JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

USING UNSTRUCTURED DATA WITH STRUCTURED DATA FOR SEGMENTATION OF NIFTY 50 STOCKS

[1]Sujata Suvarnapathaki

[1]Assistant Professor

[1]Department of Statistics

[1]Ramnarain Ruia Autonomous College, Mumbai, India

Abstract:

Nifty 50 stocks are a group of top-performing equity stocks on the National Stock Exchange of India. It's a good idea for investors to invest their capital in components of Nifty 50. It may reduce the investment risk to some extent since the Nifty 50 involves a varied range of stocks.

Investors are often interested in understanding the behaviour of stocks, in general. A prior idea of the worth of a stock or the direction of the movement of stock would always benefit the investors. The stock market analysis enables the investors to make the informed decision of "Invest", "Do not invest" or "Hold (If already invested)".

For this purpose, the investors may rely on significant fundamental analysis parameters and technical analysis parameters for understanding the behaviour of the stocks based on the historic data. But the public sentiments may also play a big role in behavioural finance and the financial decisions of investors may be influenced by the public sentiments. The economic indicators and the public sentiments are generally interrelated. With access to social media like Facebook, Twitter etc. people voice their opinions publicly about social, and political events and express their feelings. These data on social media are unstructured, hence may not be analyzed statistically but still may be useful.

This research paper includes the use of unstructured and structured data for the segmentation of Nifty 50 stocks. The structured data comprises one fundamental analysis parameter, viz. 'Price to Earnings Ratio' (P/E Ratio) and one technical analysis parameter viz. 'Relative Strength Index (RSI), for the segmentation of Nifty 50 stocks. Sentiment analysis is used as a Text Mining technique to extract information from the unstructured data from Twitter. The unstructured data in the form of tweets, for each stock in Nifty 50 is used for deriving an additional parameter Sentiment Score using which the 'Proportion of Positive sentiment' is computed. The data on 100 recent tweets about the stocks in Nifty 50 is fetched. The segmentation of Nifty 50 stocks into k clusters is then achieved using the k-means clustering method, based on the three variables, viz. RSI, P/E Ratio and the Proportion of Positive Sentiment. These "k" clusters of Nifty 50 stocks can help the investors to decide on trading actions like "Invest" or "Do not invest" or "Hold (If already invested)". The above analysis is carried out using R programming.

***Index Terms*-Unstructured data, Text Mining, Sentiment Analysis, Twitter, Stock market, Cluster analysis, R**

I Introduction:

Nifty 50 is a group or a basket of the 50 most active stocks on the National Stock Exchange of India that acts as an indicator for the overall movement of the stock market. The traders or investors are always interested in predicting the stock's movement. The economic indicators and the public sentiments are generally interrelated. The public sentiments may play a role in behavioural finance and financial decisions are influenced by the public sentiments. However, the public sentiment may exist in the form of unstructured data on social media like Facebook, Twitter, Instagram etc.

The use of social media like Facebook and Twitter has immensely contributed to the generation of unstructured data (that use natural languages) in addition to the traditional text documents. These data form a large proportion of data which is not analysed. Natural languages are different from programming languages. Natural languages are ambiguous. The meaning of a statement largely depends on the situation, tone and some other factors too.

Some common text mining applications include sentiment analysis and text classification. Text Mining is also known as Text Data Mining (TDM) and Knowledge Discovery in Textual Database (KDT). Text mining is the process of extracting significant patterns or information from text documents.

This study describes the use of Sentiment analysis with R programming, as an application of the Text Mining techniques to extract an additional parameter sentiment score viz. the "Proportion of Positive sentiment" from the Tweets about Nifty 50 Stocks.

The fundamental analysis parameter, viz. Price to Earnings Ratio (P/E Ratio) and the technical analysis parameter viz. Relative Strength Index (RSI) are the two parameters obtained from the structured data. An additional parameter viz. the Proportion of Positive sentiment computed from sentiment score, is a derived variable from the unstructured data from Twitter. These three variables are obtained for each stock of the Nifty 50 basket and are further used for segmentation of the Nifty 50 stocks. The cluster analysis using the k-means method allows the segmentation of Nifty 50 stocks. The objective is to categorize Nifty 50 stocks into "k" clusters which can help investors to decide trading actions of "Invest", "Do not invest" and "Hold (If already invested)".

II Methodology:

This study uses structured and unstructured data for the segmentation of Nifty 50 stocks. The Nifty 50 stock list is obtained from the website of the National Stock Exchange of India. The structured data comprises one fundamental analysis parameter, viz. Price to earnings Ratio (P/E Ratio) and one technical analysis parameter viz. Relative Strength Index (RSI). The unstructured data is fetched from Twitter and sentiment analysis is performed to derive a sentiment score for each tweet and for each stock in Nifty 50. The 'Proportion of positive sentiment', corresponding to each stock in Nifty 50 is computed. Further, the K-means clustering method is used for the segmentation of these Nifty 50 stocks. Thus, the three variables viz. P/E Ratio, RSI and Proportion of Positive sentiment are derived for each stock in Nifty 50.

2.1 Price to Earning Ration (P/E Ratio):

The relationship between the stock price of a company and its per-share earnings is measured by the price-to-earnings ratio. The P/E ratio allows the investors to check if a stock is undervalued or overvalued relative to others in the same sector. Investors compare the P/E ratios of competing stocks to understand the stock market better.

A low P/E ratio indicates that the stocks are available for a low price, indicating a better opportunity for investors to invest in such stocks.

The Price to Earnings Ratio is calculated as the ratio of the current price per share of stock to the company's earnings per share.

The data for the earnings and the Price of Nifty 50 stocks is obtained using the link "h*ttps://www.edelweiss.in/oyo/equity/user/screener"* and the P/E ratio is calculated for the Nifty 50 stocks.

2.2 Relative Strength Index (RSI):

Relative strength Index/Indicator aims to forecast the financial market direction which could be very helpful for the investors to decide about their course of action of "Invest" or "Do not invest" or "hold". It is a momentum oscillator that measures the speed and change of price movements of the stocks. RSI takes values from 0 to 100. Based on the values, the trader knows whether the prices of the stocks are in the overbought (RSI > 70) or oversold (RSI < 30) region.
The Relative Strength Index is calculated as follows:

$$RSI = 100 - (100/1+RS) \text{ where,}$$

$$RS \text{ (Relative Strength)} = \text{Average of 14 days up closes / Average of 14 days down closes}^{**}$$

**In this analysis, RSI based on 14-day trading period is calculated for Nifty 50 stocks, using R Programming.

2.3 Sentiment Analysis:

Text mining is getting a lot of attention these days due to an exponential increase in digital text data from web pages and social media services like Facebook, Twitter etc.

Twitter data constitutes a rich source that can be used for capturing information about any topic or event. These data can be used for finding trends related to a specific keyword, measuring sentiment, and gathering feedback about new products and services, social and political events etc. These are unstructured data.

Text analytics includes creating a data cloud (word cloud), which is the most popular way to visualize and analyze qualitative data. It's an image composed of keywords found within a body of text, where the size of each word indicates its frequency in that body of text. On the other hand, sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining as it helps to derive the opinion or attitude of a speaker.

This study uses sentiment analysis of the Twitter data on Nifty 50 stocks, using R programming. Twitter data for the Nifty 50 stocks are collected using the library "rtweet", in R. To use the library 'rtweet', one needs a Twitter account. Further, a Twitter developer account needs to be created. 'rtweet' is authorized specific account credentials. Nifty 50 stock list is mapped to the list of stocks with the Twitter handle. Stocks which do not have a Twitter handle are removed from the list of Nifty 50 stocks. 48 stocks out of Nifty 50 stocks are observed to have Twitter handles.

The most recent one hundred tweets for the stocks with Twitter handles are imported using 'search_tweets()' function in R. The sentiment score is obtained using the function 'sentiment()' available in the library 'sentimentr'. The proportion of positive sentiment is further obtained using the positive values of the sentiment score. Using 100 tweets for each stock, a sentiment score is used to obtain the Proportion of positive sentiment, thereby converting unstructured data into structured data.

The final data for the analysis consists of the Price to Earnings Ratio (P/E Ratio), Relative Strength Index (RSI) and the Sentiment Score, viz. the Proportion of positive sentiment, for 48 stocks. This data is further used for K-means cluster analysis for segmentation of Nifty 50 stocks.

2.4 k-means Clustering:

Cluster analysis is used for segmentation purpose. Cluster analysis is used to categorize objects or cases into clusters. A cluster is a group of relatively homogeneous cases or observations. Cluster Analysis is one of the unsupervised learning methods. K-Means Clustering is a popular non-hierarchical clustering method. The number of clusters (k) must be known a priori. Generally, the cluster solutions for different values of k are tried and evaluated to get the best possible cluster solution. In this case, values of k=3 and 4 are tried and the best possible and interpretable solution is obtained for k=4 in the cases discussed below. R programming is used for the k-means clustering method to analyze data with 48 stocks.

III Results and Discussion:

All the analysis variables viz. Price to Earnings Ratio (P/E ratio), Relative Strength Index (RSI) and Sentiment score: Proportion of positive sentiment are dynamic and their values change every day. The following results are based on a specific day.

3.1 Case I: The final data has 48 stocks with three variables viz. RSI, P/E Ratio and Sentiment score. Cluster analysis is performed on the final data, with 4 clusters. The data is as of 25th May 2022. The number of stocks in each cluster along with the cluster centroid (averages) are displayed in Table 1.

Table 1: The average scores of RSI, Proportion of Positive sentiment and P/E ratio for the data as of 25th May 2022

Segment	Number of Stocks (48*)	RSI	Proportion of Positive Sentiment	P/E Ratio
1	8	42.3483	0.2813	8.7267
2	8	44.8142	0.4237	83.1423
3	12	60.7587	0.4933	33.1152
4	20	34.2048	0.4550	23.0956

*Twitter data is unavailable for 2 stocks. Hence the total number of stocks is 48.

3.1.1 It is observed that in segment 3 and segment 4, the average proportion of positive sentiment is high. However, in segment 4, the P/E ratio is relatively less indicating that segment 4 could be an "Invest" segment whereas segment 3 can be a "Hold" segment. Investors can refrain from investing in segment 1 due to the low average Proportion of positive sentiment even if the P/E ratio indicates that the stocks in segment 1 are available at a low price. The second segment has a relatively high sentiment score compared to segment 1 but the stocks in this segment are very expensive hence the investors can avoid fresh investment in the stocks belonging to this segment. Table 2 displays the list of the stocks as per the cluster membership in 4 segments.

Table 2: List of the stocks in each segment i.e. Cluster Members for the data used as of 25th May 2022

Segment 1 (Number of Stocks = 8)	Segment 2 (Number of Stocks = 8)	Segment 3 (Number of Stocks = 12)	Segment 4 (Number of Stocks = 20)
COALINDIA	ASIANPAINT	BAJAJ-AUTO	APOLLOHOSP
GRASIM	BHARTIARTL	BRITANNIA	AXISBANK
HDFC	HDFCLIFE	CIPLA	BAJAJFINSV
HINDALCO	MARUTI	DRREDDY	BAJFINANCE
INDUSINDBK	NESTLEIND	EICHERMOT	BPCL
NTPC	SBILIFE	HEROMOTOCO	HCLTECH
POWERGRID	TATACONSUM	HINDUNILVR	HDFCBANK
TATAMOTORS	TITAN	ITC	ICICIBANK
		KOTAKBANK	INFY

LT	JSWSTEEL
M&M	ONGC
SUNPHARMA	RELIANCE
	SBIN
	SHREECEM
	TATASTEEL
	TCS
	TECHM
	ULTRACEMCO
	UPL
	WIPRO

3.2 Case II) The final data has 48 stocks with three variables viz. RSI, P/E Ratio and Sentiment score. Cluster analysis is performed on the final data, with 4 clusters. The number of stocks in each cluster/segment along with the cluster centroid (averages) are displayed in Table 3.

Table 3: The average scores of RSI, Proportion of Positive sentiment and P/E ratio for the data used as of 2nd June 2022.

Segment	Number of Stocks (48*)	RSI	Proportion of Positive Sentiment	P/E Ratio
1	10	46.8928	0.6270	22.0472
2	11	62.4196	0.3755	22.0045
3	12	50.3073	0.4025	76.6909
4	15	39.7551	0.3640	16.5156

*Twitter data is unavailable for 2 stocks. Hence the total number of stocks is 48.

3.2.1 It is observed that segment 1 has the highest average proportion of positive sentiment. Also, the P/E ratio is small, so this segment will be the "Invest" segment. The second segment has a low average proportion of positive sentiment also average RSI is high for this segment hence it indicates "Hold" segment. Segment 3 has a moderate average Proportion of positive sentiment and the average RSI is moderate too but the P/E ratio is high hence the stocks in this segment would be expensive. Investors may refrain from the fresh investment. Segment 4 has 15 stocks with a low average Proportion of positive sentiment, low RSI and low P/E ratio hence though the price of the stocks could be low it's not advisable to buy and invest in the stocks of this segment.
Table 4 displays the list of the stocks as per the cluster membership in 4 segments.

Table 4: List of the stocks in each segment i.e. Cluster Members for the data used as of 2nd June 2022

Segment 1 (Number of Stocks = 10)	Segment 2 (Number of Stocks = 11)	Segment 3 (Number of Stocks = 12)	Segment 4 (Number of Stocks = 15)
CIPLA	BAJAJ-AUTO	APOLLOHOSP	AXISBANK
COALINDIA	DRREDDY	ASIANPAINT	BAJAJFINSV
HCLTECH	EICHERMOT	BAJFINANCE	BPCL
LT	HDFC	BHARTIARTL	GRASIM
POWERGRID	HDFCBANK	BRITANNIA	HINDALCO
SHREECEM	HEROMOTOCO	HDFCLIFE	INDUSINDBK
TCS	ICICIBANK	HINDUNILVR	INFY
TECHM	ITC	MARUTI	JSWSTEEL
UPL	KOTAKBANK	NESTLEIND	NTPC
WIPRO	M&M	SBILIFE	ONGC
	TATAMOTORS	TATACONSUM	RELIANCE
		TITAN	SBIN
			SUNPHARMA
			TATASTEEL
			ULTRACEMCO

3.3 Case III) The final data obtained on 3rd June 2022, has 48 stocks with three variables viz. RSI, P/E Ratio and Sentiment score (Proportion of positive sentiments). Cluster analysis is performed on the final data with 4 clusters.
The number of stocks in each cluster along with the cluster centroid (averages) are displayed in Table 5 for this data.

Table 5: The average scores of RSI, Proportion of Positive sentiment and P/E ratio for the data used as of 3rd June 2022.

Segment	Number of Stocks (48*)	RSI	Proportion of Positive Sentiment	P/E Ratio
1	13	57.8042	0.4931	22.5854
2	11	38.5259	0.5664	16.7464
3	12	50.4563	0.4067	75.8478
4	12	50.1906	0.2892	19.1441

*Twitter data is unavailable for 2 stocks. Hence the total number of stocks is 48.

3.3.1 It is observed that the sentiment scores for segment 1 and segment 2 are relatively high. RSI values are moderate and P/E ratios are small indicating that the stocks are not costly and therefore investors can "Invest" in these two segments.
Investors can refrain from investing in segment 3 due to very high P/E ratio and in Segment 4 due to low sentiment.

Table 6 displays the list of the stocks as per the cluster membership in 4 segments.

Table 6: List of the stocks in each segment i.e. Cluster Members for the data as of 3rd June 2022

Segment 1 (Number of Stocks = 13)	Segment 2 (Number of Stocks = 11)	Segment 3 (Number of Stocks = 12)	Segment 4 (Number of Stocks = 12)
CIPLA	BPCL	APOLLOHOSP	AXISBANK
COALINDIA	GRASIM	ASIANPAINT	BAJAJ-AUTO
DRREDDY	JSWSTEEL	BAJFINANCE	BAJAJFINSV
HCLTECH	POWERGRID	BHARTIARTL	EICHERMOT
HDFC	SHREECEM	BRITANNIA	HEROMOTOCO
HDFCBANK	SUNPHARMA	HDFCLIFE	HINDALCO
ICICIBANK	TATASTEEL	HINDUNILVR	INDUSINDBK
ITC	TECHM	MARUTI	INFY
LT	ULTRACEMCO	NESTLEIND	KOTAKBANK
M&M	UPL	SBILIFE	ONGC
NTPC	WIPRO	TATACONSUM	SBIN
RELIANCE		TITAN	TATAMOTORS
TCS			

IV Conclusion:

The above approach uses the innovative method of combining unstructured data with structured data to analyze and segment the stocks in the Nifty 50. This approach can be useful for screening stocks of Nifty 50 before possible investment. Since the parameters are dynamic, the investors may perform this stock segmentation frequently (ideally daily). Thus, the results of the segmentation of Nifty 50 stocks can help the investors to decide trading actions on "Invest" or "Do not invest" or Hold (If already invested)".

V References:

[1] Aditya Bhardwaj et al. / Procedia Computer Science 70 (2015) 85 – 91
[2] Chetan Gondaliya et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1020 012023
[3] L. Zhang, Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, pp. 130, 2013
[4] https://www1.nseindia.com/content/indices
[5] https://www.edelweiss.in/oyo/equity/user/screener