CHAPTER-3

3. RESEARCH METHODOLOGY

3.1. Introduction

The study undertaken in the thesis adopts the quantitative path for identifying the leading indicators and developing an EWS for predicting the probability of a financial crisis in India. The study adopts three methods namely signal approach, logit regression and ANNs from the literature and attempts to compare their predictive performance in predicting a banking crisis and stock market crisis. The thesis has been divided into different sections namely-study of banking crisis and study of stock market crisis. Each of these sections has a requirement of its own set of methodology along with relevant econometric techniques that have been described in subsequent sections.

This chapter discusses in detail the Research Design, Research Framework, Data and its sources and Tools and Techniques for analysis.

3.2. Research Design

The present research follows the top-down approach to study the prediction of banking crisis and stock market crisis. It also examines the role of sentiment in estimating the probability of an approaching crisis in Indian context.

The research undertaken in this study involves identification of variables for representing macroeconomic and financial conditions, and sentiment, data collection for identified variables and analysis of the collected data for making inferences. Therefore, to implement the top-down approach, quantitative research design is followed throughout the study. Quantitative research design is suitable under this research setting because the problem in hand is quantifiable and involves empirical investigation of the objectives as document through survey of existing literature. Under quantitative research design, this study involves identification of correlational relationships between two or more variables and using models for identification of cause-effect relationships as well as collective strength of multiple variables.

Quantitative and qualitative research designs are extremely different from each other. Quantitative approach is objective in nature as it does not involve any subjectivity of researcher and rests completely on the findings deduced. Assuming the sample representative of the population, this approach establishes the cause and effect relationships among the variables in the controlled environment. This approach is purely statistical procedure dependent. On the other hand, qualitative research involves the subjectivity of the researcher and deals with exploration and developing explanations for social phenomena. The data is usually collected in the participant's setting, mainly through primary surveys, questionnaire, interviews etc. The analysis of collected primary data involves reasonable construction of general themes from the particular instances and objective interpretation of outcomes emerging from data analysis.

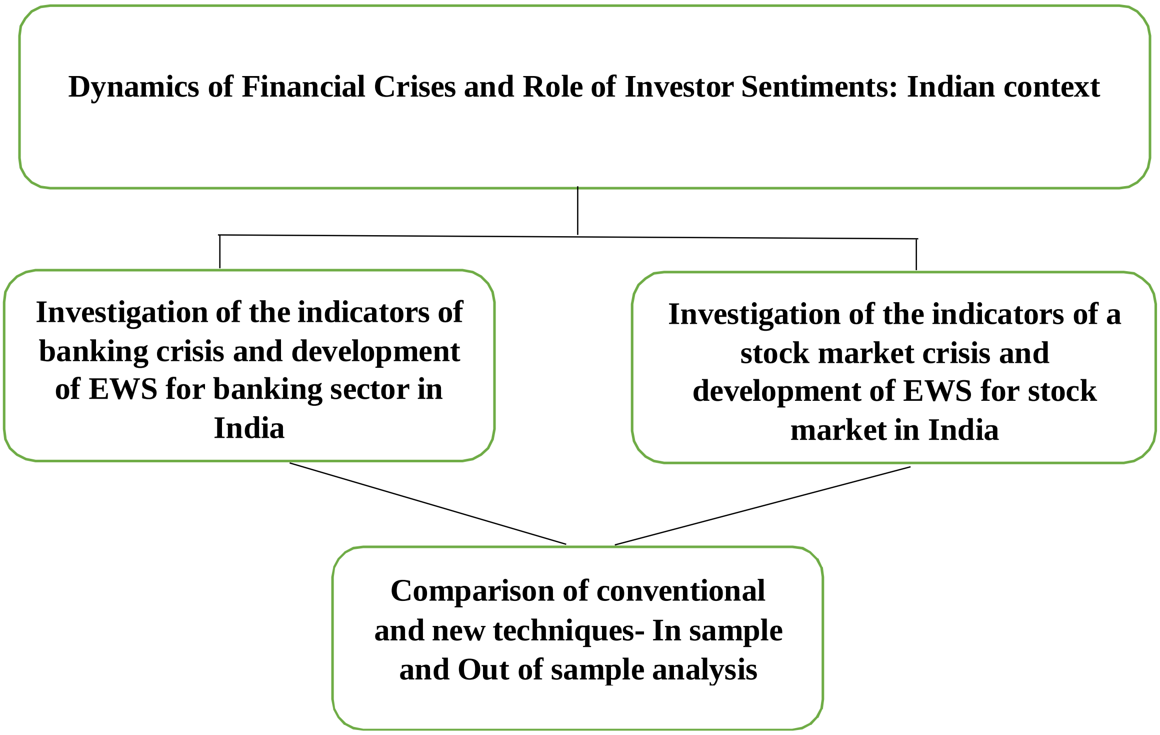
The research design in the present study is quantitative in nature and used confirmatory and exploratory methods to investigate the sources of risk and abnormal behavior of variables leading to a crisis. Exploratory research design is employed to identify the indicators to estimate the probability of an extreme event and develop a model to forewarn and caution about probable happenings of such events, so as to take proactive measures to minimize or in order to avert the huge economic costs associated with such happenings.

3.3. Research Framework

The study is divided into two sections namely- investigation of the indicators of banking crisis and development of EWS for banking sector in India, and investigation of the indicators of a stock market crisis and development of EWS for stock market in India.

For a holistic research, each of these two sections have been further divided into two subsections. The research under each sub-section deals with development of an EWS using conventional techniques and new techniques. The work attempts to identify indicators leading to crisis and check the validity and reliability of new techniques in comparison with traditional approaches as regards to their predictive performance by using in-sample and out-of-sample analysis. The research framework for the study is summarized in Fig 3-1.

Figure 3-1: Research Framework


Dynamics of Financial Crises and Role of Investor Sentiments: Indian context
Investigation of the indicators of banking crisis and development of EWS for banking sector in India
Investigation of the indicators of a stock market crisis and development of EWS for stock market in India
Comparison of conventional and new techniques- In sample and Out of sample analysis

3.4. Research Hypotheses

The following research hypotheses have been formulated for each section of the study:

Section I: Study of banking crisis

It is evident from the literature review, various indicators related to macroeconomic health and financial sector have been found to be affecting the probability of banking crises in various countries. Therefore, the present study has included potential warning indicators have been chosen based on detailed review of literature. These variables have been segregated in four categories namely: macroeconomic environment, external and global variables, credit developments and asset prices. The implications of these variables with respect to their significance or otherwise on likelihood of a banking crisis have been examined in Indian context.

H_{01}: The selected variables have no impact on the likelihood of a banking crisis.

The literature review indicates that a multitude of studies in the domain of constructing EWSs for predicting the likelihood of a banking crisis utilize logit/ probit models and only few of the studies use ANNs for the same. The second hypothesis has been formulated to examine and compare the predictive performance of these techniques.

H_{02}: The predictive performance of Logit models is better than the predictive performance of ANNs.

Section II: Study of stock market crisis

The Indian stock market gets affected by a number of factors like inflation, interest rates, FIIs investments, and other relevant variables. The third hypothesis examines the relevance of selected variables in predicting the probability of a stock market crisis in Indian context.

H$_{03}$: The selected macroeconomic variables have no impact on the likelihood of a stock market crisis.

The investor sentiment is an indicator of the perceptions of the market participants as to price development in a market and a general attitude towards the economy. Thus, the fourth hypothesis examines the role of investor sentiment in predicting the probability of a stock market crisis.

H$_{04}$: The domestic investor sentiment plays no role in predicting a stock market crisis in Indian context.

Given that India is an open emerging economy, and its stock market is integrated with other emerging economies' stock markets, it is imperative to examine the role of sentiment related to emerging countries and other developed economies in predicting the likelihood of a stock market crisis in India. The fifth and the sixth hypotheses examine the role of emerging market sentiment and the developed market sentiment in predicting the probability of a stock market crisis respectively.

H$_{05}$: The emerging market sentiment plays no role in predicting a stock market crisis in Indian context.

H$_{06}$: The developed market sentiment plays no role in predicting a stock market crisis in Indian context.

The studies investigating the prediction of stock market crisis are scant and few studies which have attempted to develop EWS are based on logit/probit models. The seventh hypothesis investigates and compares the predictive performance of logit models and predictive performance of ANN models.

H_{07}: The predictive performance of Logit models is better than the predictive performance of ANNs.

3.5. Data and Sources

The data for the research is collected from secondary sources. The sources used for collecting the data are identified from the study of literature on investor sentiment, financial crises and EWSs. Time period of the collected data is considered based on two factors, viz. number of data points that are suitable for fitting the model and the availability of the data for period of study on various variables considered.

3.5.1. Data for the study of prediction of banking crisis

The prediction of probability of a banking crisis has been investigated using different macroeconomic and financial variables. Firstly, a banking fragility index has been constructed using bank related indicators. This constructed index has been used to identify the episodes of fragility in the Indian banking sector. After the identification, variables based on comprehensive literature review like growth in Nifty 50 index prices as an indicator for stock prices for India, growth in credit - deposit ratio, growth in inflation, growth in industrial production, yield to maturity on 91 days T-bills, spread between bank rate and yield to maturity, weighted average call money rates, growth in current account balance relative to GDP, real exchange rate overvaluation (calculated by differencing the actual Real effective exchange rate (REER) and Hodrick Prescott filter of the REER) , growth in reserve money, growth in broad money relative to foreign exchange reserves, growth in Foreign Direct Investment (FDI) relative to GDP, growth in short term debt, growth in oil prices, and growth in gross fiscal deficit relative to GDP have been employed to check if these indicators act as warning indicators for an approaching banking crisis. The indicators chosen and the reason for selected variables have been discussed in detail in the following sections.

Crisis identification Variable

The sample data is monthly in nature and spans from January 2001 to March 2017. The Banking Sector Fragility (BSF) index is a composite index constituting Aggregate Time Deposits, Foreign Currency Borrowing, Net Bank Reserves and Domestic Credit as proxies for Credit risk, Liquidity risk, and Interest rate risk based on methodology adopted by Kibritçioğlu (2002) and Singh (2011), for which the data has been sourced from Monthly Reserve Bank of India (RBI) bulletin specifically from the Scheduled Commercial Bank survey published weekly by the RBI. This index is used to construct the binary dummy variable for identifying the periodsof fragility in banking sector.

Potential early warning indicators

The banking sector crisis occurs as a result of changes in the macro variables - both domestic and international and financial market indicators. Based on literature review, a set of 15 indicators have been considered in this study affecting the banking sector (Abumustafa, 2006; Ari, 2008; Ciarlone and Trebeschi, 2005; Cumperayot and Kouwenberg, 2013; Davis and Karim, 2008; Fioramanti, 2008; Frankel and Saravelos, 2012; Fuertes and Kalotychou, 2007; Holopainen and Sarlin, 2016; C. S. Lin et al., 2008; Sarlin, 2012). The selected independent variables considered for the study are growth in Nifty 50 index prices as an indicator for stock prices for India, growth in Credit - Deposit ratio, growth in Inflation, growth in Industrial production, Yield to Maturity on 91 days T-bills, spread between Bank rate and Yield to Maturity, weighted Average Call Money rates, growth in Current Account Balance relative to GDP, Real Effective Exchange rate Overvaluation, growth in Reserve Money, growth in Broad Money relative to Foreign Exchange Reserves, growth in Foreign Direct Investment relative to GDP, growth in Short term Debt, growth in Oil prices, and growth in Gross Fiscal Deficit relative to GDP. These 15 variables have been classified in four categories which are discussed below.

Stock Prices: Historically, banking crises have often been preceded by asset price booms. The most recent episode of GFC also observed a housing price boom and bust leading to collapse of the U.S. banking system. Other such examples include incidences of banking crises in number of industrial countries in the late 1970s to early 1990s, such as in Sweden, Spain, Norway, Japan, and Finland (Reinhart and Rogoff, 2008). A decline in asset prices results in loss of investor confidence, which further leads to increased defaults on loans. Hence, growth in Nifty 50 index prices had been used as an indicator of stock prices for India in this study.

Credit expansion: The financial system becomes vulnerable when the private sector indebtedness reaches high values. Especially, when asset price booms are debt- financed which could lead to defaults on account of borrowers when asset prices fall. (Kindleberger and Aliber, 2005; Jordà et. al., 2015). This may result in deleveraging by banks, particularly when the banks, relying on short term funding, face a liquidity mismatch. Deleveraging may induce a credit crunch, which can potentially result in a recession. Fire sales can amplify the effects of the asset price losses, and it may spill over to other assets as these are sold to meet the regulatory requirements such as capital and liquidity ratios. Moreover, bank runs may occur due to the decreased net worth of banks and loss of investor confidence (Allen and Gale, 2007). Thus, to capture the risks related to credit, the growth in Credit - Deposit ratio had been used as an indicator.

Macroeconomic environment: Macroeconomic developments are closely related to stock prices and credit developments. A rapid economic growth boosts risk appetite, credit growth, and asset prices (Drehmann et. al.,2011; Kindleberger and Aliber, 2005; Minsky, 1982) while an economic downturn may lead to difficulties in repayment of debt. To capture the economic environment, variables like growth in Inflation, growth in Index of Industrial production, growth in Reserve Money have been considered. Further, Yield to Maturity on 91 days T-bill (YTM), spread between Bank rate and YTM, and weighted average Call Money rates have been included as investors and banks may take on excessive risk when interest rates are low and low-risk assets are less attractive (Maddaloni and Peydró, 2011; Allen and Gale, 2007; Rajan, 2005).

External and global factors: According to early studies of Frankel and Rose (1996); Kaminsky and Reinhart (1999), the external sector was found to be having an important role in the early warning literature. Even though these studies were focused on the currency crises, it was observedthat both currency and banking crises could occur jointly and often reinforced each other and hence, known as "twin crisis" (Kaminsky and Reinhart, 1999). The external sector was found to contributing to the vulnerabilities in the banking sector when there is a sudden stop or decline in the large capital inflows from abroad. Therefore, variables like growth in current account balancerelative to GDP, real exchange rate overvaluation, growth in broad Money relative to foreign exchange reserves, growth in foreign direct investment relative to GDP, and growth in short termdebt have been included. The growth in oil prices have been included as global oil shocks can affect the domestic banking sector through contagion and various financial and trade linkages.

All the variables considered in the study are monthly in frequency. However, the variables on which availability of data had annual or quarterly frequency were converted to monthly frequency using Cubic Spline Interpolation method. All variables except interest rates and overvaluation of Real Effective Exchange Rate have been transformed into their yearly percentages. Overvaluation from real exchange rate has been calculated by differencing the real effective exchange rate and its Hodrick- Prescott (HP) filtered values. The data on different variables, except Stock prices, Oil prices and the Real Exchange Rate, has been collected from Database on Indian Economy from Reserve Bank of India[6]. Stock prices have been sourced from the official site of NSE Nifty 50, Oil prices have been collected from Federal Reserve Bank of St. Louis and Real Effective Exchange Rates have been collected from IMF. The indicators and their transformations have been shown in Table 3-1.

Table 3-1: Potential Early Warning Indicators and their data transformations

Variable Indicator/Symbol/Frequency/Transformation
Yield to Maturity-91 Days T-bill/ ***YTM****/Monthly/ none*
Spread between Bank Rate and Yield to Maturity/ ***SPREAD****/Monthly/none*
Weighted Average Call Money Rate/ ***CMR****/Monthly/none*
*Current Account Balance relative to GDP/****GCAB_GDP****/Monthly/y_o_y*
Foreign Direct Investment relative to GDP/ ***GFDI_GDP****/Monthly/y_o_y*
*Gross Fiscal Deficit relative to GDP/****GGFD_GDP****/Monthly/y_o_y*
*Index of Industrial Production(base 1993-94=100)/****GIIP****/Monthly/y_o_y*
*Wholesale Price Index(base 1993-94=100)/****GWPI****/Monthly/y_o_y*
*Reserve Money/****GRM****/Monthly/y_o_y*
*Broad Money relative to Foreign Exchange Reserves/****M3_FEX****/Monthly/y_o_y*
Oil Prices/ ***GOILP****/Monthly/y_o_y*
Stock Prices(Nifty 50)/ ***GSP****/ Monthly/y_o_y*
Short Term Debt/ ***GSTD****/ Monthly/y_o_y*
*Credit to Deposit Ratio/****GCDR****/Monthly/y_o_y*
*Overvaluation of Real Exchange Rate/****REER_DEV****/Monthly/y_o_y*

[6] *Handbook of Statistics on the Indian Economy, Time-Series Publications, Database of Indian Economy, RBI. The URL has been given: https://dbie.rbi.org.in/DBIE/dbie.rbi?site=publications*

3.5.2. Data for the study of prediction of stock market crisis

The study employs monthly data from June 2001 to December 2018. The closing prices from National Stock Exchange (NSE) have been considered to represent Indian stock market. For construction of the investor sentiment, the seven implicit market related sentiment proxies have been chosen based on literature review. All the stock market related data are extracted from Handbook of Statistics on Indian Security Market 2012 and 2018 provided by SEBI year books and Prowess IQ by Centre for Monitoring India Economy (CMIE). Monthly data has been extracted from Handbook of Statistics on Indian Economy provided by RBI, International Financial Statistics (IFS) provided by IMF, Federal Reserve Bank of St. Louis and Bruegel database. The following sections discuss the identification of stock market crisis and the choice of financial and macroeconomic variables, construction of behavioral variable and the data sources.

Crisis identification variable

For the identification of the crisis episodes in Indian stock market, monthly stock prices of NSE Nifty 50 have been used spanning from June 2001 to December 2018. The prices have been extracted from the official site of NSE. These monthly prices have been utilized to construct a CMAX index based on which a dummy variable has been deduced identifying the events of extreme stress in Indian stock market.

Potential early warning indicators

The following subsections presents the macroeconomic and financial variables used for identifying the significant indicators to predict a stock market crisis. The variables which have been included in the present study are: Net inflows by Foreign Institutional Investors(FIIs), Weighted Average

Call Money Rates (CMR), Real Interest Rate (RIRR), Credit-Deposit Ratio (CDR), Real Effective Exchange Rate (REER), and Index of Industrial Production (IIP).

Interest Rate: The interest rates have been found to be negatively related to the stock market prices. Burkart and Coudert (2002) confirmed that a relationship existed between real interest rates and stock market prices and suggested that rates declined before a stock market crisis hits. In the present study, the real interest rates have been defined as the lending rates adjusted by the inflation add the source.

Credit: Many studies have suggested that expanding domestic credit plays a major role prior to crisis. As documented in Eichler and Sobanski (2012), faster credit growth increases banking sector fragility as it encourages investors to take excessive risk. As suggested by Goldstein (1998) and Kamin (1999) banking institutions might not be fast enough to enable their institutions in screening and monitoring the new loans appropriately. This could make the economy vulnerable to financial crisis due to deterioration in bank balance sheets.

Call Money Rate: Call money rate is the rate at which short term funds are borrowed and lent in the money market. A positive relationship is expected between the call money rates and probability of a stock market crisis. As evidenced by 2008-09 crisis, call money rates rose to 20% or so due to heavy pressure on domestic banks resulting in a liquidity crisis.

Foreign Institutional Investors: FIIs are one of the most dominant investors group that have emerged to play a critical role in the overall performance of the stock market. India opened its stock markets to foreign investors in September 1992 and now it has become one of the main channels of international portfolio investment in India for foreign investors. A positive relationship is expected between the investment and stock prices. Therefore, a correction in stock market is expected to be led by FIIs pulling out the money. This in turn would increase the probability of a crisis in stock market as happened during subprime crisis of 2008-09.

Exchange Rate: The relationship between the exchange rate and stock market lacks consensus. Two theories concerning the relationship between stock prices and exchange rates are traditional approach and portfolio approach. The traditional approach supports a positive correlation between exchange rates and stock prices as depreciation of domestic currency makes local firms more competitive leading to increased exports and thus, raising the stock prices. On the contrary, the portfolio approach argues that an increase in stock prices leads to an increase in demand for domestic assets which in turn results in appreciation of the local currency. Thus, stock prices fluctuations lead to a drop in exchange rate.

Industrial Production: The index of Industrial Production is an economic tool that measures the changes in industrial activity in a country over a given period of time and is typically used as proxy for the level of real economic activity. Theoretically, a positive relationship is expected between the stock returns and industrial production as the productive capacity of an economy increases during economic growth. It also contributes to the ability of the firms in generating cash flows hence positively affecting the stock prices. Therefore, an increase in industrial production is likely to reduce the probability of a stock market crisis.

3.6. Analytical Tools and Techniques

This section deals with the tools and techniques used for analyzing the data. A set of analytical techniques have been used to examine empirically each of the objectives. The employed econometric techniques and methodology are discussed below which is followed by the discussion of tools employed to carry out analysis of data.

3.6.1. Methodology for study of banking crisis

The study combines parametric- Logit Regression, non- parametric -Signaling approach and semi-parametric- Neural Networks to identify and test the power of chosen indicators in predicting a banking crisis. The above-mentioned techniques have been selected based on the comprehensive literature review. As observed, a large number of studies on prediction of crises have employed signal extraction approach and binary response model regressions like logit and probit regressions. Despite being significantly popular, signal approach and logit/probit models have been limited in their predictive abilities of crises. Therefore, the emerging domain of AI (Artificial Intelligence) and ML (Machine learning) techniques has been examined as an alternative in addressing the limitations constituted by conventional techniques. A logit regression is an appropriate approach as the dependent variable is dichotomous in nature categorizing the periods of turbulence and tranquility. Both logit and probit models have essentially the same interpretations, however, logit entails comparatively easier interpretations of the estimated results and found to be employed more in EWS literature. ANNs, on the other hand, have significant advantages as they can implicitly detect complex non-linear relationships between the input and output variables. Further, ANNs offer flexibility in terms of modelling the data through various training algorithms and have the ability to learn based on experience. This is in contrast to the logit models, which can also be used to model nonlinear complex relationships, however, it requires an explicit search for these relationships by the model developer and may need complex data transformations. The time period considered in this study is from February 2001 – March 2017. The first step of construction of EWS constitutes defining the crisis variable for identifying the stress events. To identify the periods of vulnerability, a dependent variable has been constructed based on index identification method by Kibritçioğlu (2002) to develop Banking Sector Fragility index. Based on the index threshold, a dummy binary variable has been deduced which has been adopted as the dependent variable for the analysis. The independent variables based on a detailed literature review from different studies have been considered to constitute a comprehensive set of variables like growth in Oil prices, growth in Stock prices, growth in Reserve money, interest rates, growth in Inflation, growth in Index of Industrial Production, growth in Current Account Balance relative to GDP, growth in Foreign Direct Investment relative to GDP, growth in Gross Fiscal Deficit relative to GDP, growth in Broad Money Supply relative to Foreign Exchange Reserves. The study attempts to compare the results of three approaches by comparing their Quadratic Probability Score (QPS) and Global Squared Bias (GSB) for the predictive performance. The methodologies adopted had been discussed in the following sections.

Non-Parametric: Signal Extraction Method

The signal approach constitutes monitoring one indicator at a time and identifying the instances when it deviates from its normal level. This normal level is termed as the "threshold" which varies with potential indicator. When the indicator crosses its threshold, it is said to signal an approaching crisis. A signal could be a good or bad signal. For example, if an indicator issues a signal of an approaching crisis and the crisis event actually occurs, it is said to be a "good" signal while if a crisis does not occur in within the crisis window, it is considered as a "bad" signal or noise. The threshold for each variable is calculated based on the noise to signal ratio (NSR). The value at which NSR is minimized is used as the optimal threshold, which is calculated as the ratio of bad signals to the good signals (Edison, 2003).

Let Xi be defined as an indicator variable. $\bar{X}_i$ be the threshold of the indicator. Each indicator is converted into a signal variable, S_i. If the indicator crosses its threshold value, X is said to "signal" a crisis. The signaling state is characterized by

$$S_i = 1 \; if \; |X_i| > |\bar{X}_i| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

If X_i does not cross its threshold, then there is no signal.

$S_i = 0 \; if \; |X_i| \leq |\bar{X}_i|$……………………………………………………………………………(2)

The outcome for an indicator yields a 2x2 matrix, shown in Table3-2. For each variable, there are four possible outcomes.

The conditions equations defined in Eq.1 and Eq.2 are expressed in their absolute values because for some variables, a rise above their threshold value indicates a rise in the likelihood of a crisis, while, for some variables, a decline below their threshold values indicate an increased likelihood of a crisis. While employing the signals approach, a signaling horizon or window is set which can be 6,12,18 or 24 months. This window constitutes the period within which an indicator is expected to signal a crisis. Kaminsky, Lizondo and Reinhart (1998) used a signal window of 24 months which was arbitrarily selected. However, according to El-Shazly (2002), the window can be chosen depending upon the country specific factors and data sample. Therefore, for the present study, a 12 month signaling window has been chosen. Table 1 shows the possible outcomes for an indicator.

Table 3-2:2x2 matrix representing outcomes of an indicator

	Crisis occurs within 12 months	**No crisis within 12 months**
Signal issued	**A**	**B**
No signal issued	**C**	**D**

Source: Kaminsky et al. (1998)

Cell A represents a “good” signal outcome when a variable crosses its threshold and issues a signal and the crisis occurs within the signal window. Cell B represents a “bad” signal or “noise” outcome when a variable issues a signal but the crisis does not occur within the signaling horizon. Cell C represents a “missed” signal i.e. the variable does not issue a signal but a crisis still occurs. Cell D represents the "good silent" i.e. when an indicator does signal an approaching crisis but a crisis does not occur. Using these four outcomes, the NSR for each variable is calculated as follows.

In terms of matrix terms, the Noise to Signal ratio (NSR) is defined as:

$$NSR = {}^{(B/B+D)}\big/_{(A/A+C)} \dots\dots\dots (3)$$

= Ratio of crisis incorrectly predicted to all non-crisis episodes/ Ratio of crisis correctly predicted to all crisis episodes

In essence, the threshold value is chosen so as to strike a balance between risk of having many bad signals and of missing many crises.

Thresholds have been defined relative to the percentiles of the distribution of the observations of the indicators. The optimal threshold has been selected after performing a grid search. The NSR is calculated for a range of potential threshold values and the value which minimizes the NSR became the threshold for that indicator. This translates into examining the upper tail distribution in case an indicator is positively related to the probability of occurrence of a crisis and lower tail in case an indicator is negatively related with the crisis occurrence. The indicators can be ranked based on the following criterion:

a) The lower the value of NSR, the greater the predictive power of the indicator. For an indicator whose NSR is greater than 1, it means that the indicator produces more false alarms than good signals. The persistence of a signal is defined as the inverse of the NSR. The more persistent the indicator in pre-crisis window (i.e. 12 months), the more useful an indicator is considered.

b) The difference between the conditional probability that is conditional on a signal from the indicator and the unconditional probability of the crisis. For an indicator to be useful, the conditional probability has to be higher than the unconditional probability (Ahec-Sonje, 1999-2002).

Thus if $({}^{A}/_{A+B})/({}^{A+C}/_{A+B+C+D})$ >1 , the indicator is useful.................................(4)

Construction of Composite Indicators

Generally, the likelihood of occurrence of a crisis is high when a greater number of indicators signal a crisis. Therefore, one way to capture the information from joint signaling is to construct a composite indicator. There are several ways to construct a composite indicator as reported by Kaminsky (1998). In this study, two methods of construction of indices have been considered.

The first composite index is constructed as summation of the number of the indicators signaling at an instance. Mathematically, CI^1 is defined as follows:

$$CI_t^1 = \sum S_t^j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (5)$$

Where S_t^j = 1 indicates an approaching crisis. Thus, CI^1 is the unweighted composite index.

While CI^1 provides an aggregate information, it fails to consider the forecasting accuracy of the individual indicators which leads to loss of information related to country's banking system fragility. The second composite index CI^2 takes care of this limitation by weighing more the signals issued by the indicators having more reliable forecasting. The index is the weighted sum of the indicators where the inverse of NSR of respective indicators is used as weights. Thus, the lower is the NSR, the higher the composite index weighs in favor of that indicator. Mathematically, the CI^{2}is defined as:

$$CI_t^2 = \sum_{j=1}^{m} w_j S_t^j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (6)$$

Where, $w_j = [(1/\theta_{j)}/ \Sigma\theta_i]$ and θ_j is the NSR for jth variable.

Probability of a banking crisis

The values of CI^2 becomes useful when converted in to a conditional crisis probability. Therefore, for each value of the composite indicator an associated probability of future crisis is calculated. Following Edison, (2003), conditional probability of a banking crisis for the CI^2 can be calculated as follows:

$$\Pr\left(BCrisis_{t,t+12} \middle| CI_l < CI_t < CI_u\right) = \frac{\sum months\ where\ CI_l < CI_t < CI_u\ subject\ to\ a\ crisis\ in\ next\ 12\ months}{\sum months\ where\ CI_l < CI_t < CI_u} (7)$$

where *Pr* denotes the probability, *BCrisis*$_{t,\ t+12}$ denotes the occurrence of the banking crisis in coming 12 months, $CI_l\ and\ CI_u$ represents the lower and upper intervals for the CI. Thus, $\Pr(BCrisis_{t,t+12} | CI_l < CI_t < CI_u)$ is the conditional probability that a banking crisis will occur within 12 months under the condition that the composite index ranges between lower and upper limits.

Parametric: Logit Model

The second approach employed in the present study is Logit regression. In statistics, a Logit model is a type of regression where the dependent variable can take two values, i.e. 0 for no crisis and 1 for crisis. Using the binary Logit model, this study has attempted to estimate the probability that an observation with particular characteristics will fall into one of the categories. The probability that a crisis occurs is assumed to be a function of the vector of explanatory variables.

Mathematically, the Logit equation takes the following form: $logit\ (y_t) =\ \beta x_t + c,$

$$logit\ y_t = ln\frac{y_t}{1-y_t} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where, y_t is the dummy crisis variable (Bernoulli- distributed), x_t is the set of explanatory variables, β_t is the vector of free parameters to be estimated. F is the cumulative distribution function which ensures that the predicted outcome of the model always lies between 0 and 1. Logit regression measures the relationship between the categorical variable and the independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution The estimated coefficients cannot be interpreted as the marginal effects of the independent variable on the dependent variable. The signs of the β_t coefficients represents the direction of the effect of the change in x_t on change in the probability of y_t. Therefore, a positive value of the coefficient can be interpreted as the rise in the probability of the crisis.

The goodness- of – fit for Logit models can be measured by McFadden R^2 or pseudo R^2.

$$pseudo\ R^2 = 1 - \frac{Log\ L}{Log\ L_0} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots..(9)$$

Where Log L is the average of the Log-Likelihood(LL) function without any restriction and Log L$_0$ represents the maximized value of LL function under the restricted case that all the slope coefficients except the intercept are restricted to 0. The value of pseudo R^2 always lies between 0 and 1.

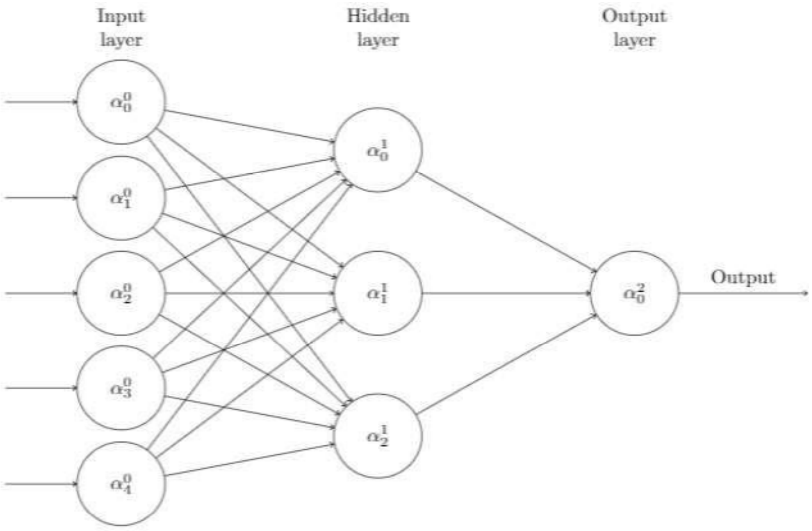
Artificial Neural Networks

A neural network is a network of connected output and input units which acquires knowledge using data provided and makes decisions based on the fed data. The neural network based modelsare preferred due to their high tolerance for noisy data thus, resulting in better predictive power compared to other models. The network constitutes three layers namely input, hidden and output layer. Each layer consists of a number of neurons/ units which are connected to each other and are assigned different weights. The weights and the input values determine the output values. These networks "learn" from the input provided at the input layer and adjust the weights till the output values get as close as possible to the actual target values thereby reducing the error (Roy, 2009). This process is called training of the network. After training the network, it is validated and tested for the remaining data in order to get most suitable fit for the sample provided.

Figure 3-2 presents the schematic diagram of single hidden layer neural network with three neurons, one input layer with 5 neurons, each neuron representing an input and one output layer with one neuron.

The two networks employed in the present study are Multilayered Feedforward Backpropagation network (MLFN) and Elman recurrent neural network. The two Neural networks differ in their structures as Elman considers lags of the inputs, where the output of hidden layers is fed back as an input in comparison to MLFN which has the inputs feeding in the forward direction only. The superiority of ANNs lies in the fact that it devises the relationship among the variables based on experience instead of hypothesizing a linear relation among the variables, like Logit or Probit models, which may not be the case.

Figure 3-2: Diagram of a single hidden layer ANN



Source: Medium.com

Construction of Crisis variable

The study has adopted the index method of identification of banking crisis. The index is based on Banking Sector Fragility (BSF) index developed by Kibritçioğlu to identify the exact months during which the Indian banking sector experienced crisis (Kibritçioğlu, 2002).

The BSF index is a composite index constituting Aggregate Time Deposits, Foreign Currency Borrowing, Net Bank Reserves and Domestic Credit as proxies for Credit risk, Liquidity risk, and Interest rate risk. Data has been sourced from Monthly RBI bulletin. The index has been labelled as BSF4 which is defined as an average of standardized values of Real Deposits, Real Foreign Currency Borrowing, Real Credit, and Real Bank Reserves. Following is the mathematical representation for the construction of the index.

$$BSF4 = \left[\left(\frac{Dep_t - \mu_{Dep}}{\sigma_{Dep}}\right) + \left(\frac{Cred_t - \mu_{Cred}}{\sigma_{Cred}}\right) + \left(\frac{FCB_t - \mu_{FCB}}{\sigma_{FCB}}\right) + \left(\frac{NBR_t - \mu_{NBR}}{\sigma_{NBR}}\right)\right]/4 \ldots\ldots(10)$$

Where $Dep = \frac{D_t - D_{t-12}}{D_{t-12}}$; $Cred = \frac{C_t - C_{t-12}}{C_{t-12}}$; $FCB = \frac{CB_t - CB_{t-12}}{CB_{t-12}}$; $NBR = \frac{BR_t - BR_{t-12}}{BR_{t-12}}$

The time series data on different variables was deflated using the Wholesale Price Index (WPI, Base: 1993-94=100). The data has been deflated to make all variables comparable as converting to the same base year helps taking care of any regime shifts in an economy. D_t, C_t, CB_t, *and* BR_t represent the Schedule Commercial Banks' real deposits, real credit, real foreign currency borrowings, and real net bank reserves respectively, in time t. Dep_t, $Cred_t$, FCB_t, *and* NBR_t, represent the annualized changes in real deposits, real credit, real foreign currency borrowings, and real net bank reserves respectively, in time t. The annual percentage change in the data has been used instead of the month- to- month variation to take care of seasonality and stationarity in the data. This transformation also considers that any difficulties in the banking sector are signaled by long-term fluctuations instead of short-term fluctuations. The mean and standard deviation of the variables have been represented by μ and σ respectively. Data on all the time series have been standardized so that no individual component may dominate the index.

When the value of BSF is greater than 0, it is a no-crisis zone. However, a value below 0 represents a condition of banking sector fragility. Based on the set threshold, the standard deviation of the index, the fragility has been distinguished between medium and high. In the case of present study, standard deviation of 0.54 is used as θ. The selection of the threshold is also based on Kibritçioğlu, (2002).

Medium: $-\theta < \text{BSF} < 0$

High: $BSF < -\theta$...(11)

This study considers the systemic banking crisis as a phenomenon of continuous alternating phases of medium and high fragility. A banking system is considered to be fully recovered when the value of BSF ≥ 0. Based on this continuum, a dummy binary variable has been deduced where a value of 0 represents a state of no crisis and a value of 1 represents a state of crisis.

Evaluation of Warning Systems

Kaminsky (1999) used QPS and GSB to evaluate the composite indicators, constructed through signal approach, for prediction of currency and banking crises. The study used QPS and GSB to test the accuracy and calibration of the indicators. Following Kaminsky (1999), Bhattacharya and Roy (2009) used QPS and GSB for assessing the predictive ability of the weighted and unweighted composite indicators in crisis prediction. Other studies which have used QPS and GSB are Budsyaplakorn et. al (2014), Roy (2009) Berg and Pattillo (1999) to name few. The present study has also employed QPS and GSB to compare the models in terms of their ability to predict a crisis

The assessment of the models is based on two attributes namely accuracy and calibration. The closeness of average of predicted probabilities and observed realizations indicates the accuracy of an indicator. Mathematically, QPS is defined as:

$$QPS = {}^{1}\!/_{T} \sum_{t=1}^{T} 2(P_t - R_t)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Where $0 \leq \text{QPS} \leq 2$. P_t is the predicted probability of the event at time t, T is the total number of observations in the sample and R_t is the realization of the event at time t. The highest possible value of QPS is 2 and the lowest is 0 which implies perfect accuracy. QPS test determines the discrepancy between the realization of an event R_t and its estimated probability P_t (Diebold and Rudebusch, 1989). The second attribute considered is Global Squared Bias (GSB) which measures the calibration. Mathematically, GSB is defined as:

$$GSB = 2(\bar{P} - \bar{R})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (13)$$

Where $\bar{R} = 1/T \sum_{t=1}^{T} R_t$, $\bar{P} = 1/T \sum_{t=1}^{T} P_t$ and $0 \leq \text{GSB} \leq 2$. The value of 0 corresponds to perfect calibration which occurs when average probability forecasts equal average realizations. Both the measures were calculated for in sample and out of sample performance. For out of sample prediction, the training set was taken from 2001 to 2014 and forecasting was done for 2015 to 2017.

Feature selection for identification of significant variables

One of the main caveats of machine learning models is that these complex techniques are subjected to the black box critique. Although their predictive performance might yield better results compared to conventional statistical models but have a limitation in identification of prominent critical factors that would contribute to occurrence of crisis.

In machine learning and statistics, feature selection entails the selection of a subset of relevant features out of a universe of predictors. There are three main categories of feature selection algorithms namely wrappers, filters and embedded methods[7].

There are two types of banking distress- individual bank distress and system-wide banking distress. The first approach focuses on micro level data relating to an individual bank like bankruptcy and insolvency of a bank which can lead to a systemic distress. The second approach takes a holistic view and relates to distress in the whole banking system at a macro level.

Bankruptcy prediction and credit scoring are the two main domains where machine learning tools have been employed to predict financial distress (Crook, Edelman and Thomas, 2007; Ravi Kumar and Ravi, 2007; Lin, Hu and Tsai, 2012). Apart from exploring sophisticated techniques for effective prediction, few studies have also examined the effect of feature selection on financial distress prediction (Liang et al., 2015). The studies which have used filter and wrapper based feature selection methods for bankruptcy prediction are Feki et al. (2012), Lin et al. (2011), Li and Sun (2011), Tsai (2009), and for credit scoring are Hajek and Michalak (2013), Chen (2012), Wang et al. (2012), Chen and Li (2010).

Thus, there are multiple studies which have used feature selection techniques for micro level financial distress prediction. However, there is no study employing feature selection methods to predict a banking crisis. This study proposes to identify the most relevant variables using stepwise selection using ANNs for effective prediction of a banking crisis on a macro level. The following section discusses about the feature selection categories and the procedure adopted in the current study.

[7] (http://jmlr.csail.mit.edu/papers/v3/guyon03a.html).

The problem of identifying the most relevant feature is tackled by using stepwise selection methods (a wrapper technique) namely forward stepwise and backward stepwise feature selectionmethods. The method is a classical stepwise method that consists of adding or rejecting an input variable step by step based on changes in their performance measure. The performance measure used is the R^2 value. The Mean Squared Error (MSE) is not used as there was very less differenceamong the values obtained for addition or elimination of the variables.

Both forward and backward selection methods were used to establish robustness in the feature selection process of the significant variables. Firstly, forward selection method was used to identify the variable which contributed to improving the model in terms of R^2, following which, the same set of 15 variables were subjected through backward selection process to remove the variables which did not contribute in improving the model. Forward selection method suffers from few drawbacks as it may produce suppressor effects. These suppressor effects occur when predictors are only significant when another predictor is held constant and therefore backward selection process helps in overcoming this effect as all the predictor variables are already added in the model.

Forward Stepwise: The first step in forward selection process entails finding out the variable with highest R^2 value. Hence, all the 15 variables are tested individually by constructing and training 15 Neural Network models and noting their respective R^2 values. Once the highest R^2 valued variable is identified, another input variable from the remaining variables is added to the neural network, generating 14 new constructs with other parameters kept constant. The new 14 models, each having two input variables are trained 10 times each and their respective R^2 values are noted. Again, the combination resulting in highest R^2value is identified. The procedure is carried out till all the variables are added sequentially in the neural network noting their respective R^2 values.

Backward Stepwise: The procedure is similar to forward stepwise but instead of addition, the input variables are eliminated one by one based on the R^2 values obtained. The deletion of the variable which results in highest decrease in R-squared is considered as the most important variable. Once the variable is identified, it is eliminated from the sample. After this step, another variable is eliminated from the remaining set of variables in combination with the first variable already eliminated. This results in new 14 networks, each having only 13 variables as input variables. The second eliminated variable which results in the highest decrease in the R^2 value is identified and this procedure is repeated until all the variables are eliminated.

Once both the procedures are completed, the variables which contributed in increasing the R^2 values in forward selection process are identified. Similarly, for the backward selection process, the variables which contributed in decreasing the R^2 value, are pulled out of the sample space. Then, a common subset of input variables is identified by taking variables obtained in forward and backward selection process. Each of the new neural network construct is trained 10 times with 1000 epochs with 70% of data for training, 15% for validation and 15% for testing. The iterative training process is followed to solve for an optimization problem that finds parameters that result in a minimum error. Finally, the performance of the discovered important variables is compared to that of the constructs involving all the variables in forward and backward selection process.

3.6.2. Methodology for the study of stock market crisis

The study compares the performance of Logit models and ANNs in predicting the probability of a stock market crisis. The first step constitutes the identification of the episodes of the stress in Indian stock markets. This has been carried out by developing an index and converting it into a binary dummy variable using a threshold. The second step constitutes identification of the variables is carried out based on comprehensive literature review. And, finally the predictive performance of logit regression is compared with that of ANNs using QPS and GSB. In addition, different logit models using combination of macroeconomic variables and investor sentiment proxies are also developed. The predictive performance of the developed Logit models is tested using different probability cutoffs and Area under Receiver Operating Curves (AuROCs).

Construction of crisis variable

The first step in developing an EWS is to identify the episodes of crisis. Patel and Sarkar (1998) were the pioneer in quantifying the definition of the stock market crisis. The study defined "a "crash" as a relative decline in the regional price index of more than 20 percent for the developed markets and more than 35 percent for the emerging markets". A ratio called CMAX, was constructed which compared the values of regional index at time t to the maximum regional index over the previous T periods, usually one to two years. A threshold was set at the mean of CMAX minus two standard deviations, such that the indicator would assume a value of unity whenever CMAX dropped below the threshold, zero otherwise.

Following Patel and Sarkar (1998), the CMAX has been constructed using both T= 12 months and T= 24 months. The CMAX ratio is calculated by dividing the current price by the maximum price over the previous T months' period.

$$CMAX_t = \frac{P_t}{\max(P_{t-T\ldots\ldots\ldots t})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(14)$$

Where P_t is the stock market price at time *t*. The denominator is the rolling maximum over a period of 12 months to avoid losing too many data points. The CMAX is also tested with a period of 24 months, however, CMAX with 12 months has been adopted as it fit the data more accurately. The CMAX assumes unity each time a crisis is detected i.e. when CMAX drops below the threshold set at the mean of CMAX minus one standard deviations calculated on whole sample.

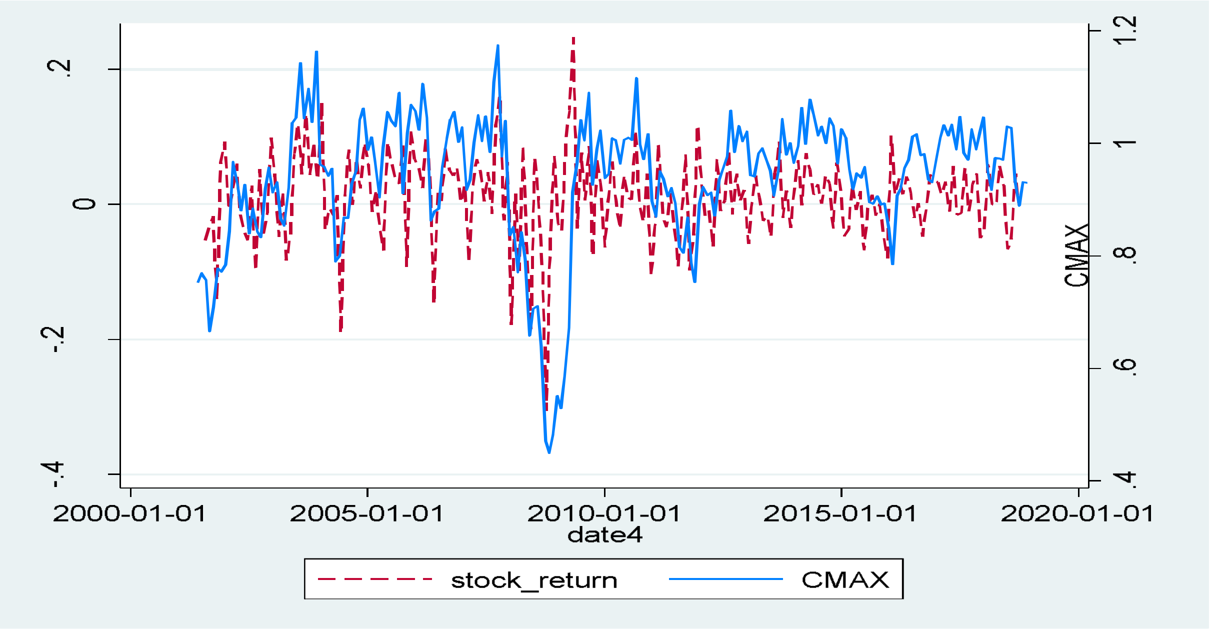
Mathematically, the crisis indicator is defined as follows:

$$C_t = 1 \; if \; CMAX_t < \overline{CMAX_t} - \sigma$$

$$C_t = 0, \qquad otherwise$$

Figure 3-3 depicts the trend of NSE stock returns and the CMAX index constructed for identifying the crisis episodes for the period from June 2001 to December 2018. It can be observed from the figure that CMAX follows same path as the stock returns and is able to identify the periods of vulnerability in the stock market. Two episodes can be identified from the Figure 3-3.

Figure 3-3: Trend of NSE Nifty 50 stock returns and CMAX from June 2001 to December 2018



Source: Based on Author's calculations

First episode is during the period from 2001 -2002 which was the result of burst of Dot com bubble. The second episode identified is during the period 2008-09, when subprime crisis hit the economies all over the world. These crises led to major (~19% during 2001-2002; >20% during

2008-09) declines in the stock returns as can be observed from the trend of stock returns in Indian context.

Construction of investor sentiment

Indian sentiment

There is no consensus on the best way to measure investor sentiment as large number of studies have adopted different proxies. As already mentioned in the literature review section, there are two ways to measure investor sentiment namely direct which involves surveys, and indirect such as using market related proxies. The following subsections discuss the construction of thebehavioral variables.

Based on the literature and availability of the data, seven market related proxies have been used to construct a composite investor sentiment index. The seven variables adopted are Advance Declining Ratio (ADR), Put Call Ratio (PCR), PE Ratio (PER), Turnover rate (TO), Trading Volume (TV), Number of IPOs (NIPO) and Mutual Funds Net Inflow (MFNI). ADR represents the ratio of number of advancing and declining stock prices which helps in gauging the trend of the stock market performance. A rising ADR indicates an upward trend of the stock market and falling values are indicative of downward trend. The ADR has been adopted by many studies such as Brown and Cliff (2004), Wang et al. (2006) and Dash and Mahakud, (2013a) as an indicator of investor sentiment. The variable PCR is representative of derivative trading activities. It is defined as the ratio of trading volumes of put-call options (Naik and Padhi, 2016). This variable indicates a bearish market when it assumes high value while in a bullish market, the values are low (Brown and Cliff, 2004).

PER is often positively correlated with the market and a high value of PER indicates high market sentiments (Sehgal et al., 2009; Zhu, 2012). It reflects both the financial situation of listed companies and the stock market in the prevailing macroeconomic environment. TO and TV represent the stock market liquidity. Baker and Stein (2004) suggests that market liquidity can be a sentiment indicator. According to the study, "An unusually liquid market is one in which pricing is being dominated by irrational investors". Studies like Qiang and Shu E (2009), Zhu (2012) and Bu and Pi (2014) used turnover rate as a sentiment indicator while Chuang et al. (2010) used trading volume as the investor sentiment index. Lastly, Mutual funds net inflow has been used as another sentiment proxy since mutual funds flows are one way to see how investors are feeling about the market. For example, when the perception about the market is bullish, the investors become more risk loving and the net inflows to equity based funds increase. Conversely, when there is a risk averse attitude prevailing in the market, the flows get directed towards bonds from equities. Few studies which have used mutual funds flow as sentiment indicator are Brown and Cliff (2004), Chi et al. (2012) and Dash and Mahakud (2013a).

Baker and Wurgler (2006) argued that some proxies take longer time to reveal the sentiment. Hence, following Baker and Wurgler (2006), a composite sentiment index has been constructed using variables at their levels as well as their lags. The index is constructed using Principal Component Analysis (PCA) with 14 variables which includes seven level variables and seven first lags of each variable. The following steps are adopted: First, each variable is regressed over the macroeconomic variables namely Inflation, industrial production, real interest rates, Foreign Institutional investors inflows and term spread and their respective residuals are taken as the non-fundamental component. Using these residuals, a raw sentiment has been constructed using the seven variables and their respective lags. Second, the correlation coefficient among the 14 variables and the first stage raw sentiment has been computed. Third, those variables (out of the levels and their respective lags) are selected which have higher correlation with the first stage raw sentiment for the construction of final sentiment index. The first principal component explains 34.68% of the overall variance and gives the following measure of sentiment index:

$$SENT = 0.4227\, PER_t + 0.3101\, NIPO_t + 0.6072\, TO_{t-1} + 0.5698\, TV_{t-1} + 0.0782 ADR_{t-1} - 0.1097\, PCR_t + 0.1169\, MFNI_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots. (15)$$

The results for PCA are presented in Table 3-3 and Table 3-4 below. It is observed that the first principal component explains 34.68% of the variance which is the highest among all the other components. Therefore, the first principal component has been selected which is a linear combination of all the proxies included.

Table 3-3: PCA analysis for construction of Indian sentiment variable

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.427	1.153	0.346	0.346
Comp2	1.274	0.271	0.182	0.528
Comp3	1.002	0.155	0.143	0.672
Comp4	0.847	0.102	0.121	0.793
Comp5	0.744	0.111	0.106	0.899
Comp6	0.633	0.564	0.090	0.990
Comp7	0.069	.	0.009	1

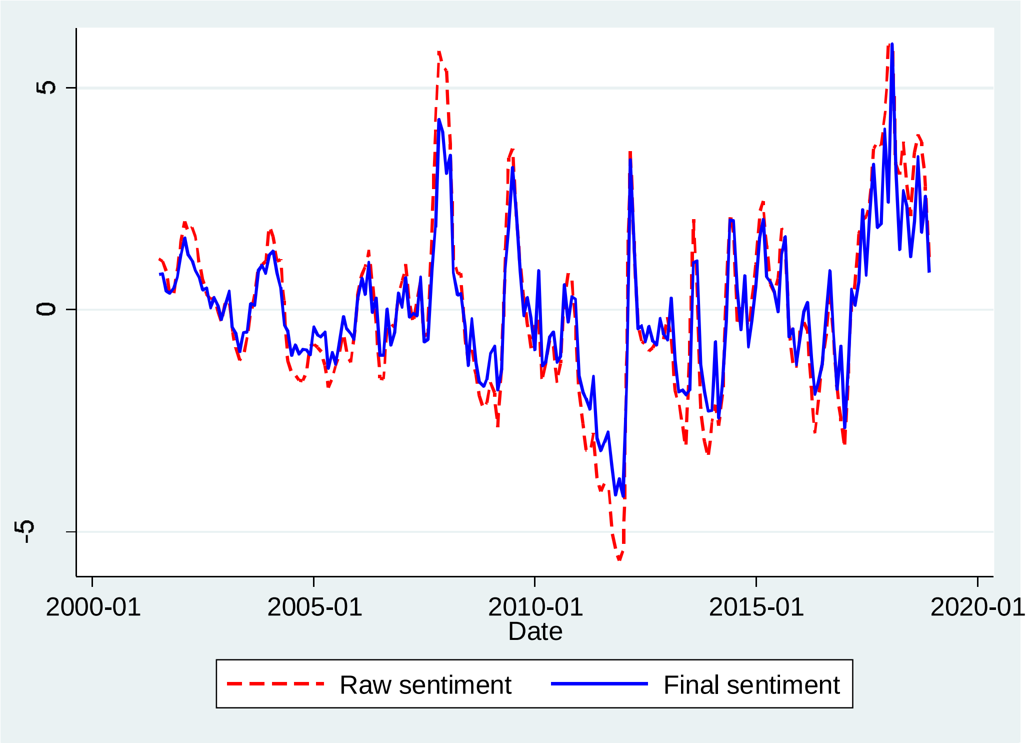
Table 3-4:Principal Components calculated for Indian sentiment

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
res_PER	0.422	-0.187	0.289	0.105	-0.212	-0.787	0.163	0
res_NIPO	0.310	-0.372	0.091	-0.424	0.757	0.033	0.031	0
res_PCR	-0.109	-0.440	0.728	0.332	-0.089	0.379	-0.012	0
L1_resTO	0.607	-0.025	-0.106	0.095	-0.172	0.200	-0.735	0
L1_resTV	0.569	0.099	-0.154	0.181	-0.109	0.414	0.651	0
L1_resADR	0.078	0.623	0.280	0.487	0.522	-0.097	-0.083	0
L1_resMFNI	0.116	0.481	0.513	-0.646	-0.241	0.117	0.0107	0

Table 3-4 represents the principal components calculated for constructing the Indian sentiment. The first component has been selected as the proxy for Indian investor sentiment based on the highest variance explained. The negative loading to the few variables under the first principal component can be attributed to the negative correlation between the variables. Given that the variable Put-Call ratio is a contrarian indicator i.e., high value of the ratio indicates that the market participants are bearish and therefore, opt for buying put options compared to call options. This results in high trading volume of put options over call options thereby, increasing the ratio value (Kumari and Mahakud, 2016). Even though eigenvalues for first three components are greater than 1, only first component has been considered due to the variance proportion explained. Also, the first component was selected because it constituted all the variables with correct sign. The first component carries highest loadings for variables related to market turnover (0.607) and trading volume (0.569). This is expected as in any market where there are constraints on short sales, liquidity is added in the markets by irrational participants due to high prevailing optimism. As expected, a positive relationship is found between the sentiment and market liquidity. This is followed by the PE ratio (0.422) and number of IPOs (0.310) which again, are expected to have a positive relationship with the market sentiment. High IPO numbers are indicative of the investor's enthusiasm and signify high sentiment.

The correlation coefficient between the raw sentiment index comprising 14 terms and the final sentiment index having seven terms was found to be 95.61% which suggested that the risk of losing substantial information due to dropping of the other seven proxies was very less. Figure 3- 4 depicts the trend between the raw and final stage sentiment. It can be observed that both trends are highly correlated. Figure 3-5 depicts the trend of the Nifty 50 index and the final sentiment for the period from June 2001 to December 2018. It can be observed from the figure that both the final sentiment and closing price follow same trend. However, the sentiment is more volatile than the closing prices. It can also be observed that during subprime crisis of 2008, both sentiment and Nifty stock index fell drastically simultaneously and recovered as well.

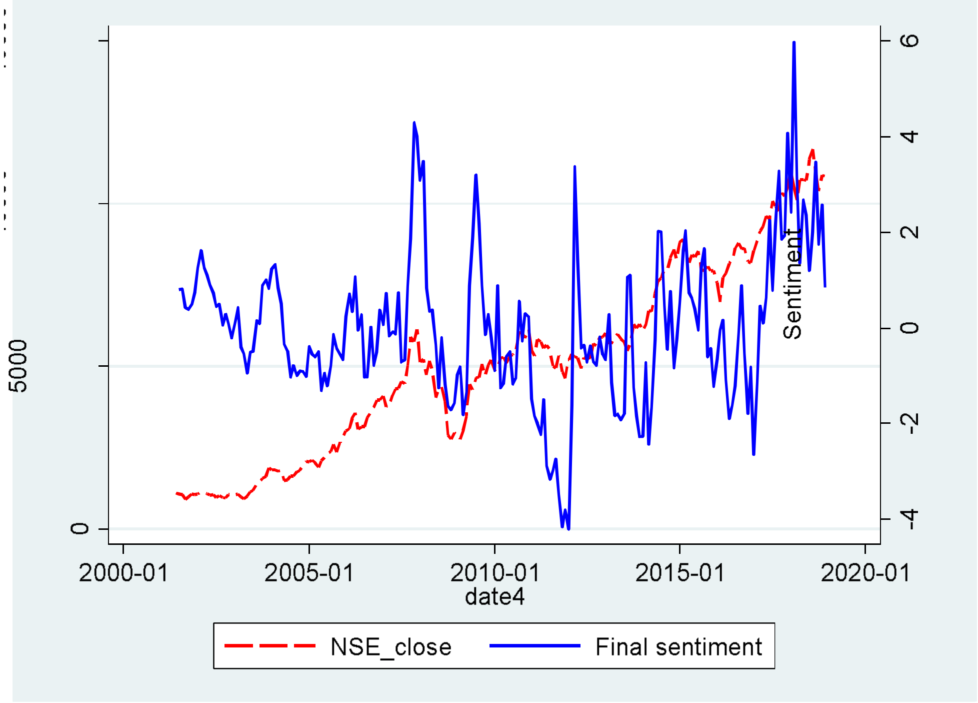
Figure 3-4:Trend of raw and final sentiment for India



Source: Based on Author's calculations

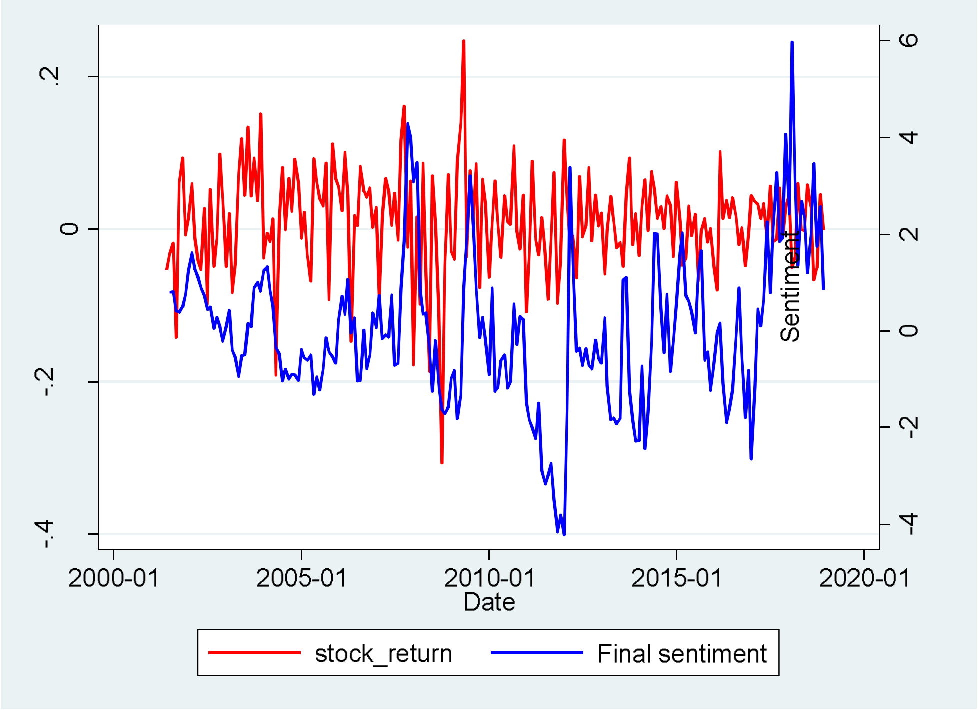
Figure 3-4 shows the trends of raw and final stage sentiment constructed based on methodology proposed by Baker and Wurgler (2007). The correlation between the raw and the final stage sentiment is found to be 95.61% which is confirmed by the trends shown above. Figure 3-5 represents the trends for closing prices of Nifty 50 index and the investor sentiment. Investor sentiment is more volatile as compared to the closing prices of the index. Both the trendshave a positive relationship. Figure 3-6 shows the trends for Nifty 50 stock returns and investor sentiment. It can be seen that apositive relationship exists between the two. Whenever there is a decline in stock returns, it is accompanied by a decline in the value of investor sentiment as well.

Figure 3-5:Closing price of Nifty 50 and investor sentiment



Source: Based on Author's calculations

Figure 3-6:Nifty 50 stock returns and Final stage sentiment



Source: Based on Author's calculations

Emerging market sentiment

The aggregate emerging market sentiment (EMERSENT) has been constructed using the third principal component of the orthogonal CCI data for 5 emerging market countries namely Mexico, Poland, South Korea, Russia and South Africa. The countries selected are the prominent emerging countries. However, China, which is one of the dominant emerging economy, has been dropped from the study owing to the unavailability of the data. The emerging countries namely Poland, Mexico and South Korea have been chosen based on their trade linkages with India while, countries namely Russia and South Africa have been chosen as they constitute the BRICS nations. Following the Baker and Wurgler (2006), the country specific CCI is regressed over the country specific macroeconomic variables namely Inflation, Industrial production, term spread, and money supply. This is known as orthogonalizing the variables to extract the non-fundamental or 'irrational' component known as the sentiment. Once all the non-fundamental components are extracted, PCA is utilized to construct an aggregate emerging market sentiment. The third component of the PCA is taken as the variable representing the emerging market sentiment.

Developed market sentiment

To study the effect of foreign investor sentiment on the probability of a stock market crisis in India, Consumer Confidence index (CCI) data for the U.S. market has been taken from the University of Michigan Survey of Consumers which provide a monthly survey since 1978.

Similarly, the effect of the sentiment in U.K. has been studied as well. The Eurozone Consumer Confidence Index (CCI) has been used to represent the Eurozone sentiment. The CCI has been orthogonalized to only study the impact of sentiment in predictive performance of the model.

Logit Models

The logit models have been estimated using different combinations of the macroeconomic and sentiment variables. Firstly, the model with only selected macroeconomic variables is estimated. Subsequently, this model is tested by including different sentiment variables one by one. After checking the incremental predictive power of each sentiment variable, the models with different combinations of sentiment variables are estimated.

Artificial Neural Network

After estimating the models using logit, ANNs have been used to assess the performance of the models. The study employs pattern classification neural network to identify the instance of extreme stress in stock market using macroeconomic and sentiment variables. The network that has been used for pattern recognition is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer and the number of output neurons is set to 1. The optimal number of neurons in the hidden layer has been decided based on the k-fold cross validation process. Once the optimal number of neurons is decided, the network has been trained on the sample with random division, with 70% used for training, 15% for validation, and 15% for testing. The construct for the neural network employed has been discussed in the Results section.

Cross Validation

Cross validation is a statistical resampling procedure which helps in model selection for the predictive modelling problem at hand. In machine learning models, this process helps in evaluation of models for a limited data set. The procedure is often called k-fold cross validation because it entails specification of a single parameter which refers to the number of groups into which a given dataset has to be split. For example, a value of k=10 means it’s a 10-fold cross validation and the data sample will be split into 10 groups for evaluation of the model. The idea behind this process is that a limited sample is used to estimate the expected performance of the model when it is used to make predictions for the data other than the one which was used during its training. The general procedure consists of four steps. First, the data gets shuffled randomly and gets split into k groups. Since the categorical variable in the present study is unbalanced in nature, i.e. number of crisis instances is less than the number of tranquil instances, the procedure used constitute "Stratified" division of the dataset. Once the data is divided into k groups, for each group, one group is taken as the test data set and the remaining groups are taken as the training data set. Using training dataset, a model is fitted which is then evaluated for the test dataset. This process is repeated and the average of the evaluation scores is taken as the measure of the skill of the model. This is done for different number of neurons in the hidden layer. For each n number of neurons in the single hidden layer, the mean performance calculated 5 times and the average of these 5 performance measures, in present case- classification accuracy, is taken to determine the best performing neural network for the given sample. The process of cross validation is depicted in the Figure 3-7 below.

Figure 3-7: Process of cross-validation

Split 1	***Fold-1***	**Fold-2**	**Fold-3**	**Fold-4**	**Fold-5**	**Metric 1**
Split 2	**Fold-1**	***Fold-2***	**Fold-3**	**Fold-4**	**Fold-5**	**Metric 2**
Split 3	**Fold-1**	**Fold-2**	***Fold-3***	**Fold-4**	**Fold-5**	**Metric-3**
Split 4	**Fold-1**	**Fold-2**	**Fold-3**	***Fold-4***	**Fold-5**	**Metric-4**
Split 5	**Fold-1**	**Fold-2**	**Fold-3**	**Fold-4**	***Fold-5***	**Metric-5**

Training Data ***Test Data***

Source: towardsdatascience.com

The model forecasting ability

Traditional evaluation of binary classification problems constitutes confusion matrix specifying four values namely i) True Positive (TP) which denotes the number of positive cases correctly classified as positive, ii) False Positive (FP) which denotes the number of negative cases classified incorrectly as positive, iii) True Negative (TN) which denotes the number of negative cases correctly classified as negative, and iv) False Negative (FN) which denotes the number of positive cases classified incorrectly as negative. Based on these notations, Type I error and Type II error are defined where Type I = FN/(FN+TP) [misclassified positive cases to the total positive cases] and Type II = FP/(FP+TN) [misclassified negative cases to the total negative cases]. Based on these four parameters, a measure of model predictive performance is defined called Accuracy which is calculated as: (TP+TN)/(TP+TN+FP+FN). This calculates the proportion of correctly classified cases.

To evaluate the performance of the model, two approaches have been utilized. First technique is comparing models using different probability cutoffs and second is comparing models using Receiver Operating Characteristic (ROC) for in-sample and out of sample. Area under Curve (AUC) is a popular measure based on ROC. In an ROC, the vertical axis shows the true positive rate (sensitivity) i.e. the proportion of positive cases correctly identified, and the x-axis shows the false positive rate (specificity) i.e. the proportion of negative cases incorrectly identified as positive cases. A perfect model would have a value of 1 indicating all the positive cases identified correctly. Therefore, AUC closer to 1 indicates that the model is assigning the right class to the data point.