consists of a vector β in d-dimensional feature space, then for a point x in feature space, project it onto β to convert it into a real number z in the range -θ to + θ.

$$z = \alpha + \beta x = \beta_1 x_i + ... + \beta_d x_d \tag{5.4}$$

Maps z to the range 0 to 1 using the logistic function using equation 5.2

5.5 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes, and it cares only about the data points near the class boundary and finds a hyperplane that maximizes the margin between the classes.

A Support Vector Machine models the situation by creating a feature space, which is a finite dimensional vector space, each dimension of which represents a "feature" of a particular object. In the context of our case, each "feature" is the prevalence or importance of a particular field.

Let us assume that we have n labelled examples $(x_1, y_1), \ldots, (x_n, y_n)$ with labels $y_i \in \{1, -1\}$. We want to find the hyperplane $< w, x > +b = 0$ (i.e. with parameters (w, b)) satisfying the following three conditions:

1. The scale of (w, b) is fixed so that the plane is in canonical position w.r.t. $\{x_1, \ldots, x_n\}$. i.e.,

$$\min_{i \leq n} | < w, x_i > +b| = 1$$

2. The plane with parameters (w, b) separates the $+1$'s from the -1's. i.e.,

$$y_i(< w, x_i > +b) \geq 0 \text{ for all } i \leq n$$

3. The plane has maximum margin $\rho = 1/|w|$. i.e., minimum $|w|^2$.

Clearly 1 and 2 combine into just one condition:

$$y_i(< w, x_i > +b) \geq 1 \text{ for all } i \leq n.$$

Thus, we want to solve the following optimization problem,

$$\text{minimize } \frac{1}{2}|w|^2$$

over all $w \in R^d$ and $b \in R$ subject to,

$$y_i(< w, x_i > +b) - 1 \geq 0 \text{ for all } i \leq n.$$

This results in a very simple quadratic programming problem with already existing algorithms.

5.6 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), originally developed in 1936 by R. A. Fisher (James et al., 2000) is a means to assess how perfectly a group of variables retains an a prior classification of objects. The goal is to project a dataset onto a lower dimensional space with good class separability in order avoid over-fitting and also reduce computational costs. This model assumes that the conditional distribution of the features given in the class is Gaussian (normal distribution). This assumption is not explicitly true for the dataset used in this research. Nevertheless, a fairly good result is expected because much variability is not present in the class. For this multiclass formulation of LDA. A one-vs-one approach is used, which involves training a classifier for each pairwise combination of classes

5.6.1 Experimental Dataset

This dataset was simulated based on the literature obtained from drill operations as stated in 1.6.1. Some relevant features were selected to form the sample dataset. Descriptions of situations in these reports contain the occurring problems and their proposed solutions resulting in 278,040 observations. This dataset was then used in the experiments to determine the goodness of the created model and used some statistical methods to make estimate on the accuracy of the models that was created. To test the performance of the model, 10-fold cross validation was applied to the dataset and all the four classifiers were experimented. The datasets were partitioned into training set for training the model and test set for testing the model. For the choice of samples in training the percentage of each class was balanced statistically, while for testing portions, the percentage of each class in each portion is preserved. The training and the testing portions were further adjusted in the ratio of 10:90, 30:70, 50:50, 70: 30 and 90:10.

The second was a labelled dataset from a clinical trial that was used in comparing two treatments for a respiratory disorder dataset as stated in 1.6.2. It was originally published in (Everitt et al., 2017), which involves eligible patients that were randomly assigned to

active treatment or placebo, each at four monthly visits.The trial recruited 111 participants (54 in the active group, 57 in the placebo group). Their respiratory status was determined as being "poor" or "good", which constitutes the categorical output. Covariates such as centre, sex and age were also recorded. The goal was to evaluate the effect of the treatment on the respiratory status based on the output.

5.6.2 Data Pre-processing

The pre-processing was done to identify a subgroup of features that have a greater significance over the others because it has a very significant impact in the precision on estimate models and to enhance the modelling characteristics.This procedure was carried out through the following steps:

1. Critical examinations on all relevant records with the aim of identifying and isolating incomplete or missing data. These records were ignored if found.

2. Identifying and removing redundancy in the data.

3. Assessing the relevance of each feature to ensure maximum entropy.

4. If necessary, to transform the features corresponding to the model to enhance a normality.

The First experiment was done with the objective of learning from the simulated data as well as have good generalization performance. By so doing, a way to manage model complexity and a method to measure performance of the chosen model is also required. A common approach of achieving the set objective is to divide the data into three sets, training, validation and test (Dobbin and Simon, 2011).

1. Training set: Training data are used to learn or develop candidate models. It consists of a randomly selected proportion of the transaction data where the class is known

2. Validation set: This actually can be regarded as a part of training set, because it is used to build the model and consists of a sample of data where the outcome is known and is used only once the model has been completed to validate the accuracy of the model. It is a sample of data held back from training the model that is used to give an estimate of model skill while tuning model's hyperparameters.

3. Test set: Consists of a sample data that is outside the training set that is used to provide an unbiased evaluation of a final model fit on the training dataset, or used to evaluate how well the model does with data outside the training set.

However, in many applications as well as in this research, only two sets are created; training and test.

To test the performance of this model, 10-fold cross validation was applied to the dataset and were experimented with selected interpolation methods to be used as classifiers. The datasets were partitioned into training set for fitting the model and test set for measuring the accuracy of the model.

The samples in training the percentage of each class were balanced statistically, while for testing portions, the percentage of each class in each portion was preserved. The training and the validating portions were further adjusted in the ratio of 10:90, 30:70, 50:50, 70: 30 and 90:10.

To determine which algorithm will be more appropriate for our model, some selected interpolants were used; Two simple linear algorithms; Linear discriminant analysis (LDA) and logistic regression (LR), and two nonlinear k-Nearest Neighbours (kNN) and Support vector machines (SVM) algorithms. LDA and LR are widely used multivariate statistical methods for analysis of data with categorical outcome variables. Both are appropriate for the development of linear classification models associated with linear boundaries between the groups. While the kNN, determines the nearest k training instances to a target instance. SVM on the other hand, attempts to find a hyper-plane separating the different classes of the training instances, with the maximum error margin.

Furthermore, random number seed were reset before each run to ensure that the evaluation of each algorithm is performed using the same data splits. This is to ensure that the results are directly comparable. The different models were then later run directly on the validation set and the results summarized as a final accuracy score, a confusion matrix and the outcome of the classification presented.

5.7 Results and Evaluation

This section presents some experimental results on how the selected interpolatants perform on the experimental drill dataset.

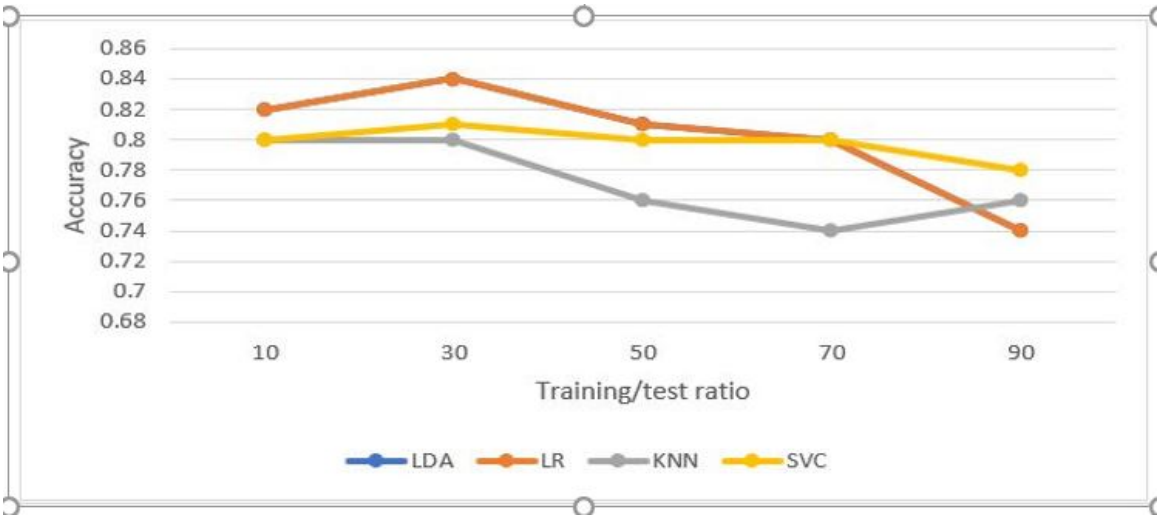

Figure 5.1: Accuracy for the Classifiers

Figure 5.1 shows results from the comparison of the classification accuracy for the selected interpolatants used as classifiers. The results obtained shows that LR and LDA are overlapping in terms of performances, and in identifying the failed process with steady decrease as the percentage ratio is increased. KNN, and SVM, also show steady decrease as the percentage ratio was increased. The overall best performance was seen at 70:30 training/test ratio variation.

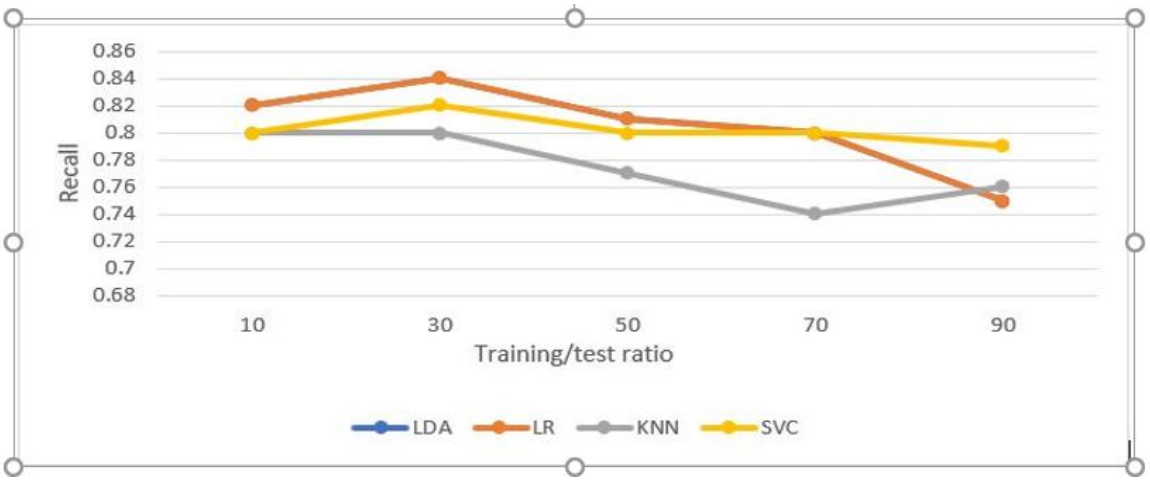

Figure 5.2: Recall for the Classifiers

The result for the recall in figure 5.2 also show similar performances with overlapping characteristics with LDR and LR as well. k-NN exhibits a sharp drop as the variation in training/test ratio is increased but SVM seems to have a steady performance as the Training/test ratio is increased. k-NN seems to perform poorly as the variation increases.

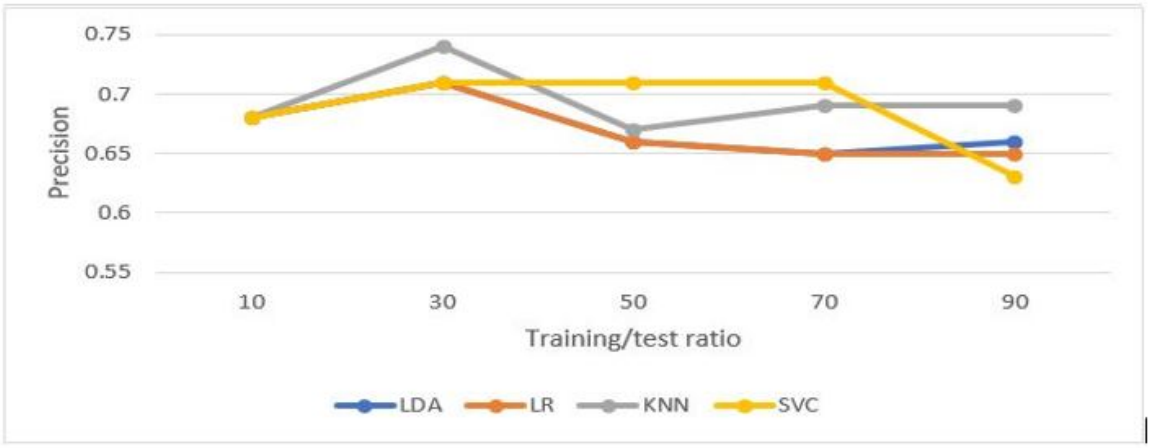

Figure 5.3: Precision for the Classifiers

Figure 5.3 shows the result for the precision. In terms of precision, k-NN has the highest performance which is at 30:70 training/test ratio, whereas SVM seems to have a stable outcome in increasing the training/test ratio, but with a sharp decrease at the ratio of 70:30

5.7.1 Testing with Data from a clinical trial comparing two treatments for a respiratory illness

Further experiment was also carried to examine the interpolation procedures using a respiratory disorder dataset.

The results from the experiments in terms of how the classifiers were able to accurately identify the disease diagnosis is presented. The preceding charts show the comparison with 10-fold cross validation of the training and test data variation:

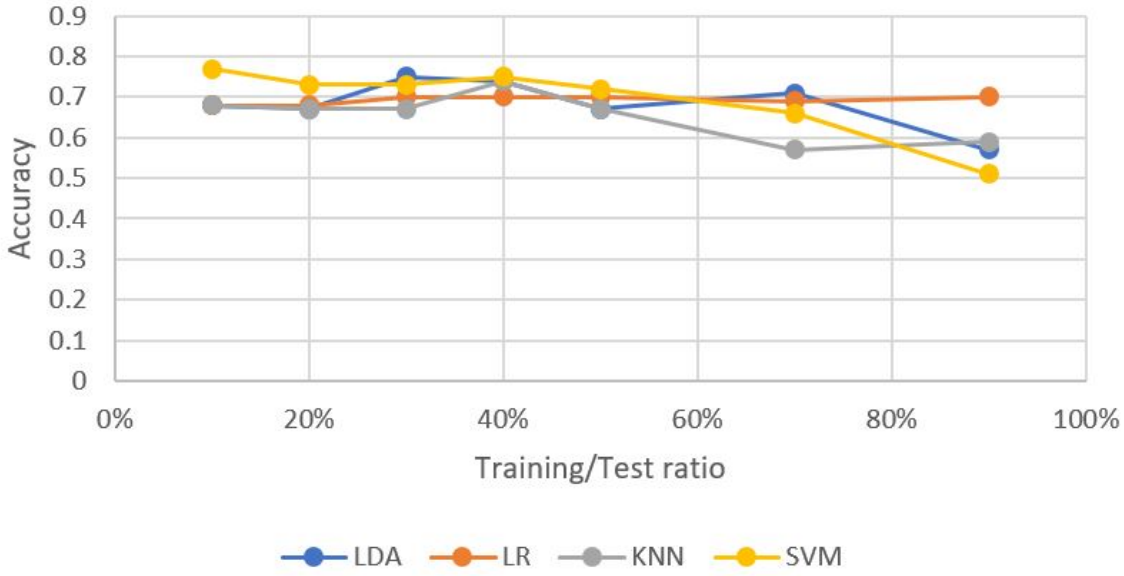

Figure 5.4: Accuracy for the Respiratory diseases Classifiers

5.4 Shows results from the comparison of the classification accuracy for the four classifiers. In overall, the results obtained shows that LR has a better performance in identifying the disease with steady increase as the percentage ratio is increased as compared to LDA, KNN, and SVM, where they show a steady decrease as the percentage ratio was increased. Good performance is attained with higher training and test set ratio for the LR.

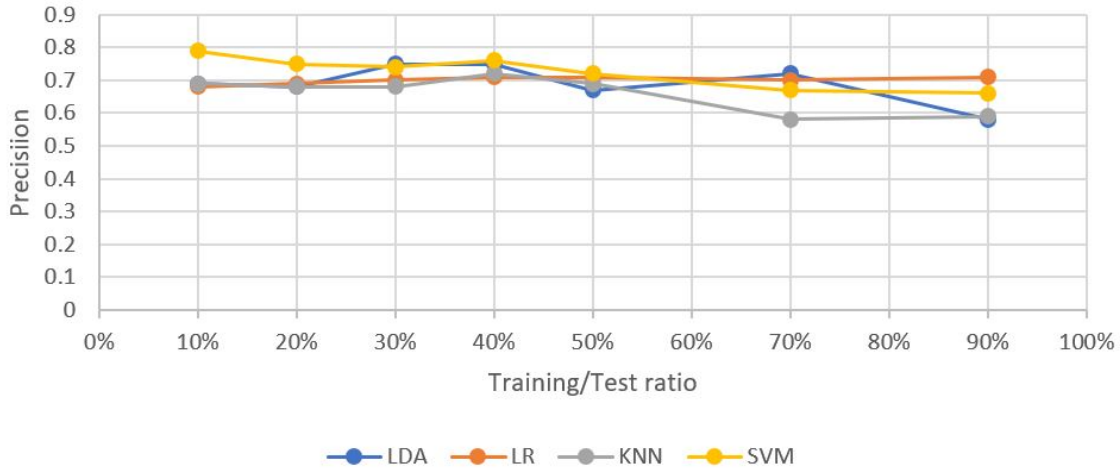

Figure 5.5: Precision for the Respiratory diseases Classifiers

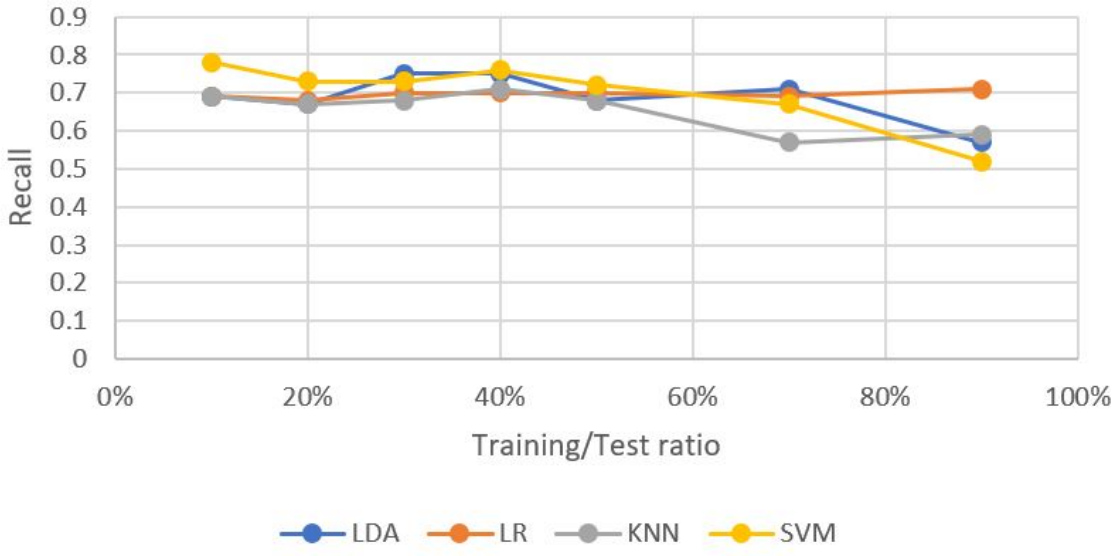

Figure 5.6: Recall for the Respiratory diseases Classifiers

Lastly, figure 5.5 and 5.6 show the Recall and Precision result. It was observed that the linear models offered a better performance over the non-linear models, though all informed models significantly outperformed the SVM.

5.7.2 Conclusions and future work

This chapter presents a brief survey of some popular interpolation methods as well as experiments relating to their accuracy and efficiency which serves as a guide on the suitable choice in building the proposed model. It could be deduced from the results that the variation in the test and training ratio has a great impact on the general performance of the algorithm.

The results from the two experiments show how the classifiers were able to accurately identify and perform a given classification tasks by using comparison with 10-fold cross validation of the training and test data variations.

From the first experiments, the variation of the percentage ratio for the training dataset has significant effect on the behavioural pattern of the various performance metrics used. The result also shows that there is no significant difference between the two linear algorithms (LR and LDA), while the algorithms have their overall best performances at 30:70 data validation/train ratio. The performances also seem to decrease with increase in the validation data. kNN performs better in terms of precision. It can also be seen that LDA and LR show overall better performances.

But when compared with the second experiments on respiratory disorder, the linear model LR has an overall best performances.

These results will guide the choice of selecting the best performing algorithm for the model at hand.

It is also pertinent to point that the drill dataset used was characterised with linearity, incomplete information, fuzziness and uncertainty owing to the legal, corporate and societal impact of having confidential information such as this drill operations data in public domain. This impact negatively on exploring the experiment in greater depth. To prove

the efficiency of our method, synthetic simulated data were used in evaluating their performances.

For future work, the plan would be to model some of the uncertainties to create a knowledge pool of distinct types of stock patterns and apply CBR in computing the similarities and characteristics of the case using controlled experiment and the result tested against real historical stock datasets.

6 The Forward and Inverse Solution

6.1 Case Retrieval

The performance of the CBR is determined by the effectiveness with case retrieval, which in turn depends of the good case representation. Because this work is dealing with unlabelled data, and the case is represented as time points in form of curves. The Euclidean distance measure is adopted from a list of available metrics because of its simplicity and ability to be incorporated with other popular distance measures.

6.1.1 Experimental dataset (Stock dataset)

This dataset as stated in 1.6.3 was drawn from recorded stock market repository; New York Stock Exchange (NYSE) obtained from Yahoo finance. For the purpose of this research, the monthly stock prices of sixteen companies were considered; TOTAL, APPLE, ASPEN, BA, CAT, CSCO, CVX, DIS,FDX, FRD, GSK, IBM,INTC, JPM, NSRGY, TM. The period under consideration is from year 2000 to year 2018. The procedure is to collect the daily index historical data(to cover as many details as possible), and to pre-process them for outlier, missing value, and standardization of the data using Python. Our focus is on the Adjusted price rather than the Closing price. Although they both provide different information that can be used for analysis, the closing price is the raw price and only indicates end of sales price whereas the Adjusted price mirrors stock value after adjustments for any corporate actions like all applicable splits and dividend distributions, which better reflects the assets' perceived value by investors (Ganti, 2020).

The experiment was commenced using the standard correlation coefficient which statistically measures the strength of a linear relationship between paired data (Gogtay and Thatte, 2017; Beaumont, 2012) and given by:

$$\rho(x,y) = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} \tag{6.1}$$

Because our data is preprocessed with (n observations, p features) such that each feature has $\mu = 0$ and $\sigma = 1$, then correlation reduces to cosine:

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \mathbb{E}[XY] = \frac{1}{n}\langle X,Y\rangle \tag{6.2}$$

And under the same conditions, squared Euclidean distance also reduces to cosine:

$$d^2_{\text{Euclid}}(X,Y) = \sum (X_i - Y_i)^2 = \sum X_i^2 + \sum Y_i^2 - 2\sum X_i Y_i = 2n - 2\langle X,Y\rangle = 2n[1 - \text{Corr}(X,Y)] \tag{6.3}$$

Also, rather than applying directly to the Adjusted Close fields, we performed Differencing so as to remove the seasonal component of the series, and the graph shown in figure 6.1

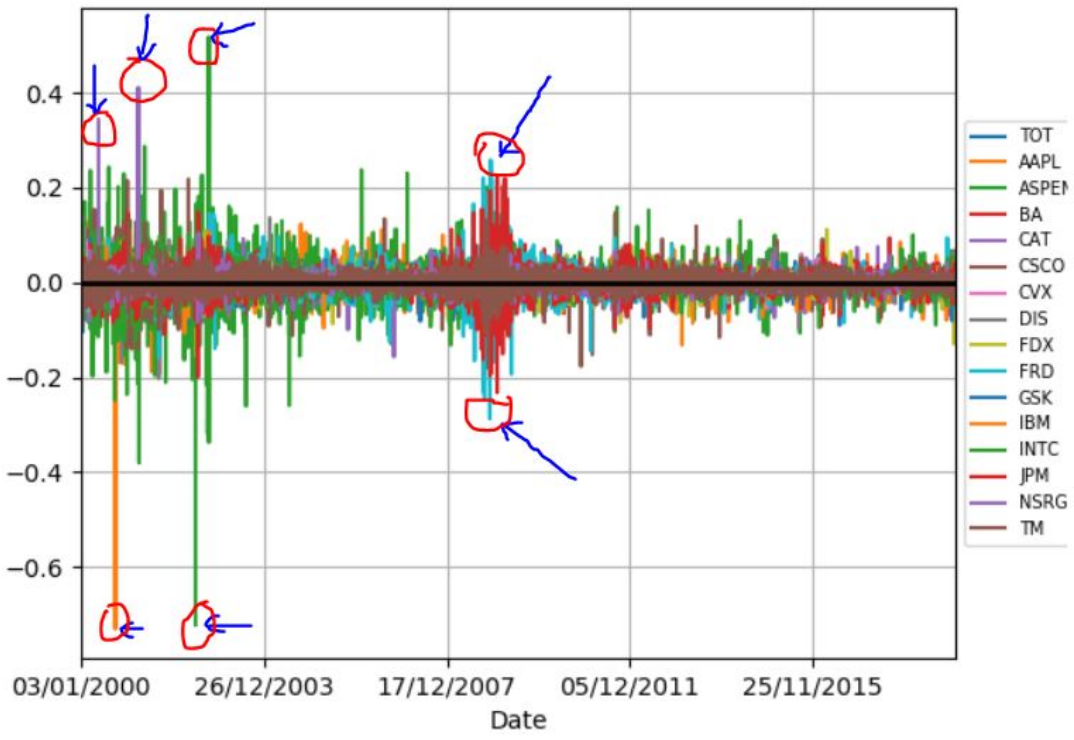

Figure 6.1: correlation Plot

In addition to showing how related the stock movements are, some visible spikes are seen, the presence of the spikes reveal some element of bubbles to be further investigated. For it to be done, a combination of techniques was used to further present the visuals of how the stocks are correlated by first using the sparse inverse covariance estimation which was necessary to determine which stocks have strong or weak correlations with each other (Pedregosa et al., 2011). This method has been successfully applied in Johnson et al. (2011); Friedman et al. (2008).

The sparse inverse covariance depicts a graphical list representation of related stock symbols, with the strength of the edges showing how closely related they are, the bolder the edges, the more strongly correlated the stocks. This sort of features is capable of giving some sort of explanations to any existing fluctuations. Secondly, we applied the Affinity

Propagation technique (Zhang and Gu, 2014; Givoni and Frey, 2009) which concurrently takes all data points as potential exemplars (Refianti et al., 2016) (an instance of the input set that represents clusters), then swapping real-valued information between data points until it eventually arrives at admissible set of patterns and corresponding clusters. The technique does not enforce clusters of equal size as opposed to other popular clustering algorithms like k-means or k-medoids. In order to tackle the issue of high dimensionality, the manifold learning algorithm was used, which is an emerging and promising approach in non-parametric dimension reduction (Cayton, 2005; Qiao et al., 2013), with the help of non-linear version of PCA in reducing the dimensions. The source for generating this was modified from Buitinck et al. (2013b) and the results are visualized in graphical display of figure 6.2

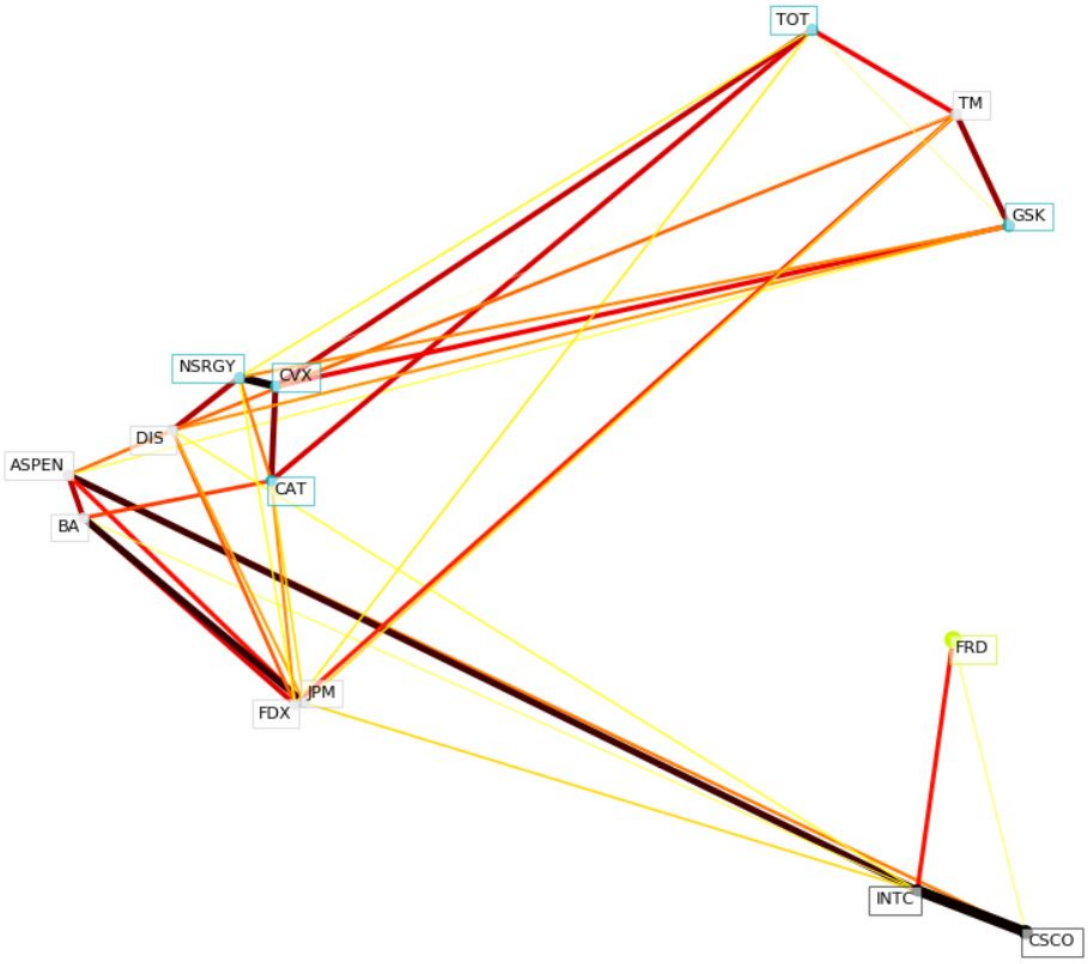

Figure 6.2: Clustering Result

Figure 6.2 represents a graph $G = (V, E)$ consisting sets of objects $V = v_1, v_2, \ldots v_n$ called vertices(stocks), and $E = e_1, e_2, e_3, \ldots e_k$ called edges(relations), such that e_k is identified with an unordered pair (v_i, v_j) of vertices. It is weighted by a function $m : E(G) \mapsto T$, where T represents sets of integers(degree of thickness). As such, the situation where each stock relates in terms of fluctuations level is considered, a strongly positive related vertices are connected with thick lines (CVX and NSRGY), negatively related vertices are joined with very thin lines(CSCO and FRD), and no related have no connecting edges (CAT and FRD). The relevant cluster result is summarised in the table 6.1.1 below.

Clusters	Stocks	
Cluster 1	*CSCO*	
Cluster 2	CAT,CVX CVX NSRGY TOT	
Cluster 3	FRD	
Cluster 4	ASPEN, BA, DIS FDX,JPM,TM	

In order to make clusters from this data that can segment similar fluctuations together, the Agglomerate Hierarchical Clustering technique was applied, as used in Murtagh and Contreras (2012) and then the two results were compared.

To have a better picture of how well the clustering performs, The Cophenetic Correlation Coefficient was used, also applied in (Farris, 1969; Saraçli et al., 2013). The cophenetic distance, simple put is the height of the dendrogram where two branches merge into a single branch. The Cophenetic Correlation Coefficient compares or correlates the actual pairwise distances of all the samples to those implied by the hierarchical clustering (Saraçli et al., 2013). The result of $cc > 0.85$ that is received implies our cluster actually preserves the original distances, which is considered a reasonable cluster fit.

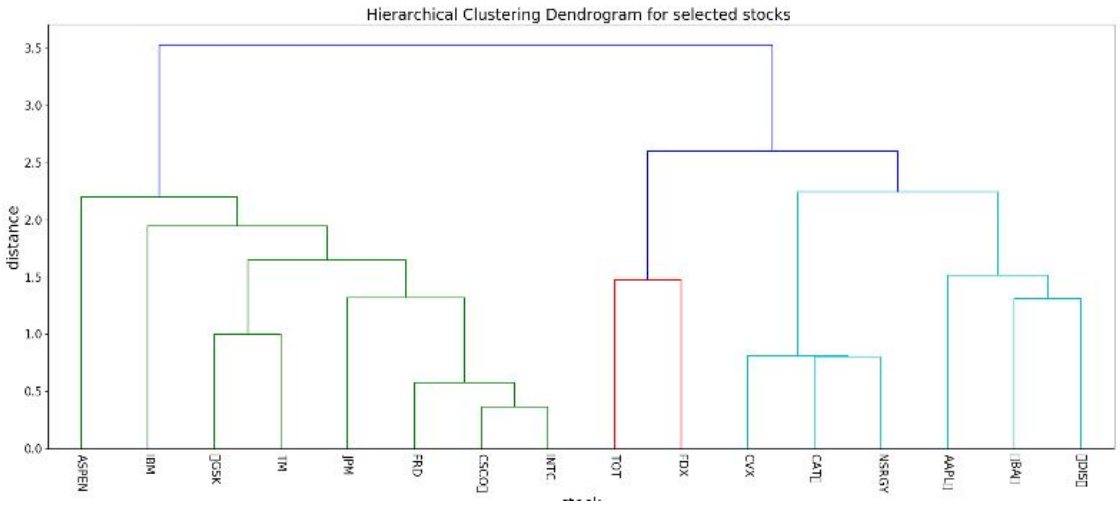

Figure 6.3: Hierarchical Clustering Dendrogram for selected stocks

Figure 6.3 displays a cluster in a tree form, the leaves represents individual stocks based on fluctuating pattern while the root shows the final single cluster. The distance between each cluster is shown on the y-axis, The higher the distance of the vertical lines in the dendrogram, the higher the distance between those clusters, and the less correlated the clusters are. A critical look at the dendrogram simply shows that stocks that are related most in terms of fluctuating patterns are Intel and Cisco, indicating they would be naturally correlated since they are both in the IT sector, while ASPEN and DISC show a very poor correlation .

To decide the actual number of clusters that is suitable, the threshold was set to 6 and then applied it on the scaled data, which eventually identifies 4 distinct clusters as shown in figure 6.4.

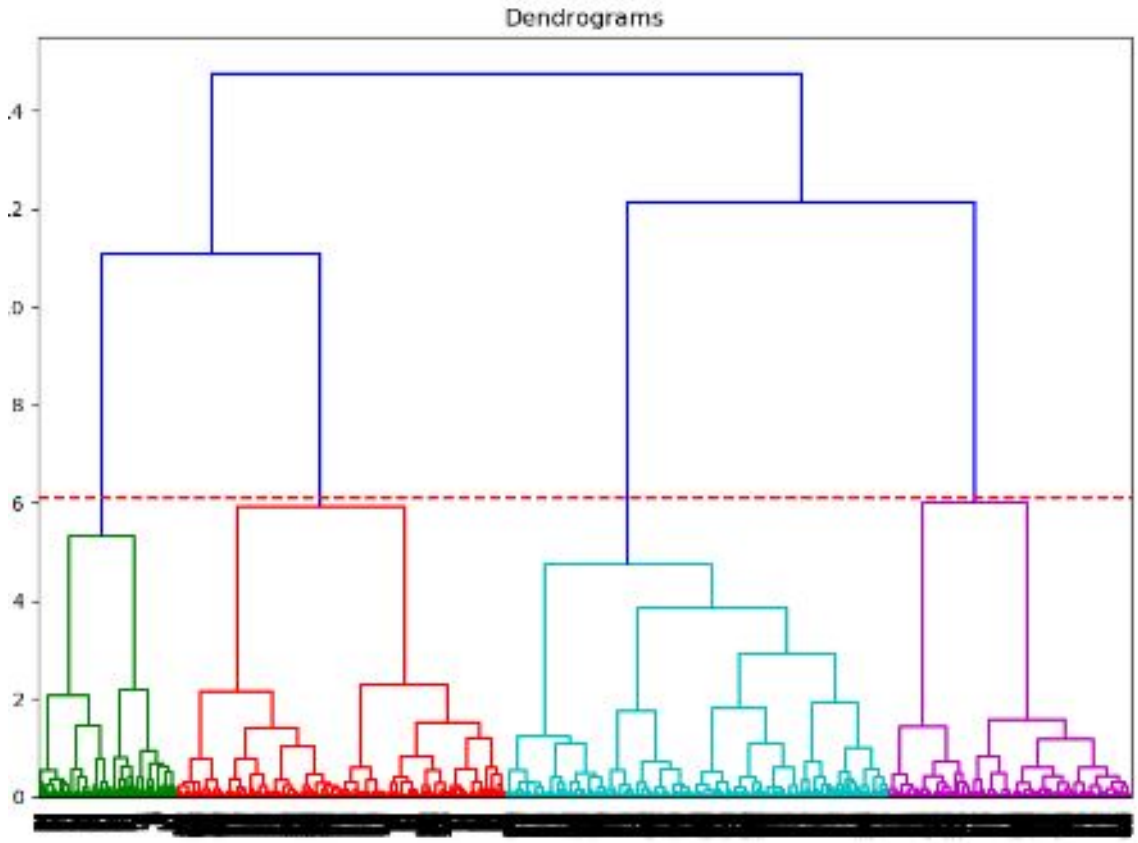

Figure 6.4: Dendrogram Threshold

The result in figure 6.5 finally reveals the actual behaviour of the clusters.

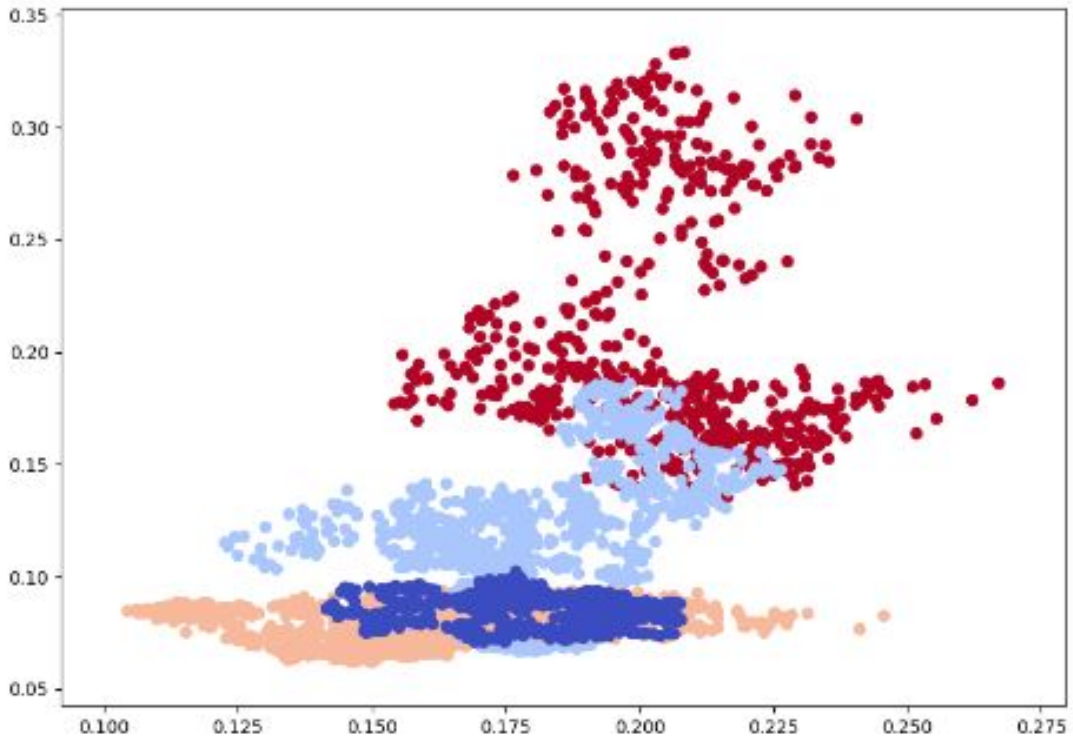

Figure 6.5: Cluster Display

This result in figure 6.5 has shown that fluctuations in stocks of each cluster behave similarly across observed market conditions as expected. Although the overall visuals show some overlaps in the clusters which is typical, considering the high dimensionality, but separate clusters were clearly identified.

6.2 Case Adaptation for Sentiment Analysis

The result is then adapted for use in the Inverse problem Phase. This is achieved by conducting a Sentiment analysis on the sliced portion of data representing the period of bubble.

Sentiment Analysis belongs to a field of Natural Language Processing (NLP) that builds models capable to computationally identifying and categorizing opinions expressed in a piece of body work with the aim of determining whether the writer's attitude towards a

chosen subject, has positive, negative, or neutral tone.

From the Machine learning perspective, Sentiment Analysis is a supervised learning task, where a group of phrases is equipped with labels of their respective sentiments to the machine learning model, and the model is tested on unlabelled phrases.

1. Polarity: if the said attribute implies a positive or negative opinion,
2. Subject: what is being discussed,
3. Opinion holder: the person, or entity that expresses the opinion.

6.2.1 Corpus Collection

For the inverse solution to be accomplished, there is strong need to gather information with which to build a model. To build a decent model, it is mandatory to find data with strong relevance, and for this model to have adequate performance, there is need for a relatively good size of training dataset to train, which is harnessed in the news and tweets that were gathered. As such, there were 3 major channels of dataset used to prove this concept. First was a crawled news headlines from Reddit WorldNews Channel from 2008-06-08 to 2016-07-01, (the period coinciding with the investigated fluctuations). The dataset headlines were ranked by Reddit users' votes, where only the top 25 headlines for a single date were considered. Also, the business times news and tweets were also crawled to obtain data specific to stock and their creation time for each tweet, occurring within the investigating regions. These consist of relevant tweets obtained through Twitter Search API, where the search query consists of stock hash-tag, achieved through the Tweepy (Roesslein, 2019) - an open-source Python library for accessing the Twitter API. Tweepy allows for filtering based on hash-tags or words. Stocks of interest were extracted in a most direct way through the use of hashtag ("#") followed by "stock name", with some others, the abbreviations were used. The selected news sources are financial times, Boomleg, Business times.

The Second dataset consists of data for the purpose of model building, that would serve as training data for the sentiment analyser. This comprises of labelled tweets and news for

classification. The dataset was retrieved from the "Sentiment140", which originated from Stanford University. Detailed information on the dataset is provided at Go et al. (2009).

Thirdly, the generality of the labelled data was observed, bearing in mind that sentiments analysis is a domain specific problem, we also sourced for domain related dataset that consist of labelled news from the finance sector. making a train size of 158,000. Table 6.6, shows a sample of the training data.

0	1467810369	Mon Apr 06 22:19:45	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a
0	1467810672	Mon Apr 06 22:19:49	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it...
0	1467810917	Mon Apr 06 22:19:53	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to
0	1467811184	Mon Apr 06 22:19:57	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. v
0	1467811372	Mon Apr 06 22:20:00	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only
0	1467811795	Mon Apr 06 22:20:05	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09	NO_QUERY	mimismo	@twittera que me muera ?
0	1467812416	Mon Apr 06 22:20:16	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
0	1467812579	Mon Apr 06 22:20:17	NO_QUERY	pardonlauren	I just re-pierced my ears
0	1467812723	Mon Apr 06 22:20:19	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought th
0	1467812771	Mon Apr 06 22:20:19	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never
0	1467812784	Mon Apr 06 22:20:20	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a
0	1467812799	Mon Apr 06 22:20:20	NO_QUERY	HairByJess	@iamjazzyfizzle I wish I got to watch it with you!! I miss
0	1467812964	Mon Apr 06 22:20:22	NO_QUERY	lovesongwriter	Hollis' death scene will hurt me severely to watch on filr
0	1467813137	Mon Apr 06 22:20:25	NO_QUERY	armotley	about to file taxes
0	1467813579	Mon Apr 06 22:20:31	NO_QUERY	starkissed	@LettyA ahh ive always wanted to see rent love the sou
0	1467813782	Mon Apr 06 22:20:34	NO_QUERY	gi_gi_bee	@FakerPattyPattz Oh dear. Were you drinking out of the
0	1467813985	Mon Apr 06 22:20:37	NO_QUERY	quanvu	@alydesigns i was out most of the day so didn't get muc
0	1467813992	Mon Apr 06 22:20:38	NO_QUERY	swinspeedx	one of my friend called me, and asked to meet with her
0	1467814119	Mon Apr 06 22:20:40	NO_QUERY	cooliodoc	@angry_barista I baked you a cake but I ated it
0	1467814180	Mon Apr 06 22:20:40	NO_QUERY	viJILLante	this week is not going as i had hoped
0	1467814192	Mon Apr 06 22:20:41	NO_QUERY	Ljelli3166	blagh class at 8 tomorrow
0	1467814438	Mon Apr 06 22:20:44	NO_QUERY	ChicagoCubbie	I hate when I have to call and wake people up
0	1467814783	Mon Apr 06 22:20:50	NO_QUERY	KatieAngell	Just going to cry myself to sleep after watching Marley a
0	1467814883	Mon Apr 06 22:20:52	NO_QUERY	gagoo	im sad now Miss.Lilly
0	1467815199	Mon Apr 06 22:20:56	NO_QUERY	abel209	ooooh.... LOL that leslie.... and ok I won't do it again so

Figure 6.6: Raw Sample

6.2.2 Tweets Preprocessing and Cleaning

Because raw data are normally populated with so many slang words and punctuation marks, there was need to get rid of them before they can be used for training the machine learning model. Preprocessing is an essential step as it makes this raw text ready for mining. The following steps were followed to carry out the preprocessing.

Tokenization

This step involves splitting the text by spaces, forming a list of individual words per text. This is also called a bag of words. Each word in the tweet were later used as features to train the classifier.

Removing Stopwords

The Python's Natural Language Toolkit library stop word dictionary was deployed to pull out stopwords from each text. If a word is a stop word, then it will be filtered out. The list of stopwords contains articles, some prepositions, and other words that was felt would not add any value to our analysis.

Text Symbols

Many texts that contain extra symbols such special characters as well as URLs were filtered out entirely, as they add no sentiment meaning to the text.

6.2.3 Imbalanced Data

Class imbalance problem is one of the most fundamental challenges faced by the machine learning community (Dattagupta, 2018). The count is 54439 positives and 51465 negatives as also shown in figure 6.7, there is slightly more positives than negatives, making the dataset slightly imbalanced which introduces biasses to the classifier. The data was balanced by using SMOTE (Zheng et al., 2015; Lawrence O. Hall, 2006). which creates balanced set of 43684 0s and 43684 1s.

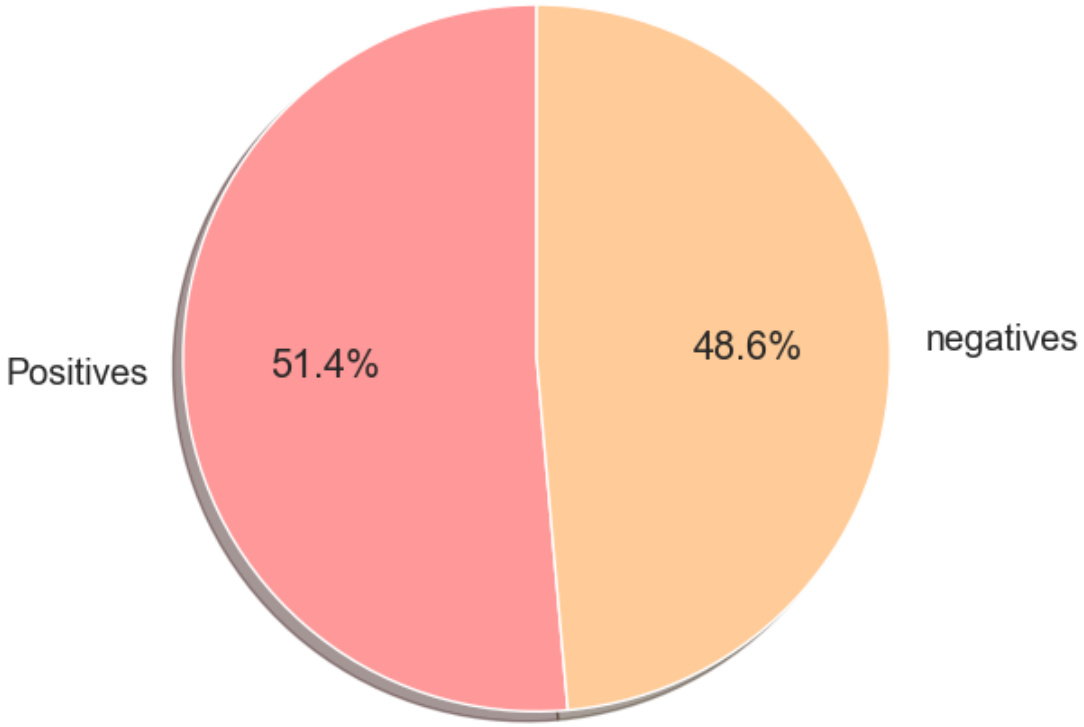

Figure 6.7: Percentage of positives and negatives in training set

At the completion of the stated process, cleaned texts were obtained with balanced set that was saved to a CSV file. To carry out further analysis, there was need to transform the cleaned data into a format that can be processed by the machine learning models through:.

Vectorization

Which carries out text preprocessing, tokenizing and filtering of stop words and it builds a dictionary of features and transform documents to feature vectors.

Features Extraction

Feature extraction is a necessary step for us to make good analysis of a preprocessed data. Based upon the applicability, text features can be constructed using varieties techniques,

some of which are: Bag-of-Words, TF-IDF, and Word Embeddings. To build feature set, each texts was processed to extract meaningful feature and create feature matrix by the TF-IDF although other methods can be used as well.

TF-IDF Features method is based on the frequency method but it is different to the bag-of-words approach in the sense that it takes into account, not just the occurrence of a word in a single tweet, but in the entire corpus. It works by penalizing the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus but appear in good numbers in few tweet. the related terminology is given by

- **TF** = (Number of times term t appears in a tweet)/(Number of terms in the tweet)
- **IDF =** log(N/n), where, N is the number of tweets and n is the number of tweets a term t has appeared in.
- **TF-ID**F = TF*IDF

6.2.4 Model building

Accurate classification is still an interesting problem in machine learning and data mining. In this case, the aim is to construct a model that is trained on the "positive" and "negative" labeled corpus. From this, the model will be able to label raw news as either "positive" or "negative," based on the selected attributes or features.

Model training

Prior to training, and even vectorizing, there was need to split the data into training and testing sets. This was important so as to have a fresh test set. The test size was set to 0.2, or 20 %. Which implies that X_test and y_test contains 20 % of the data which was reserved for testing. Thereafter, the highly fast and simple Naive Bayes was applies to build the first model.

The Random Forest and logistic regression were further used to build the models, the logistic regression predicts the probability of occurrence of an event by fitting data to a logit function.

6.2.5 Model Metrics and Evaluation

There are many ways in which one can obtain performance metrics for evaluating a classifier and to understand how accurate a model is. One of the most frequently used is known as cross-validation.

What cross-validation does is splitting the training data into a certain number of training folds (with 75 of the training data) and a the same number of testing folds (with 25 of the training data), use the training folds to train the classifier, and test it against the testing folds to obtain performance metrics (see below). The process is repeated multiple times and an average for each of the metrics is calculated.

If the testing sets always kept the same, it might be be overfitting to that testing set, which means one might be adjusting analysis to a given set of data so much that it might fail to analyse a different set, a situation that is prevented by Cross-validation helps prevent that.

6.2.6 Model Performance

The standard metrics used to evaluate the performance of a classifier are the Precision, Recall, and Accuracy and the F1 measure

Precision (P) is defined as the fraction of the retrieved elements that are relevant. In other words, precision measures how many elements are predicted correctly as belonging to a given category out of all of the ones that were predicted (correctly and incorrectly) as belonging to the category. And is represented as:

$$P(c,o) = \frac{\mid c \cap o \mid}{\mid c \mid} = \frac{\text{Number of relevant elements retrieved}}{\text{Number of retrieved elements}} \tag{6.4}$$

where c and o represent the set of the retrieved elements and the set of relevant elements respectively.

Recall (R) is defined as the fraction of the relevant elements that are retrieved, mathematically represented as :

$$R(c,o) = \frac{\mid c \cap o \mid}{\mid o \mid} = \frac{\text{Number of relevant elements retrieved}}{\text{Number of relevant elements}} \tag{6.5}$$

which by implication measures how many elements were predicted correctly as belonging to a given category out of all the elements that should have been predicted as belonging to the category. We also know that the more data we feed our classifiers with, the better recall will be.

A more robust measure is the F-measure which combines both recall and precision. The F-measure is defined as:

$$F_1(c,o) = \frac{2.P(c,o).R(c,o)}{P(c,o)+R(c,o)} \tag{6.6}$$

The F-measure is termed F_1 because the recall and precision both have equal weights. But in a case where one might want to assign different weight to the Recall or the Precision. The equation would be represented as:

$$F_\alpha(c,o) = \frac{(1+\alpha).P(c,o).R(c,o)}{\alpha.P(c,o)+R(c,o)} \tag{6.7}$$

Accuracy measures how many elements were predicted correctly (both as belonging to a category and not belonging to the category) out of all of the elements.

Most frequently, precision and recall are used to measure performance since accuracy alone does not say much about how good or bad a classifier is.

6.2.7 Results

The model was first tested using kNN and Naive Bayes, and run a quick comparison. KNN successfully fit the model with an accuracy of 65.23% and F1 score of 62.22 while the NB fits with an accuracy of 74.83 % a higher F1 score of 74.83.

The results show that the NB does a better job as compared with the kNN although they both are generating lots of false predictions. With this result, further improvements were tried on these metrics by running 10 fold cross validation on the data and then compare the results of the two classifiers, and the results shown in figure 6.8, and figure 6.9.

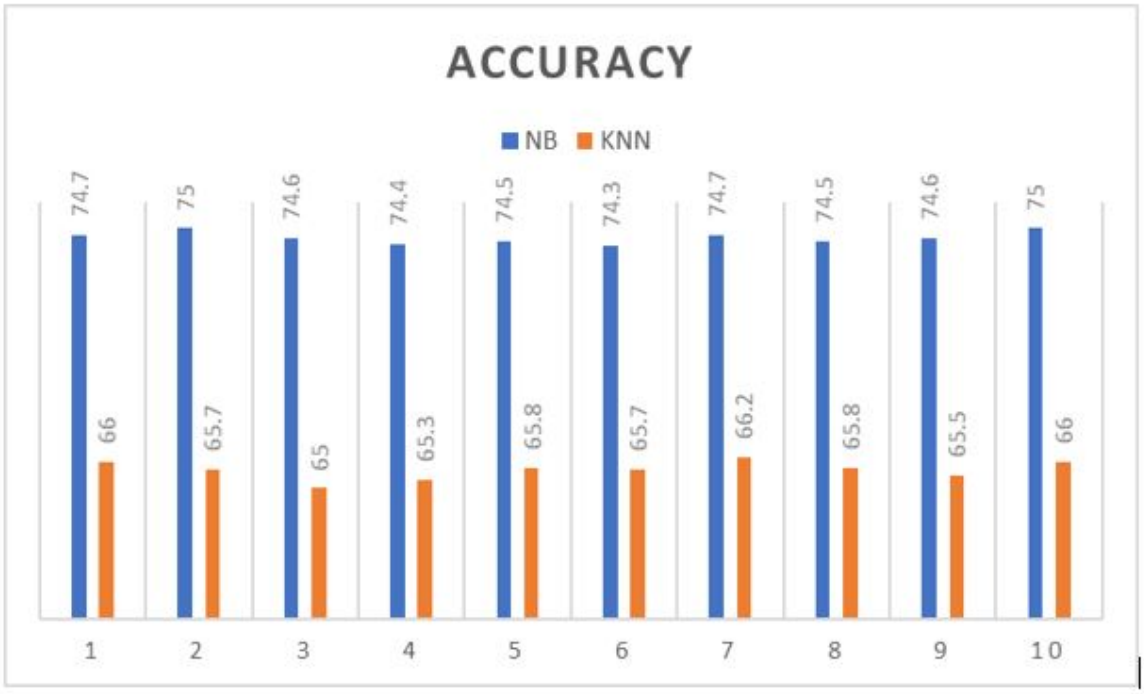

Figure 6.8: Accuracy comparison on 10 fold cross validation

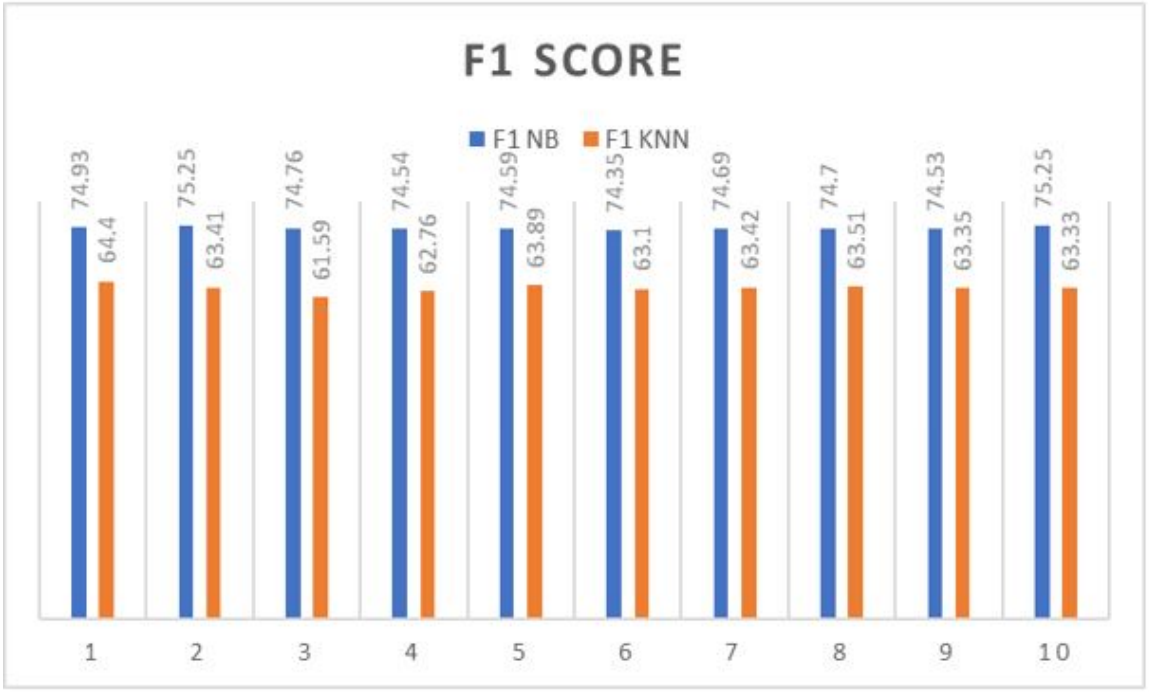

Figure 6.9: F1 score comparison on 10 fold cross validation

The result shows that the NB classifier outperforms the kNN classifier. but there seems to be no significant difference from the results obtained from the single fold. The results were compared using other classifiers to if it would give an improved result. The selected classifiers includes; MultinomialNB Classifier, BernoulliNB Classifier, Logistic Regression Classifier, stochastic gradient descent(SGD) Classifier, LinearSV classifier, and RandomForest Classifier. The obtained result is shown in figure 6.10

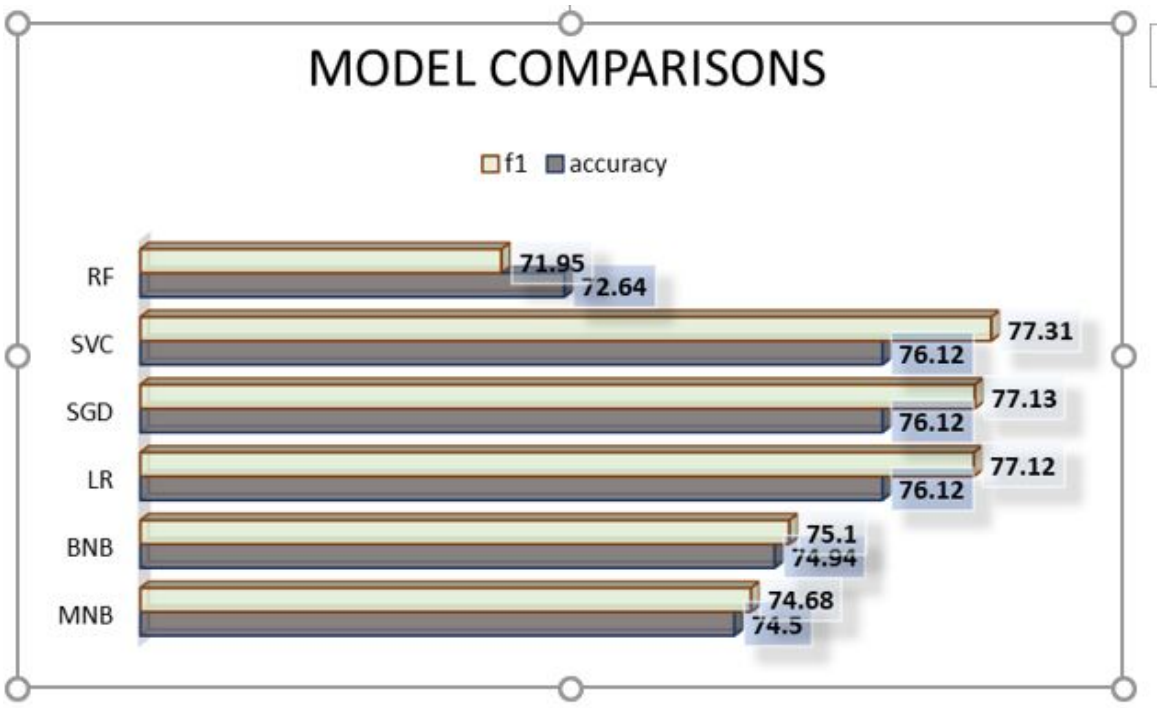

Figure 6.10: models comparison on 10 fold cross validation

From the result above, SVC produced an improved result with an accuracy of 76.23 over that of kNN and NB, also with a higher F1 score of 77.31.

In order to gain more insights on individual performance of the model, some other experiments were conducted to show the behaviour of the labels. Fig 6.11, Fig 6.12 and fig 6.13 show the results of Recall, Precision and F1-score of the classifiers.

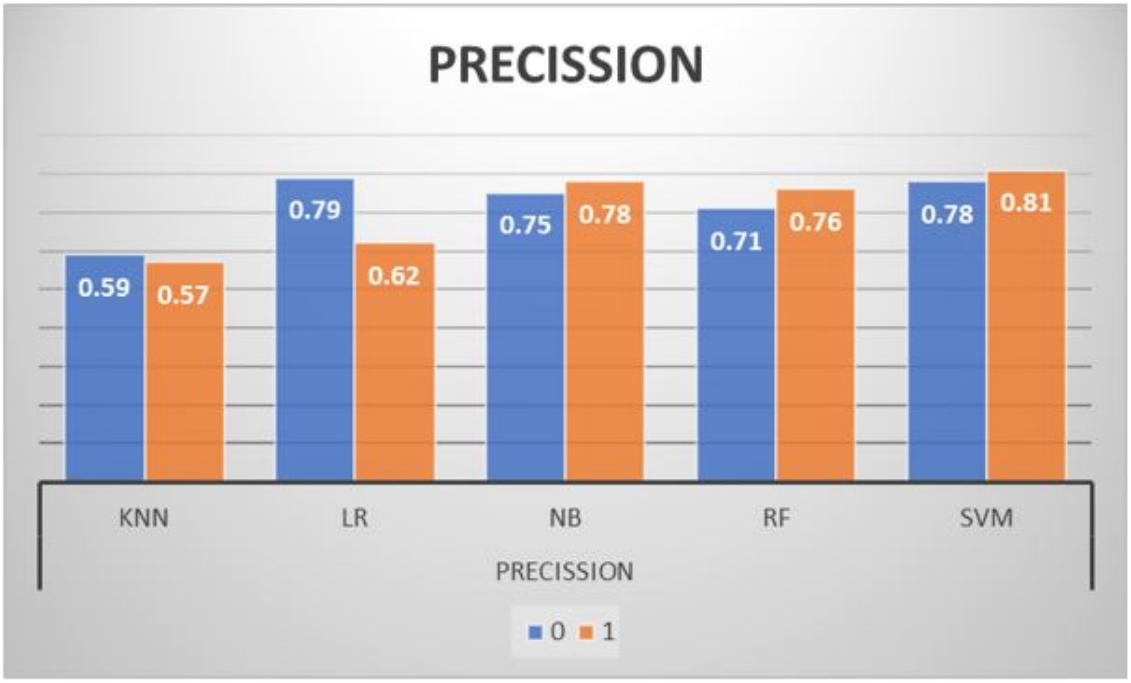

Figure 6.11: Model precision across the classifiers

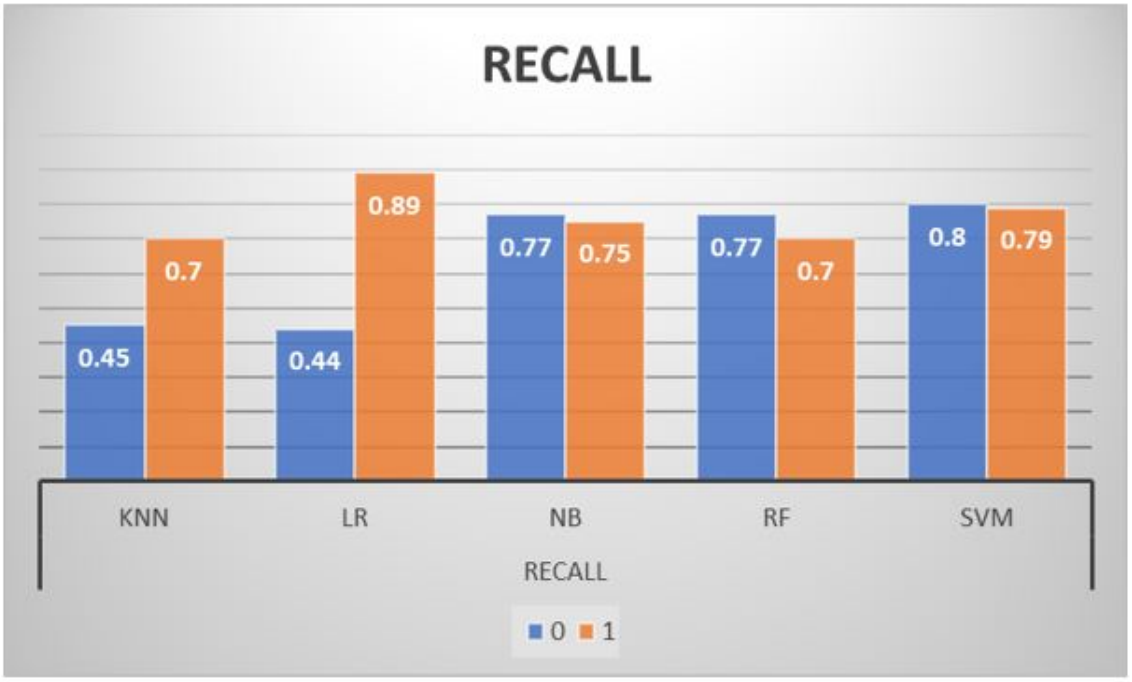

Figure 6.12: Recall across the classifiers

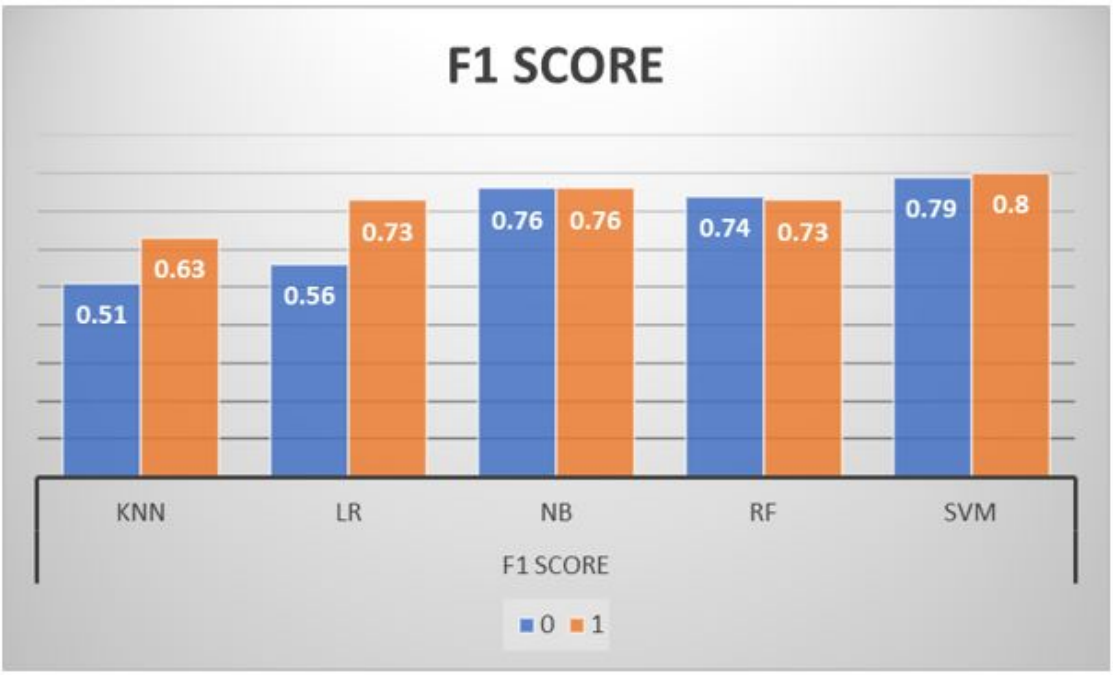

Figure 6.13: F1 score across the classifiers

The overall result from the labels still shows that the SVM scores more in Recall, Precision and in the F_1 score than the other selected classifiers, while the kNN performances were low. The LR gave the highest Recall rate.

Further experiments were conducted to determine how the classifiers perform on the labels using n-grams and up to n=3, and the results presented as follows:

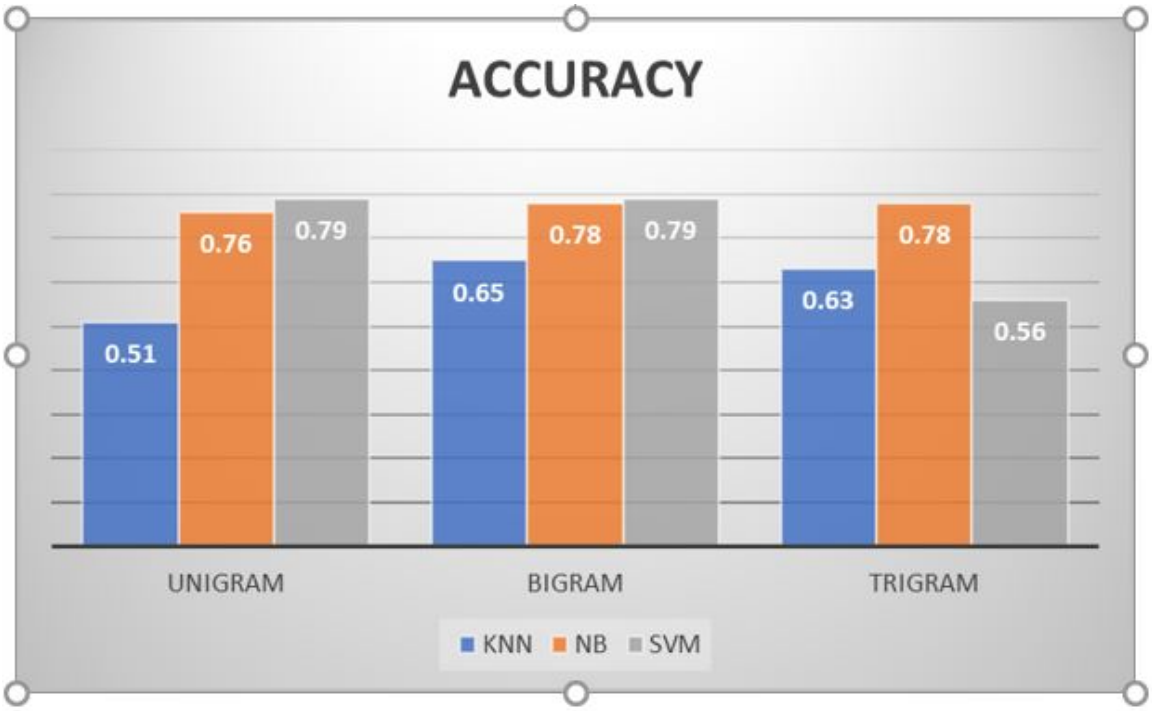

Figure 6.14: n-gram model comparisons

From the result, SVM shows overall superior performance over the kNN and the NB, giving the same high score with uni-gram and Bi-gram although accuracy declined in Tri-gram. The NB seems unaffected in accuracy with all the N-grams. While the kNN performs low in terms of accuracy compared to the BN and the support Vector machine.

6.2.8 Stock movement correlation

To further prove this concept, a short window event around the periods of fluctuations was carried out. The results obtained from both sentiment analysis and stock are combined and analysed. As a result, each correlation compares two time series, the retrieved stock from the forward problem and tweet sentiment by associating result derived from model and each stock index with a corresponding timestamp and then perform a correlation using person correlation given by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{6.8}$$

The sliced the region of the affected fluctuation alongside the time stamps was tracked, their percentage change computed, and then compute the initial correlation based on their percentage change, and the result shown in figure 6.15.

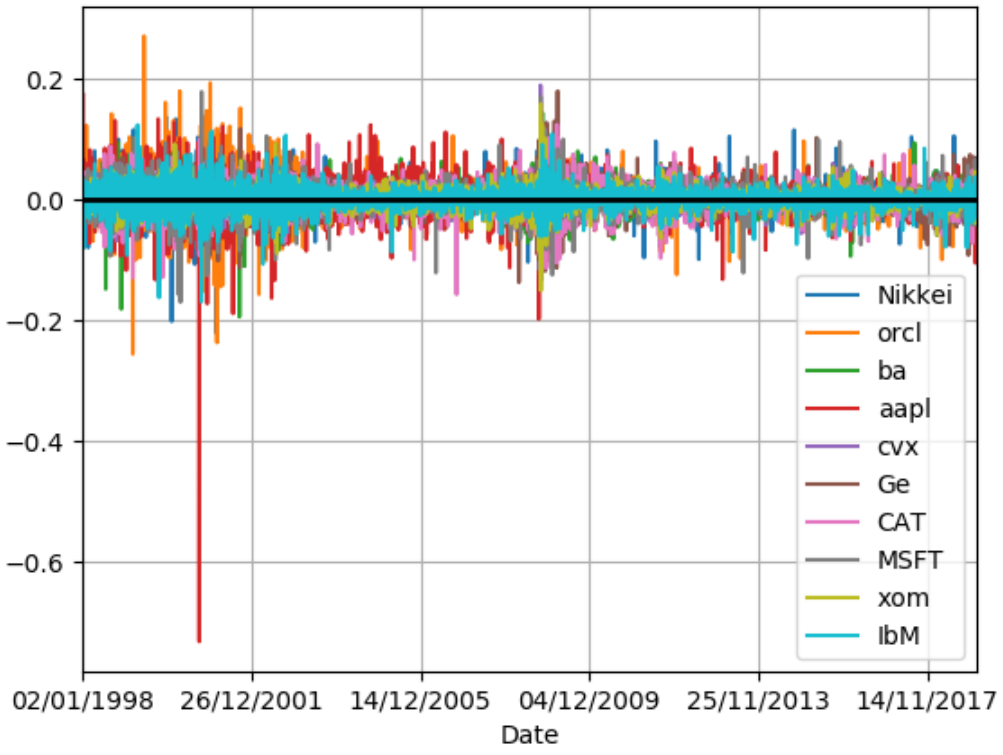

Figure 6.15: Stock/sentiments correlations

After this, a new column that is derived from the model outcome was added, which was also achieved through the use of Panda function which takes the previously built financial stocks prices dataframe as argument and left joins it with our derived sentiment dataframe.

A correlation plot of this real new feature is presented as in figure 6.16:

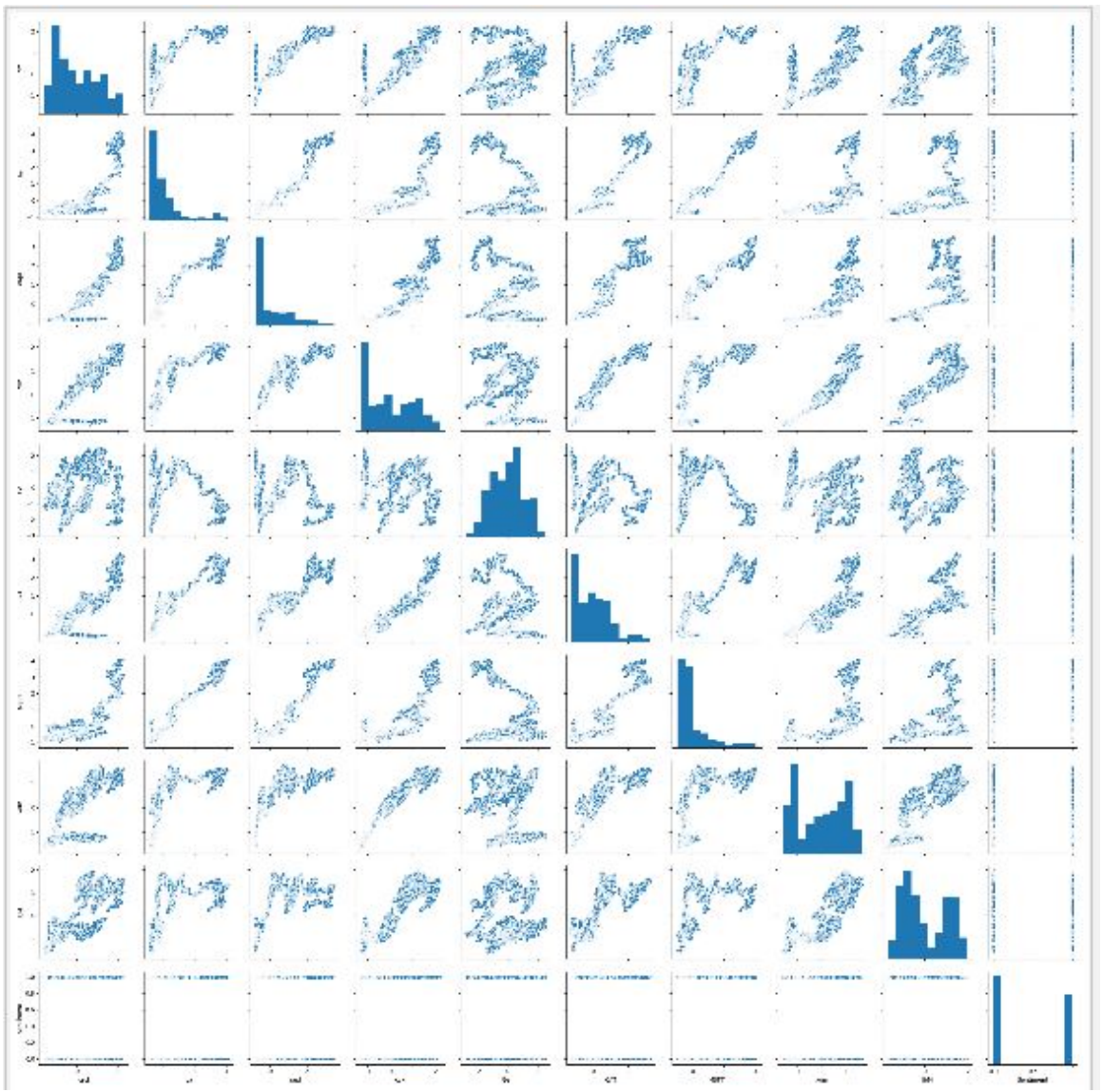

Figure 6.16: Stock/sentiments correlations

The general observation of the graph reveals existence of positive but very week relationship between the sentiments results and the stocks as a whole. What could be deduced at this point of the experiment is that the sentiments outcome do not have any predictive value for stocks.

6.3 Validation and Evaluation

To examine how the IPCBR methodology might work on practical applications, and to measure its performance, a trial was done by first embedding (adapting) the result of the inverse Solution into the CBR framework. Then perform some further preprocessing, followed by the splitting of the dataset into its attributes and labels X and y, X being the columns representing stock (i.e. attributes) while y contains the labels (sentiment score in our case).

The second component being the learning models. Based on the learning models that was worked on previously, a very interesting results was derived. Partitioning was done on 70% of it for training set and the rest 30% for test set. This measure is to avoid over-fitting, and to give a better idea as to how our algorithm performed during the testing phase. The features were also scaled so that all of them can be uniformly evaluated. Training was first trained with the kNN Classifier due to its simplicity (Zhang et al., 2011; Elkan, 2011), precision, recall and f1 score were again used to evaluate the performance of the classifier. A correct classification rate: 0.52. was recorded, while a comparative evaluation of some other classifiers were also done and the result shown in figure 6.17.

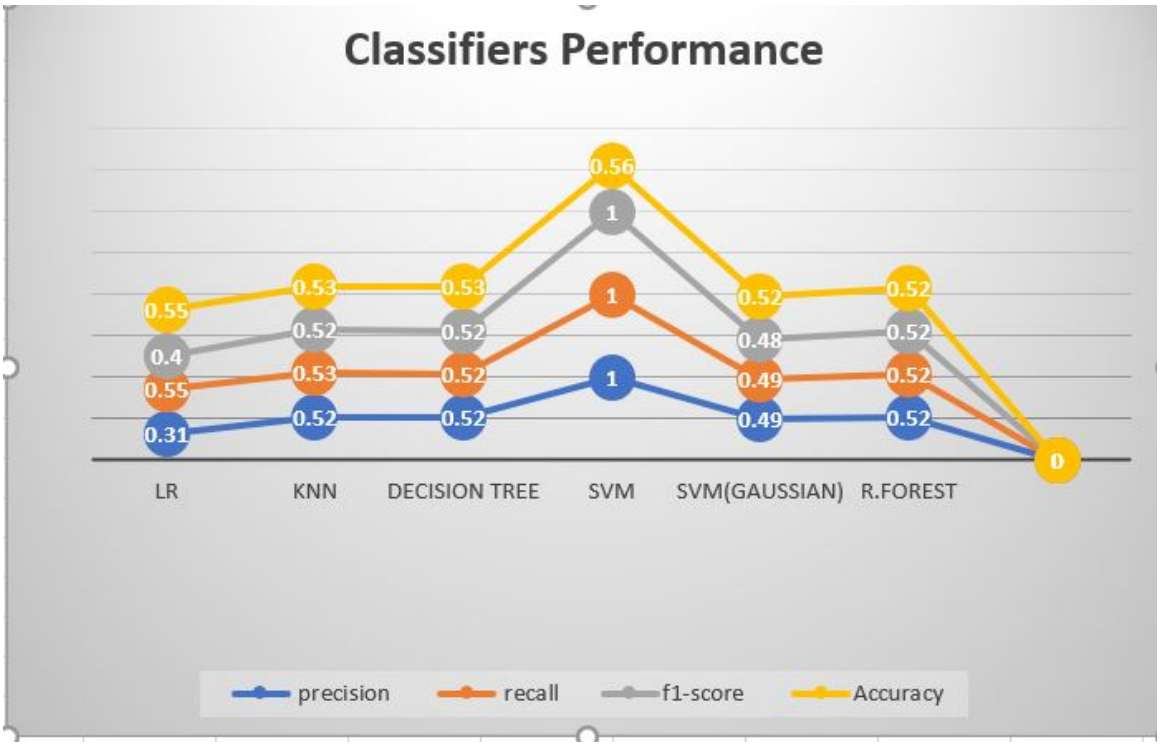

Figure 6.17: Performance of selected classifiers

The result from different classifiers shows that the Support vector Machine gave the best performance, with the value of 0.56 while the least performance is displayed by the linear Regressor. But the fascinating result of rescission, recall and the F1 score 1 values gave room for suspicion of over-fitting in the SVM (which requires further investigation), given that its Gaussian variant gave a lower performance, possibly due to the inherent assumption of feature independence of these algorithms (Kulkarni, 2017). With this, we made do with the kNN classifier.

To verify the above argument and further improve the accuracy and decrease the bias of our data set the performance was improved by randomly K-folding our data set. Using the standard 10-fold cross validation test for analysis with different k values where k is the number of nearest neighbours retrieved.

Since the accuracy of the kNN algorithm deeply depends on the method of distance calculation between time series sequences (Guo et al., 2016). We decided to implement the k-

Nearest Neighbour using Dynamic Time Warping (DtW) as its distance measure as applied in Kulkarni (2017). This is because the DTW has been successfully applied to automatically accommodate time distortions and different speeds associated with time-dependent data as reported in Switonski et al. (2019); Oregi et al. (2017).

In the 10-fold cross validation test, cases in the dataset were randomized and then partitioned into 10 sets. Each set is in turn used as a test set while the other sets are used as training set. This was repeated 10 times since there are 10 partitions. The test results presented here are the average results of the 10 test sets. Also the same partitions were in all tests. The tuning of framework parameters performed based on a grid search. The instances are reserved in time order in order to ensure that they maintain enough independence in time span. The obtained result accuracy is the mean accuracy derived from the 10 experiments.

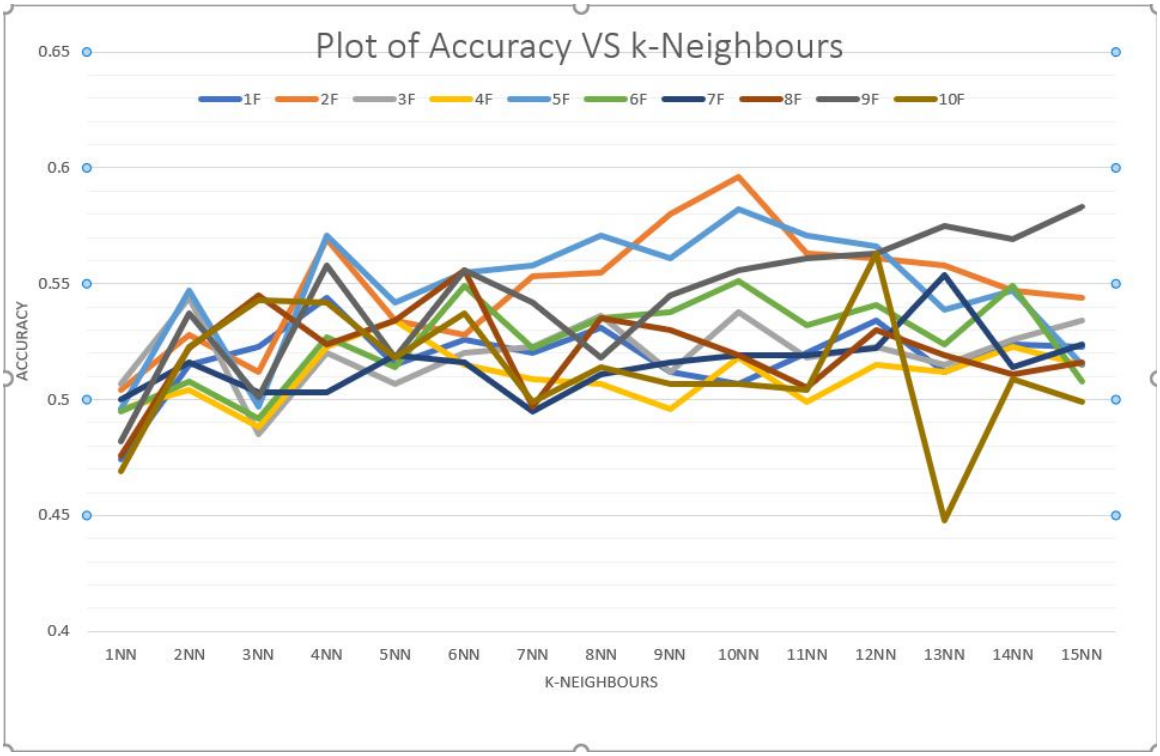

Figure 6.18: Accuracy Vs kNN/DTW of selected classifiers

The experimental result of the 10 fold cross validation on 15 neighbours(1-15), fitting 10

folds for each of 15 candidates, totalling 150 fits is shown in figure 6.18 above. The result shows a peak at the 10-NN of the 2nd fold.

Further experiment was conducted to ascertain the best kNN since there was no way to choose this, we only selected 1 to 15 in the training and prediction section.

The best value of K was determined by plotting the graph of K value and the corresponding error rate for the dataset.The graph of the mean error for the predicted values of test set for all the K values between 1 and 50 is presented below in figure 6.19

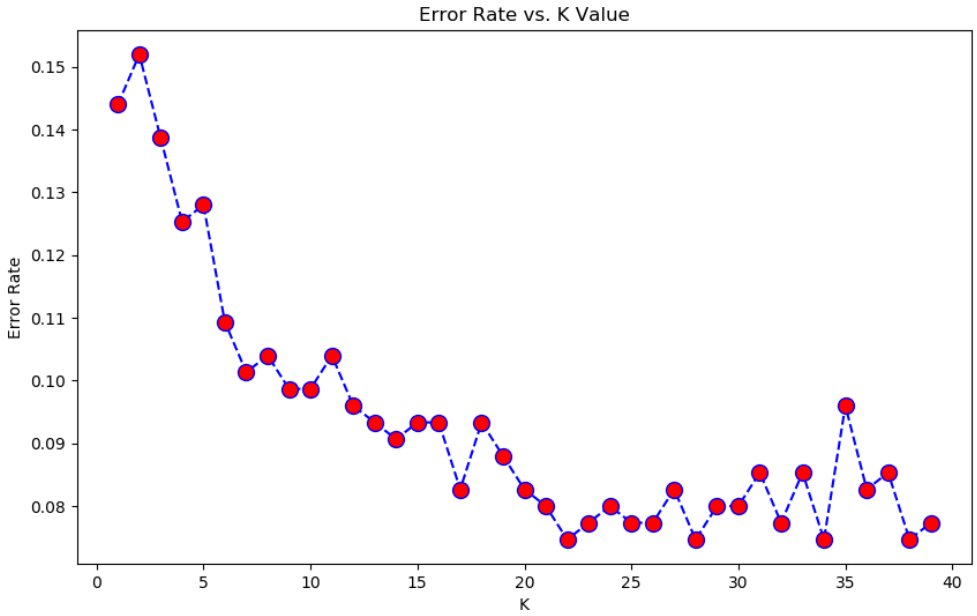

Figure 6.19: mean error Vs K value of k-NN/DTW classifiers

The graph shown above reveals the mean error is lowest when the value of the K is between 22 and 34 neighbours.

This result is important as it gives the optimum number of clusters at the lowest error value. This is because Clustering is an unsupervised approach,the process of identifying the optimum number of clusters is one of the biggest issues affecting Clustering, as such determining the suitable number of clusters for a data set is generally an important procedure.

6.3.1 Summary

Clustering has proven to be a vital tool for data analysis owing to its flexibility, simplicity and robustness. It has served as efficient technique in time series data analysis which is directly applied to financial data. The results have also shown that the existence of numerous clustering algorithms generates several options for evaluating a clustering technique. And that there is no single clustering method that is entirely appropriate, which approach is preferred is specific to the application. Selecting a suitable clustering algorithm and a proper measure for the appraisal highly depends on the clustering objects and the clustering task. As such, acquiring an ample knowledge of both the clustering problem and the clustering technique is essential to apply a proper method to a specific problem.

The results empirically evaluate the ability of Clustering techniques to recognize patterns in multivariate time series of stock data. The results also establish the effectiveness of combining clustering algorithms for modelling stock data. Trained only on raw time series, our models shows strong baseline ability to identify patterns associated with certain fluctuations. Also, the different validity measures presented enables me to obtain the optimal number of clusters.

7 Research Summary and conclusion

7.1 Research Overview

Increasing equivocal concerns in asset value predictability has generated various econometric techniques/models to predict fluctuations in financial asset prices (often associated with asset bubbles). While these models promise improved predictive capability, they are yet to receive wider acceptance in practice.

This is because the stochastic nature of asset and its fuzziness make it difficult to mathematically formulate such problems, resulting in the choice of parameters to be set through heuristics, which could make it difficult to build reliable models. In effects, to give professional explanations to such models become difficult, specifically because it becomes very hard to identify which parameters need to be optimized, and in what way, in order to improve the descriptive power of the model.

The primary research aim of this thesis was to develop a robust and ensemble IPCBR framework to investigate financial bubble which is characterised with fuzziness.

This aim was achieved by developing an ensemble that utilises the simplicity and applicability of CBR to deliver a more robust representation of asset value fluctuation patterns (and their subsequent classification as potential asset bubbles) and the successes of the Inverse Problem to identify the factors that most likely cause such patterns.

Case-based reasoning is a unique problem solving methodology that in many respects is different from the other major AI approaches. This uniqueness in problem-solving technique lies in its ability of utilizing specific knowledge of the previously solved problems

otherwise referred to as cases rather than depending totally on a broad knowledge of the problem domain or making relationships between the problem descriptors and conclusions.

A major research questions were raised which have been successfully solved. These questions and their solutions are illustrated with simplified chart represented in figure 7.1

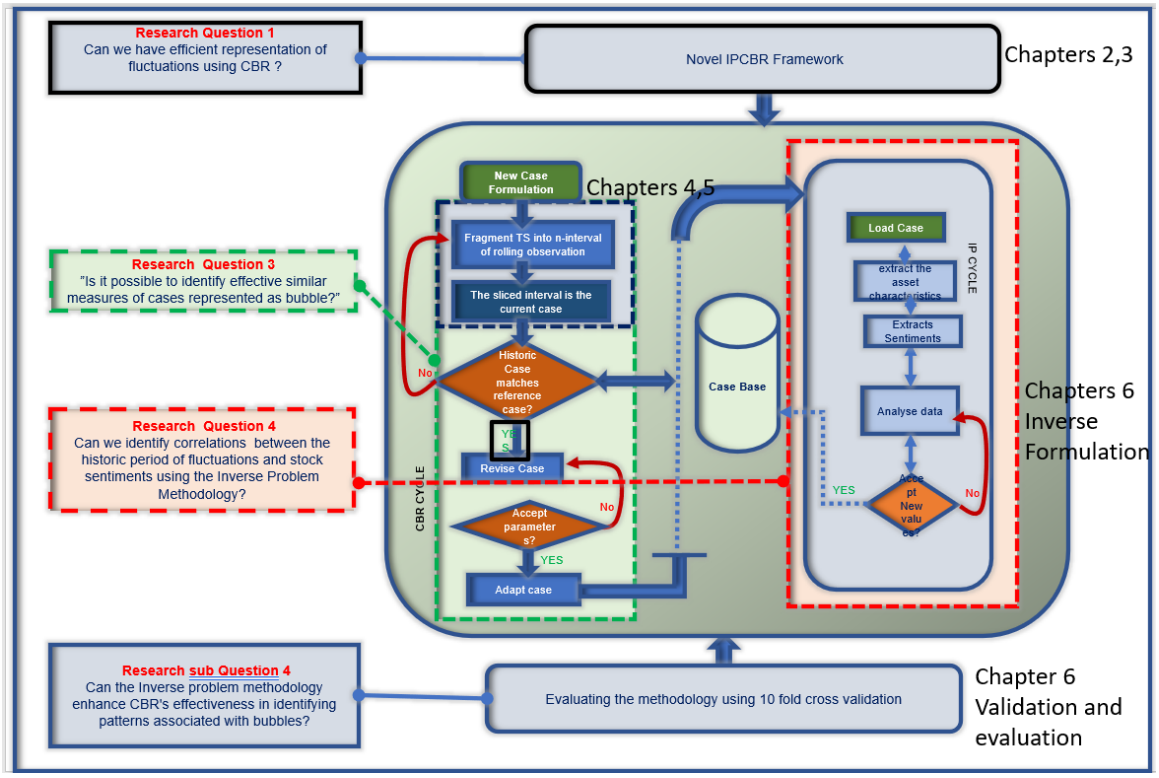

Figure 7.1: Research Problems and Solutions

The thesis has been carried out through a literature study, a survey and an experiment. The literature study conducted during this thesis laid the foundation to derive the research questions outlined in chapter one. With the in-depth study of the existing literature, a framework was developed through the use of Case-based Reasoning and the Inverse problems.

Then came the first research question in trying to get suitable representation of the financial bubble which says:

> Can we have efficient representation of fluctuations using CBR?

This question arose because the forward phase is the Case based reasoning phase. The effectiveness of the CBR lies in its ability to retrieve the most similar cases to the target case. The retrieval effectiveness is based on how well the case is represented. The answer to this salient question is provided in chapters two and three.

This was achieved by forming a library pattern of observations, thereby treating every group as a case category. In effect, the entire Time series was split into smaller sequences of patterns, by decomposing the series into a sequence of rolling observation patterns or rolling windows in which case, every observation in the pattern constitutes the case which may attain a predefined upward, steady or declining patterns as illustrated in table 2.1 and figure 2.3.

Research question two which says:

> Is it possible to identify effective similar measures of cases represented as bubble?

was addressed in chapters four and five where several similarity measures are carefully studied and a simple experiment conducted to ascertain which, among the plethora of similarity measures will be suitable for our model respectively.

Euclidean distance is one of the most commonly used methods for calculating the similarity of time series (Liu et al., 2018), and it has the advantage of being a metric. However, a major demerit of Euclidean distance is that it requires both input sequences be of the same length, and it is sensitive to distortions.

It was discovered in the course of this research that despite the wide us of the Euclidean distance measures, it was not quite suitable for this model because of its inability to handle shifts in time, and may not also process noise and outliers in a desirable way. The Dynamic Time Warping therefore became a suitable measure.

The research question three which states:

> Can we identify correlations between the historic period of fluctuations and stock sentiments using the Inverse Problem Methodology?

was answered in chapter 6 through the sentiment analysis implementation using the stock news/tweets. Also, series of techniques were used to show various degree of correlations on the derived results, it shows how correlated the stocks are in terms of fluctuations within the selected window period that show all the bubble characteristics as well.

Research question four stated:

> Can the Inverse problem methodology enhance CBR's effectiveness in identifying patterns associated with bubbles?

Section 6.3, " Validation and Evaluation" proffers solution to this question by extending Case Based Reasoning methodology framework to deal with cases in which the evaluation of an act may also depend on past performance of different, but similar acts. The implementation of which sees the adaptation of the retrieved cases in the forward solution to finally complete the CBR phase. This enhances the learning ability of the CBR by generating and storing the case in the knowledge container for further use, hence boosting the CBR knowledge base.

In this thesis the use of alliance machine learning methods has demonstrated that the case retrieval accuracy can be achieved using a simple yet efficient approach that is based on assortment algorithms, to steer the course of producing the optimum combinations of classifiers to give accurate classification results.

This thesis has proven that CBR methodology can be successfully applied financial domain, because despite that the CBR has received popularity in many application areas like engineering, medicine, business administration, meteorology, physics amongst others,its full potential is yet to be harnessed in time series domain.

The thesis has also shown that this concept IPCBR is feasible and promising. A suitable architecture for the integration of these two methodologies have been provided and some experimental evaluations given. The IPCBR framework has demonstrated to enhance the Case-based Reasoning both as a methodology and application paradigm.

This research is very significant and timely because the Data Mining techniques are just recently applied in time series analysis, therefore, few scientific efforts are channelled in

this course.

The ensemble approach of combining CBR with the Inverse problems in time series data is a promising area of research. The application of the proposed system in bubble prediction process also has no doubt shown some interesting merits.

The framework enables better and more focused understanding of the behaviour of the bubble with respect to real historical data.

7.2 Research Contributions to knowledge

The research at hand contributes to both the theoretical and practical knowledge on the application of Artificial Intelligence in investigating financial bubbles, which includes:

1. The creation of ensemble IPCBR framework to assist investors in identifying bubbles so as to make well informed decisions.

2. An effective case representation of time series data for a more precise case retrieval in CBR.

3. A combination of Euclidean Distance metrics with Dynamic Time Warping on time series data for a more effective case retrieval strategy using a combination of clustering algorithms.

4. Successful adaptation of the Sentiment analysis results as input to the CBR adaptation phase to enhance CBR's effectiveness in pattern matching.

This research is a novel study because no similar work has been done by any researcher, or any known work done in financial domain using CBR and the Inverse problems according to the literature.

Closely related to this is a work from Woon et al. (2004) and Jenny Freeman (2018).

Woon et al. (2004) showed how the CBR systems can be formulated from numerical models, in order to enhance their usability. The study also presented a novel method for

interpolation over nominal values, called Generalised Shepard Nearest Neighbour method (GSNN), which can utilise distance metrics defined on the solution space of a CBR system.

Jenny Freeman (2018) presented early-warning signalling through optimization-supported tools, which tackles the bubble idea geometrically by establishing and evaluating ellipsoids. The report used ellipsoids to study the bubble patterns by applying the inverse randon transform from the Inverse Problems theory to study the pattern of the episode as bubble time approached.

Although their work and the IPCBR share some concepts, they differ in various ways: While the work of Woon et al. (2004) used CBR and the Inverse problems, both the concept and application is different from this work. It is applied to solve numerical problems while this work investigated financial bubble. On the other hand, Jenny Freeman (2018) investigated bubbles, but with the use of inverse transform form the Inverse problem theory whereas, this work utilises the Sentiments analysis for the inverse solution.

Also, both works from Woon et al. (2004) and Jenny Freeman (2018) used the traditional forward approach of input to output, whereas, this research is applies both the forward and inverse approach.

As such, this thesis establishes a footing framework for future research around trends and future developments in investment behaviour using Artificial Intelligence, hence defining a starting point for models with the capability to study and identify patterns in financial data for future economic developments.

7.3 Limitations of the study

This research has orchestrated a worthwhile course in financial domain and developed a novel IPCBR model framework for financial bubble investigation, but a lot more can still be done to improve on the footings provided by this research to further improve effective response to financial bubble investigations. The following limitations and recommendations are identified, and some useful directions for further improvements are given.

- Lack of prior research studies on the topic – This research is a novel study, at the time of this study, there were very few publications of Case Based Reasoning and the Inverse Problems application in financial domains, and there are also very few literature on CBR and Inverse problem, the closest is the work of (Woon et al., 2004), Using the ensemble in numerical model. This makes it difficult to make direct comparison on the performance of this model.

- The re-integration of Inverse solution into the CBR phase through adaptation introduced another level of complexity during the retrieval process. This operation has a relatively high computational cost. That is because the number of comparison between incident events and the existing cases can be very large. Future research can explore more efficient indexing algorithms to reduce the computational cost involved in the integration process.

- Another challenge is initial knowledge modelling that is required for integration of the CBR methodology and the Inverse Problem technique in creating the ensemble. Most especially with the financial bubble concept that is characterised with fuzziness, The nature of time series that introduces new aspects that is not supported in the processing of the traditional "attribute-value" case representation, this effort can be very high. It would be worthwhile to explore other methods of dimensionality reductions using Symbolic Aggregate approximation (SAX) (Lin et al., 2007) to further improve the result in terms case retrieval.

- The IPCBR framework is a learning model that requires to accrue more cases to guarantee that it can offer appropriate and relevant knowledge in the area of bubble investigations. However, the results have shown that the framework has potential to exploit a small amount of the existing knowledge and deliver reasonably comprehensive evaluation of the financial fluctuations within a short window. Nevertheless, in order to ensure that the IPCBR framework can cover a broad range of financial bubble situations, constant efforts will be required to create and store more cases into

the knowledge base. These efforts may be minimised by integrating the knowledge representation efforts into existing bubble investigation processes.

Moreover, due to the restrictions impacted by the COVID-19 pandemic, precious time that would have been used to get more sample data, run extensive tests on the model framework was lost. Hence the result is evaluated based on the standards metrics, future plans would be to extend this process to domain experts for a more robust evaluation. The limited time factor also inhibited the effort to try out other inverse solution like the Inverse Transforms applied in Jenny Freeman (2018) to further study the behaviour of the fluctuations.

7.4 Future Research Directions

It is apparent from the applicability study described above that there is work to do in trying the architecture on other practical problems since there is a good practical reason to try to solve the inverse problem.

Future direction can be specific to work on Neural Network applications. The reason for not venturing into that during the course of this PhD was due to the framework complexity and the volatile modelling of the financial domain. Further work should be able to provide a formidable benchmark for future reference and research in the area.

As stated earlier, it would be worthwhile to explore more efficient indexing algorithms and other methods of dimensionality reductions to reduce the computational cost involved in the integration process.

Furthermore, this complexity did not give ample time for complete comparison of this framework to wider application areas. As such, future work will be needed to ascertain how this model would perform on a wider variety of examples from the machine learning community. This exercise will be fairly time consuming in view of the necessity to construct alternative metrics on the sets of selected domains. Furthermore, the IPCBR technique would also benefit from a thorough comparative evaluation on a wider set of cases.

References

Aamodt, A. and E. Plaza (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications. IOS Press 7*(1), 39–59.

Abraham, J., D. Higdon, and J. Nelson (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review 1*(3).

Abreu, D. (2003). BUBBLES AND CRASHES By Dilip Abreu and Markus K. Brunnermeier 1. *Econometrica 71*(1), 173–204.

Abutair, H. Y. A. and A. Belghith (2017). Using Case-Based Reasoning for Phishing Detection. *Procedia Computer Science 109*, 281–288.

Acevedo, N. I. A. and N. C. Roberty (2010). An explicit formulation for the inverse transport problem using only external detectors – Part I : Computational Modelling. *Transport 29*, 343–358.

Adler, J. and O. Öktem (2017). Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems 33*(12).

Al-Anaswah, N. and B. Wilfling (2011). Identification of speculative bubbles using state-space models with Markov-switching. *Journal of Banking and Finance 35*(5), 1073–1086.

Al-Jamal, M. F. (2012). Numerical solutions of elliptic inverse problems via the equation error method.

Allanki, S., M. Dixit, P. Thangaraj, and N. K. Sinha (2017). Analysis and modelling of septic shock microarray data using Singular Value Decomposition. *Journal of Biomedical Informatics 70*, 77–84.

Argoul, P. (2012). Overview of Inverse Problems, Parameter Identification in Civil Engineering. pp. 1–13.

Arridge, S., P. Maass, O. Öktem, and C. B. Schönlieb (2019). Solving inverse problems using data-driven models. *Acta Numerica 28*(2019), 1–174.

Asako, Y., Y. Funaki, K. Ueda, and N. Uto (2017). Centre for Applied Macroeconomic Analysis Symmetric Information Bubbles : Experimental Evidence.

Baker, D. The housing bubble and the financial crisis. *real-world economics review,* (46), 73–81.

Bal, G. (2012). Introduction to Inverse Problems.

Bankó, Z., L. Dobos, and J. Abonyi (2011). Dynamic Principal Component Analysis in... 11. *Conservation, Information, Evolution 1*(1), 11–24.

Barido-Sottani, J., S. D. Chapman, E. Kosman, and A. R. Mushegian (2019). Measuring similarity between gene interaction profiles. *BMC Bioinformatics 20*(1), 1–13.

Barlevy, G. (2007). Economic theory and asset bubbles;. pp. 44–59.

Beaumont, R. (2012). An Introduction to Correlation. (September), 1–28.

Berlin, E. and K. Van Laerhoven (2010). An on-line piecewise linear approximation technique for wireless sensor networks. *Proceedings - Conference on Local Computer Networks, LCN*, 905–912.

Berthold, M. R. and F. Höppner (2016). On Clustering Time Series Using Euclidean Distance and Pearson Correlation.

Bharathi, S. and A. Geetha (2017). Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems 10*(3), 146–154.

Bichindaritz, I. MNAOMIA : Improving Case-Based Reasoning for an Application in Psychiatry. pp. 1–2.

Brabham, D. C. (2019). Further Readings. *Crowdsourcing*, 688–690.

Brice, M. (2007). Economic Analysis of Asset Prices. pp. 1–12.

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux (2013a). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux (2013b). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Cao, D., Y. Tian, and D. Bai (2015). Time Series Clustering Method Based on Principal Component Analysis. (Icimm), 888–895.

Caroline Thomas (2003). The South Sea Bubble. *17*, 17–37.

Cassidy, S. (2002). Speech Recognition: Dynamic Time Warping. *Department of Computing, Macquarie University*, 11.2.

Cayton, L. (2005). Algorithms for manifold learning. *Univ of California at San Diego Tech Rep 44*(CS2008-0923), 973–80.

Chakrabarti, K. (2002). Chakrabarti et al. - 2002 - Locally adaptive dimensionality reduction for indexing large time series databases - ACM Transactions on Dat.pdf. *27*(2), 188–228.

Chang, V., R. Newman, R. J. Walters, and G. B. Wills (2016a). Review of economic bubbles. *International Journal of Information Management 36*(4), 497–506.

Chang, V., R. Newman, R. J. Walters, and G. B. Wills (2016b). Review of economic bubbles. *International Journal of Information Management 36*(4), 497–506.

Chaplot, D. S., E. Rhim, and J. Kim (2016). Personalized Adaptive Learning using Neural Networks. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*, 165–168.

Chen, G., Y. Chen, and T. Fushimi (2017). Application of Deep Learning to Algorithmic Trading.

Chen, S. H., P. P. Wang, and T. W. Kuo (2007). Computational intelligence in economics and finance: Volume II. *Computational Intelligence in Economics and Finance: Volume II* (January), 1–227.

Choi, H. R. and T. Kim (2018). Modified Dynamic Time Warping Based on Direction Similarity for Fast Gesture Recognition. *Mathematical Problems in Engineering 2018*.

Contessi, S. and U. Kerdnunvong (2015). Asset Bubbles: Detecting and Measuring Them Are Not Easy Tasks. *The Regional Economist* (July), 5–9.

Cooper, A. (2008). *Review of the book Tulipmania: Money, Honor, and Knowledge in the Dutch Golden Age*. Renaissance Quarterly, vol. 61 no. 1, 2008, p. 220-222. Project MUSE muse.jhu.edu/article/233735.

Cornillon, P., W. Imam, and E. Z. Matzner-l (2008). Forecasting time series using principal component analysis with respect to instrumental variables. *52*, 1269–1280.

Craw, S., N. Wiratunga, and R. C. Rowe (2006). Learning adaptation knowledge to improve case-based reasoning. *Artificial Intelligence 170*(16-17), 1175–1192.

Dale, R. S., J. E. Johnson, and L. Tang (2005). Financial markets can go mad: Evidence of irrational behaviour during the South Sea Bubble. *Economic History Review 58*(2), 233–271.

Dash, T. and T. Nayak (2013). Parallel Algorithm for Longest Common Subsequence in a String. pp. 66–69.

Dattagupta, S. J. (2018). A performance comparison of oversampling methods for data generation in imbalanced learning tasks. pp. 28.

Degutis, A. and L. Novickytė (2014). The Efficient Market Hypothesis: A Critical Review of the Literature. *IUP Journal of Financial Risk Management 93*(2), 7–23.

Devika, M. D., C. Sunitha, and A. Ganesh (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science 87*, 44–49.

Dobbin, K. K. and R. M. Simon (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics 4*(1), 31.

Dokmanic, I., R. Parhizkar, J. Ranieri, and M. Vetterli (2015). Euclidean Distance Matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine 32*(6), 12–30.

Duan, W. Q. and H. E. Stanley (2010). Cross-correlation and the predictability of financial return series. *Physica A: Statistical Mechanics and its Applications 390*(2), 290–296.

Dunbar, S. R. and A. Hall (2016). Stochastic Processes and Advanced Mathematical Finance Models of Stock Market Prices Rating Key Concepts. pp. 1–19.

Dvhg, D. V. H., H. Eulg, V. Iru, R. Q. Ixqgdphqwdo, D. Q. G. Whfkqlfdo, D. Vlv, and I. R. U. Ghflvlrq (2016). A Case-Based Reasoning-Decision Tree Hybrid System for Stock Selection. *10*(6), 1181–1187.

Elkan, C. (2011). Nearest Neighbor Classification.

Elshafiey, I. M. (1991). Neural network approach for solving inverse problems.

Erber, G. (2010). The problem of money illusion in economics. *Journal of Applied Economic Sciences 5*(3), 196–216.

Esling, P. and C. Agon (2012). Time-series data mining. *ACM Computing Surveys 45*(1).

Estimation, L. S. (2008). Chapter 7 Least Squares Estimation. (4), 1–13.

Everitt, A. B. S., T. Hothorn, and D. Functions (2017). Package ' HSAUR '.

Fakhrazari, A. and H. Vakilzadian (2017). A survey on time series data mining. *IEEE International Conference on Electro Information Technology*, 476–481.

Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology 18*(3), 279–285.

Finnie, G. and Z. Sun (2002). Similarity and metrics in case-based reasoning. *International Journal of Intelligent Systems 17*(3), 273–287.

Floyd, M. W. and B. Esfandiari (2009). An active approach to automatic case generation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5650 LNAI*, 150–164.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Friedman, M. (2012). Implications Of In Stock. *1989*(2), 137–189.

Fuchs, B., J. Lieber, A. Mille, and A. Napoli (2000). An algorithm for adaptation in case-based reasoning. *Ecai 2000* (January), 45–49.

Ganti, A. (2020, 12). Adjusted closing price. Available at: `https://www.investopedia.com/terms/a/adjusted_closing_price.asp`, accessed 2021-02-20.

Garber, P. M. (2000). *Famous First Bubbles: The Fundamentals of Early Manias*. The MIT Press Cambridge, Massachusetts London, England.

Garber, P. M. (2018). Famous First Bubbles. *Famous First Bubbles 4*(2), 35–54.

Garcia, E. (2015). Cosine Similarity Tutorial. *Information Retrieval iItelligence* (April), 4–10.

Givoni, I. E. and B. J. Frey (2009). Semi-Supervised Affinity Propagation with Instance-Level Constraints. *International Conference on Artificial Intelligence and Statistics. 5*, 161–168.

Go, A., R. Bhayani, and L. Huang (2009). Twitter sentiment classification using distant supervision.

Goel, A. and A. Mittal (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229. *Cs229.Stanford.Edu* (June).

Gogtay, N. J. and U. M. Thatte (2017). Principles of correlation analysis. *Journal of Association of Physicians of India 65*(MARCH), 78–81.

Gomez-ramirez, J. (2003). Inverse Thinking in Economic Theory : A Radical Approach to Economic Thinking 2 . Four problems in classical economic modeling.

Gontis, V., S. Havlin, A. Kononovicius, B. Podobnik, and H. E. Stanley (2016). Stochastic model of financial markets reproducing scaling and memory in volatility return intervals. *Physica A 462*, 1091–1102.

Guo, A. Y. and H. Siegelmann (2004). Time-warped longest common subsequence algorithm for music retrieval. *Proc. ISMIR*, 27–32.

Guo, G.-c., K.-s. Huang, and C.-b. Yang (2016). TIME SERIES CLASSIFICATION BASED ON THE LONGEST COMMON SUBSEQUENCE SIMILARITY AND ENSEMBLE. pp. 5–7.

Gurkaynak, R. S. (2005). Econometric Tests of Asset Price Bubbles : Taking Stock Econometric Tests of Asset Price Bubbles :.

H., S. and M. Elmogy (2015). Case Based Reasoning: Case Representation Methodologies. *International Journal of Advanced Computer Science and Applications 6*(11), 192–208.

Hasna, O. L. and R. Potolea (2016). The Longest Common Subsequence Distance using a Complexity Factor. *1*(Ic3k), 336–343.

He, H., J. Chen, H. Jin, and S. Chen (2006). Stock trend analysis and trading strategy. *Proceedings of the 9th Joint Conference on Information Sciences, JCIS 2006 2006*(January).

He, X., Z. Zhang, J. Yu, N. Ren, and Y. Xie (2010). Trial-and-error approach to the bioluminescence tomography inverse problem. *Proceedings - 2010 3rd International Conference on Biomedical Engineering and Informatics, BMEI 2010 2*(Bmei), 661–665.

Hidalgo, E. G. (2011). Statistical Physics in the Modeling of Financial Markets. *Erasmus Mundus Master in Complex Systems esteban_guevarah@yahoo.es*.

Hochreiter, S. and J. Schmidhuber (1997a). Long Short-Term Memory. *Neural Computation 9*(8), 1735–1780.

Hochreiter, S. and J. Schmidhuber (1997b). Long Short-Term Memory. *Neural Computation 9*(8), 1735–1780.

Hu, C., Q. Wu, H. Li, S. Jian, N. Li, and Z. Lou (2018). Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *article*, 1–16.

Hu, S. (2019). A Review of Case-Based Decision Theory. *American Journal of Industrial and Business Management 09*(01), 82–90.

Huang, H., W. Zhang, G. Deng, and J. Chen (2014). Predicting Stock Trend Using Fourier Transform And Support Vector Regression. *2014 IEEE 17th International Conference on Computational Science and Engineering*, 213–216.

Iglesias, F. and W. Kastner (2013). Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. *Energies 6*(2), 579–597.

Iizuka, H., D. Marocco, H. Ando, and T. Maeda (2013). Experimental study on co-evolution of categorical perception and communication systems in humans. *Psychological Research 77*(1), 53–63.

Ince, H. (2014). Short term stock selection with case-based reasoning technique. *Applied Soft Computing Journal 22*, 205–212.

Iulian, M. M. (2013). Direct Problems and Inverse Problems in Biometric Systems. *III*(5), 1–14.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2000). *An introduction to Statistical Learning*, Volume 7.

Jeewana, C. (2000). Domain Decomposition Based Algorithms for Some Invers E Problems. (August).

Jenny Freeman, T. Y. (2018). Early warning on stock market bubbles via methods of optimization, clustering and inverse problems. *Annals of Operations Research 260*(1-2), 293–320.

Ji, S., M. Park, H. Lee, and Y. Yoon (2010). Similarity measurement method of case-based reasoning for conceptual cost estimation. *Proceedings of the International Conference on Computing in Civil and Building Engineering*.

Jiang, Z. Q., W. X. Zhou, D. Sornette, R. Woodard, K. Bastiaensen, and P. Cauwels (2010). Bubble diagnosis and prediction of the 2005-2007 and 2008-2009 Chinese stock market bubbles. *Journal of Economic Behavior and Organization 74*(3), 149–162.

Jiawei, X. and T. Murata (2019). Stock market trend prediction with sentiment analysis based on LSTM neural network. *Lecture Notes in Engineering and Computer Science 2239*, 475–479.

Jin, K. H., M. T. Mccann, E. Froustey, and M. Unser (2017). Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE TRANSACTIONS ON IMAGE PROCESSING 26*(9), 4509–4522.

Johnson, C. C., A. Jalali, and P. Ravikumar (2011). High-dimensional Sparse Inverse Covariance Estimation using Greedy Methods.

Kapetanakis, S., M. Petridis, J. Ma, and L. Bacon (2010). Providing explanations for the intelligent monitoring of business workflows using case-based reasoning. *CEUR Workshop Proceedings 650*, 25–36.

Kapetanakis, S., G. Samakovitis, and P. V. G. B. D. Gunasekera. The Use of Case-Based Reasoning for the Monitoring of Financial Fraud Transactions The Business Domain : Intelligent Approaches to Financial Fraud Detection. (i).

Katja Taipalus (2012). *Detecting asset price bubbles with time-series methods Detecting asset price bubbles with time-series methods*.

Kaur, M., T. Sharma, and J. February (2015). Evaluation of Inventory Cost in Supply Chain using Case Based Reasoning. *4*(1), 198–204.

Kaushik, A., A. Kaushik, and S. Naithani (2015). A Study on Sentiment Analysis: Methods and Tools. *International Journal of Science and Research (IJSR) 4*(12), 287–292.

Keogh, E. and C. A. Ratanamahatana (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems 7*(3), 358–386.

Khan, R., M. Ahmad, and M. Zakarya (2013). Longest common subsequence based algorithm for measuring similarity between time series: A new approach. *World Applied Sciences Journal 24*(9), 1192–1198.

Khoshrou, A. and E. J. Pauwels (2019). Data-driven pattern identification and outlier detection in time series. *Advances in Intelligent Systems and Computing 858*(July), 471–484.

Kindleberger, C. P., R. Z. Aliber, and J. Wiley (2005). *Manias, Panics, and Crashes*. John Wiley & Sons, Inc., Hoboken, New Jersey Published simultaneously in Canada.

Knight, B., M. Petridis, and F. L. Woon (2010). Case Selection and Interpolation in CBR Retrieval. *10*(1), 31–38.

Ko, W., K. Lee, S. Lee, and S. Lee (2010). Auditory-visual speech recognition error detection using longest common subsequence matching of vowel sequences. *International Conference on Applied Mathematics and Informatics - Proceedings*, 45–50.

Krasnopolsky, V. M. and H. Schiller (2003). Some neural network applications in environmental sciences. Part I: Forward and inverse problems in geophysical remote measurements. *Neural Networks 16*(3-4), 321–334.

Kubicová, I. and L. Komárek (2011). The Classification and Identification. *Finance a úvěr-Czech Journal of Economics and Finance 61, no. 1*(403), 34–48.

Kulakov, A., M. Zwolinski, and J. Reeve (2015). Fault Tolerance in Distributed Neural Computing. pp. 1–9.

Kulkarni, N. (2017). Effect of Dynamic Time Warping using different Distance Measures on Time Series Classification. *International Journal of Computer Applications 179*(6), 34–39.

Kumar, M. and R. Kumar (2015). Classification Rule Discovery for Diabetes Patients Using SVM. *2*(2), 17–24.

Lawrence O. Hall, W. P. K. N. V. C. K. W. B. (2006). snopes.com: Two-Striped Telamonia Spider. *Journal of Artificial Intelligence Research 2009*(Sept. 28), 321–357.

Lee, D., M. Lim, H. Park, Y. Kang, J. S. Park, G. J. Jang, and J. H. Kim (2017). Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus. *China Communications 14*(9), 23–31.

Lee, H. and R. Singh (2012). Symbolic representation and clustering of bio-medical time-series data using non-parametric segmentation and cluster ensemble. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*.

Lee, Suk Jun, Jeong, S. J. (2012). Trading Strategies based on Pattern Recognition in Stock Futures Market using Dynamic Time Warping Algorithm. *Journal of Convergence Information Technology 7*(10), 185–196.

Lei, Y., Y. Peng, and X. Ruan (2001). Applying case-based reasoning to cold forging process planning. *Journal of Materials Processing Technology 112*(1), 12–16.

Leone, V. and O. R. de Medeiros (2015). Signalling the Dotcom bubble: A multiple changes in persistence approach. *Quarterly Review of Economics and Finance 55*, 77–86.

Liao, Z., X. Mao, P. M. Hannam, and T. Zhao (2012). Adaptation methodology of CBR for environmental emergency preparedness system based on an Improved Genetic Algorithm. *Expert Systems with Applications 39*(8), 7029–7040.

Liberti, L., C. Lavor, N. Maculan, and A. Mucherino (2014). Euclidean distance geometry and applications. *SIAM Review 56*(1), 3–69.

Limjaroenrat, V. (2017). Monetary Policy and Housing Bubbles : Some Evidence when House Price is Sticky. (May), 1–47.

Lin, J., E. Keogh, L. Wei, and S. Lonardi (2007). Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Discov. *Data Mining and Knowledge Discovery 15*(2), 107–144.

Lin, J. and Y. Li (2009). Finding structural similarity in time series data using bag-of-patterns representation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5566 LNCS*, 461–477.

Lin, Jessica, Sheri Williamson, Kirk Borne, D. D. (2011). *Chapter 1 Pattern Recognition in Time Series*.

Liu, L., W. Li, and H. Jia (2018). Method of time series similarity measurement based on dynamic time warping. *Computers, Materials and Continua 57*(1), 97–106.

Liu, P., X. Huang, C. Zhu, X. Chen, and X. Qiu (2015). Long Short-Term Memory Neural Networks for Chinese Word Segmentation. (September), 1197–1206.

López, B. (2013). Case-Based Reasoning: A Concise Introduction. *Synthesis Lectures on Artificial Intelligence and Machine Learning 7*(1), 1–103.

Lovett, F. (2006). Rational choice theory and explanation. *Rationality and Society 18*(2), 237–272.

Magoulas, G. D. and M. N. Vrahatis (2006). Adaptive algorithms for neural network supervised learning: a deterministic optimization approach. *International Journal of Bifurcation and Chaos 16*(07), 1929–1950.

Mansar, S. L. and F. Marir (2003). Case-Based Reasoning as a Technique for Knowledge Management in Business Process Redesign. *Knowledge Management 1*(2), 113–124.

Manzoor, J., S. Asif, M. Masud, and M. J. Khan (2012). Automatic Case Generation for Case-Based Reasoning Systems Using Genetic Algorithms. *2012 Third Global Congress on Intelligent Systems*, 311–314.

Marcela, D. and S. Velandia (2006). A CASE-BASED REASONING METHODOLOGY TO FORMULATING POLYURETHANES.

Marketos, G., K. Pediaditakis, Y. Theodoridis, and B. Theodoulidis (1999). Intelligent Stock Market Assistant using Temporal Data Mining. *Citeseer* (May 2014), 1–11.

Marling, C., M. Sqalli, E. Rissland, H. Muñoz-avila, and D. Aha (2002). Case-Based Reasoning Integrations. *AI Magazine 23*(1), 69–86.

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt and J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference*, pp. 51 – 56.

Merelli, E. and M. Luck (2004). Technical Forum Group on Agents in Bioinformatics. *Knowledge Engineering Review 20*(2), 117–125.

Merigó, J. M. and M. Casanovas (2011). A new minkowski distance based on induced aggregation operators. *International Journal of Computational Intelligence Systems 4*(2), 123–133.

Michael V, M. V. K. (2000). The Cost Minimizing Inverse Classification Problem : A Genetic Algorithm Approach. *Science Direct Volume 29,*(3), 283–300.

Miller, S. (1992). The Method of Least Squares and Signal Analysis. pp. 1–7.

Milunovich, G., S. Shi, and D. Tan (2019). Bubble detection and sector trading in real time. *Quantitative Finance 19*(2), 247–263.

Miskolczi, P. (2017). Note on simple and logarithmic return. *Applied Studies in Agribusiness and Commerce 11*(1-2), 127–136.

Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. W. Phillips, A. Sachs, and E. E. Schadt (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics 75*(6), 1094–1105.

Moreira, M. and E. Fiesler (1995). Neural Networks with Adaptive Learning Rate and Momentum Terms. *Technique Report 95 4*, 1–29.

Morris, J. J. and P. Alam (2012). Analysis of the Dot-Com Bubble of the 1990s. *SSRN Electronic Journal* (June).

Murtagh, F. and P. Contreras (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*(1), 86–97.

Naqvi, N. (2019). Manias, Panics and Crashes in Emerging Markets: An Empirical Investigation of the Post-2008 Crisis Period. *New Political Economy 24*(6), 759–779.

Nedelcu, S. (2014). *Mathematical Models for Financial Bubbles*. Ph. D. thesis.

Nisar, T. M. and M. Yeung (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science 4*(2), 101–119.

Norouzi, M., D. J. Fleet, and R. Salakhutdinov (2012). Hamming distance metric learning. *Advances in Neural Information Processing Systems 2*(October), 1061–1069.

Olden, M. (2016). Predicting Stocks with Machine Learning. pp. 12–55.

Oregi, I., A. Pérez, J. Del Ser, and J. A. Lozano (2017). On-Line Dynamic Time Warping for Streaming Time Series. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10535 LNAI*, 591–605.

Oth, H. A. N. S. O. V. (2003). Riding the South Sea Bubble.

Ou, P. and H. Wang (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science 3*(12), 28–42.

Paper, D. (2021). Sentiment Analysis. *TensorFlow 2.x in the Colaboratory Cloud 5*(6), 203–228.

Pecar, B. (2002). Case-based Algorithm for Pattern Recognition and Extrapolation (APRE Method). *SGES/SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Phan, T.-t.-h., E. P. Caillault, A. Lefebvre, A. Bigand, T.-t.-h. Phan, E. P. Caillault, A. Lefebvre, and A. Bigand (2017). Dynamic time warping-based imputation for univariate time series data To cite this version : HAL Id : hal-01609256. *Pattern Recognition Letters, Elsevier, 2017*.

Picasso, A., S. Merello, Y. Ma, L. Oneto, and E. Cambria (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications 135*, 60–70.

Press, P. and T. Profit (2017). Bursting Bubbles: Finance, Crisis and the Efficient Market Hypothesis. *The Profit Doctrine*, 125–146.

Price, S. M., J. Shriwas, and S. Farzana (2014). Using Text Mining and Rule Based Technique for Prediction of. *International Journal of Emerging Technology and Advanced Engineering 4*(1).

Protter, P. (2016). Mathematical models of bubbles. *9502*.

Qi, J., J. Hu, and Y. Peng (2012). Expert Systems with Applications A new adaptation method based on adaptability under k -nearest neighbors for case adaptation in case-based design. *Expert Systems With Applications 39*(7), 6485–6502.

Qian, C., Y. Wang, G. Hu, and L. Guo (2015). A novel method based on data visual

autoencoding for time series similarity matching. *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, 2551–2555.

Qiao, H., P. Zhang, D. Wang, and B. Zhang (2013). An explicit nonlinear mapping for manifold learning. *IEEE Transactions on Cybernetics 43*(1), 51–63.

Ranco, G., D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE 10*(9), 1–21.

Refianti, R., A. Mutiara, and A. Syamsudduha (2016). Performance Evaluation of Affinity Propagation Approaches on Data Clustering. *International Journal of Advanced Computer Science and Applications 7*(3).

Reinhart, C. M. and K. S. Rogoff (2014). This time is different: A panoramic view of eight centuries of financial crises. *Annals of Economics and Finance 15*(2), 215–268.

Ritter, A., F. Hupet, R. Mun, S. Lambot, and M. Vanclooster (2003). Using inverse methods for estimating soil hydraulic properties from ® eld data as an alternative to direct methods. *Agricultural Water Management 59*, 77– 96.

Roelofsen, P. (2018). Business analytics time series clustering.

Roesslein, J. (2019). tweepy Documentation.

Romanycia, H. J. and J. Pelletier (1985). What is a heuristic? *Computer Intelligence* (1), 47–58.

Romli, A., M. P. D. L. Pisa, R. Setchi, and P. Prickett (2015). Eco-Case Based Reasoning (Eco-CBR) for Supporting Sustainable Product Design. *Second International Conference on Sustainable Design and Manufacturing*, 1–12.

Roofe, A. J. A. (2005). ICT and the efficient markets hypothesis. *Encyclopedia of Developing Regional Communities with Information and Communication Technology*, 353–359.

Saha, P. (2002). A heuristic algorithm for computing the max – min inverse fuzzy relation. *30*, 131–147.

Salvador, S. and P. Chan (2018). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis 11*(5), 561–580.

Saraçli, S., N. Doğan, and I. Doğan (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications 2013*, 1–8.

Schanze, T. (2018). Using singular value decomposition for generalized linear autoregression of signals. *4*(1), 375–378.

Schetinin, V. G. (1998). Self-Organizing Multilayered Neural Networks of Optimal Complexity. *Computer* (4).

Search, H., C. Journals, A. Contact, M. Iopscience, and I. P. Address (2008). Inverse problems Problems in in Machine Learning : machine learning : an an application Interpretation application to activity interpretation. *Theory and Practice 012085*.

Segura, D. M., R. Heath, and A. West (2007). Formulating Polyurethanes using Case-Based Reasoning. *Journal of Plastics, Rubber and Composites 36*(6).

Sengupta, S., P. Ojha, H. Wang, and W. Blackburn (2012). Effectiveness of similarity measures in classification of time series data with intrinsic and extrinsic variability. *Proceedings of the 11th IEEE International Conference on Cybernetic Intelligent Systems 2012, CIS 2012*, 166–171.

Seo, Y., D. Sheen, and T. Kim (2007). Block assembly planning in shipbuilding using case-based reasoning. *Expert Systems with Applications 32*(1), 245–253.

Sever, A. (2015). An inverse problem approach to pattern recognition in industry. *Applied Computing and Informatics 11*(1), 1–12.

Sever, A. (2017). A Machine Learning Algorithm Based on Inverse Problems for Software Requirements Selection. *Journal of Advances in Mathematics and Computer Science 23*(2), 1–16.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives 17*(1), 83–104.

Shokouhi, S. V., P. Skalle, and A. Aamodt (2014). An overview of case-based reasoning applications in drilling engineering. *Artificial Intelligence Review 41*(3), 317–329.

Sikarwar, R. and M. Appalaraju (2018). The Impact of Stock Market Performance on Economic Growth in India. *Asian Journal of Research in Banking and Finance 8*(5), 49.

Sitikhu, P., K. Pahi, P. Thapa, and S. Shakya (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019*.

Sornette, D. and P. Cauwels (2014). Financial bubbles : mechanisms and diagnostics. (January), 1–24.

STADNIK, B., J. RAUDELIUNIENE, and V. DAVIDAVICIENE (2016). Fourier Analysis for Stock Price Forecasting: Assumption and Evidence. *Journal of Business Economics and Management 17*(3), 365–380.

Stone, B. J. M. and H. Clinton (2016). The Case for Universal National Service. pp. 17–19.

Su, Y., S. Yang, K. Liu, K. Hua, and Q. Yao (2019). Developing a case-based reasoning model for safety accident pre-control and decision making in the construction industry. *International Journal of Environmental Research and Public Health 16*(9).

Sun, S., Z. Cui, X. Zhang, and W. Tian (2020). A hybrid inverse problem approach to model-based fault diagnosis of a distillation column. *Processes 8*(1).

Sutton, O. (2012). Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction. *Introduction to k Nearest Neighbour Classification*, 1–10.

Switonski, A., H. Josinski, and K. Wojciechowski (2019). Dynamic time warping in classification and selection of motion capture data. *Multidimensional Systems and Signal Processing 30*(3), 1437–1468.

Tang, M., Y. Yu, W. G. Aref, Q. M. Malluhi, and M. Ouzzani (2015). Efficient processing of hamming-distance-based similarity-search queries over MapReduce. *EDBT 2015 - 18th International Conference on Extending Database Technology, Proceedings*, 361–372.

Tarantola, A. (1987). CHAPTER 1: INTRODUCTION 1.1 Inverse Theory: What It Is and What It Does. *Albert Tarantola, Elsevier Scientific Publishing Company 1*, 1–11.

Technologies, E., R. Makhijani, and R. Gupta (2013). I Solated W Ord S Peech R Ecognition S Ystem Using. *6*(3), 352–367.

Tsz, Y., C. L. Wee, and Y. Hong (2010). Analysis of Mouse Periodic Gene Expression Data Based on Singular Value Decomposition and Autoregressive Modeling. *proceedings of the International MultiConference and Computer Scientists I*, 17–20.

Ulrych, T. J., M. D. Sacchi, and A. Woodbury (2001). A Bayes tour of inversion: A tutorial. *Geophysics 66*(1), 55–69.

Van Rossum, G. and F. L. Drake Jr (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wongsai, N., S. Wongsai, and A. R. Huete (2017). Annual seasonality extraction using the cubic spline function and decadal trend in temporal daytime MODIS LST data. *Remote Sensing 9*(12).

Woon, F., B. Knight, M. Petridis, P. Chapelle, and M. Patel (2004, 01). Enhancing the usability of numerical models with case-based reasoning. *Journal of Expert Update 7*, 17–20.

Xihao, S. and Y. Miyanaga (2013). Dynamic time warping for speach recognition with training part to reduce the computation. *ISSCS 2013 - International Symposium on Signals, Circuits and Systems*.

Xing, G., J. Ding, T. Chai, P. Afshar, and H. Wang (2012). Hybrid intelligent parameter estimation based on grey case-based reasoning for laminar cooling process. *Engineering Applications of Artificial Intelligence 25*(2), 418–429.

Yao, Z. and W. F. Eddy (2014). A statistical approach to the inverse problem in magnetoencephalography. *The Annals of Applied Statistics 8*(2), 1119–1144.

Yiu, M. S., J. Yu, and L. Jin (2013). Detecting bubbles in Hong Kong residential property market. *Journal of Asian Economics 28*, 115–124.

Yu, G., G. Sapiro, and S. Mallat (2012). Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 21*(5), 2481–2499.

Zeitschrift, M. A., S. Gesellschaft, P. Link, E. Dienst, and E.-b. Eth (2019). Tulip Mania ? : The Dutch Tulip Bulb Episode (1636-1637) Revisited Tulip Mania ?

Zgurovsky, M. Z. and Y. P. Zaychenko (2017). *The Fundamentals of Computational Intelligence: System Approach*, Volume 652.

Zhang, C., Y. Chen, A. Yin, Z. Qin, X. Zhang, and L. G. Jun (2019). An Improvement of PAA on Trend-based Approximation for Time Series. pp. 1–15.

Zhang, C., Y. Chen, A. Yin, and X. Wang (2019). Anomaly detection in ECG based on trend symbolic aggregate approximation. *16*(December 2018), 2154–2167.

Zhang, K. and X. Gu (2014). An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. *Mathematical Problems in Engineering 2014*, 1–8.

Zhang, X., J. Liu, Y. Du, and T. Lv (2011). A novel clustering method on time series data. *Expert Systems with Applications 38*(9), 11891–11900.

Zheng, Z., Y. Cai, and Y. Li (2015). Oversampling method for imbalanced classification. *Computing and Informatics 34*(5), 1017–1037.

Zhou, Y., W. Lin, P. Wang, X. Pang, and V. Chang (2018). Stock Market Prediction based on Deep Long Short Term Memory Neural Network. (Complexis), 102–108.

Zhu, C. (2011). P Atterns in F Inancial M Arkets : D Ynamic T Ime W Arping. (January).

Zhu, J., H. Jiang, Y. Shi, C. Zhang, X. Chen, and S. Liu (2015). Fast and accurate solution of inverse problem in optical scatterometry using heuristic search and robust correction Fast and accurate solution of inverse problem in optical scatterometry using heuristic search and robust correction. (May).

List of Figures

1.1 Roadmap of the research 9

2.1 Bubble representation 26
2.2 Case based Reasoning Cycle modified from Aamodt and Plaza (1994) . . . 30
2.3 Sample case patterns 33
2.4 Dependency between input and output parameters 37
2.5 Forward Transformation 38
2.6 Inverse Transformation 39
2.7 Inverse Sentiment Pipeline 45

3.1 CBR/IP Model framework 50
3.2 Case Matching in bubble structure 53

5.1 Accuracy for the Classifiers 82
5.2 Recall for the Classifiers 83
5.3 Precision for the Classifiers 83
5.4 Accuracy for the Respiratory diseases Classifiers 84
5.5 Precision for the Respiratory diseases Classifiers 85
5.6 Recall for the Respiratory diseases Classifiers 85

6.1 correlation Plot 90
6.2 Clustering Result 92
6.3 Hierarchical Clustering Dendrogram for selected stocks 94

6.4 Dendrogram Threshold 95
6.5 Cluster Display 96
6.6 Raw Sample 98
6.7 Percentage of positives and negatives in training set 100
6.8 Accuracy comparison on 10 fold cross validation 104
6.9 F1 score comparison on 10 fold cross validation 105
6.10 models comparison on 10 fold cross validation 106
6.11 Model precision across the classifiers 107
6.12 Recall across the classifiers 107
6.13 F1 score across the classifiers 108
6.14 n-gram model comparisons 109
6.15 Stock/sentiments correlations 110
6.16 Stock/sentiments correlations 111
6.17 Performance of selected classifiers 113
6.18 Accuracy Vs kNN/DTW of selected classifiers 114
6.19 mean error Vs K value of k-NN/DTW classifiers 115

7.1 Research Problems and Solutions 118